Extractive Summarization via ChatGPT for Faithful Summary Generation

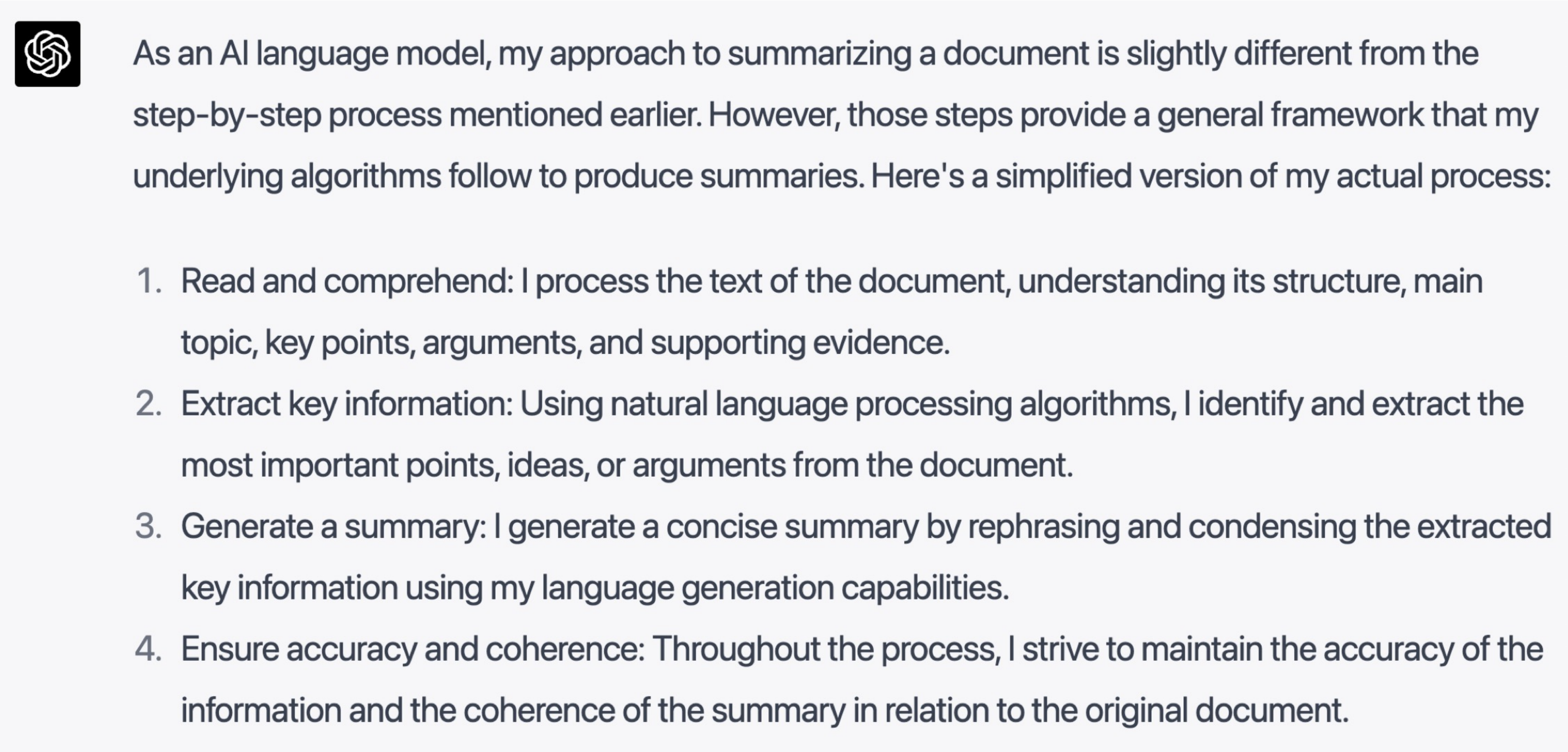
- ChatGPT 性能 (Rouge) 明显低于监督模型
- In-context Learning 和 Chain-of-Thought Reasoning 有帮助
- Extract-then-Generate 框架显著提高事实性

ChatGPT vs. Fine-tuning

- News summarization and evaluation in the era of gpt3 [Goyal 2022]
  - GPT-3 Rouge 分数略低，但是人类标注者倾向于 GPT-3 生成的文本
- Benchmarking large language models for news summarization [Zhang, 2023]
  - 新闻摘要（生成式），LLM 已经可以媲美人类
- 可能出现胡编乱造，难以实用（事实性）
  - 基于 LLM 的抽取式摘要

Extract-then-Generate

- How do you summarize a document?



As an AI language model, my approach to summarizing a document is slightly different from the step-by-step process mentioned earlier. However, those steps provide a general framework that my underlying algorithms follow to produce summaries. Here's a simplified version of my actual process:

1. Read and comprehend: I process the text of the document, understanding its structure, main topic, key points, arguments, and supporting evidence.
2. Extract key information: Using natural language processing algorithms, I identify and extract the most important points, ideas, or arguments from the document.
3. Generate a summary: I generate a concise summary by rephrasing and condensing the extracted key information using my language generation capabilities.
4. Ensure accuracy and coherence: Throughout the process, I strive to maintain the accuracy of the information and the coherence of the summary in relation to the original document.

In-context Learning and Explanation

- In-context Learning
  - 给出一些“输入-输出”示例，辅助模型预测

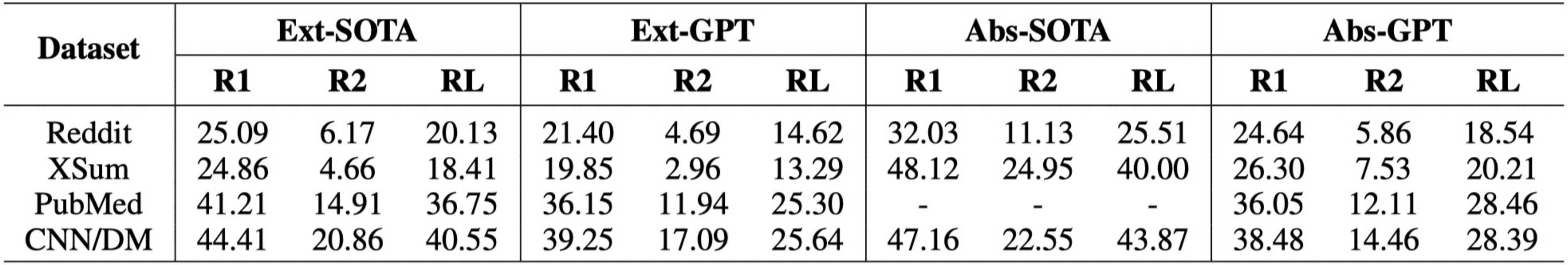
$$\hat{a} = \arg\max_{a} p_M(a \mid q, \{(q_1, a_1) \ldots (q_m, a_m)\}).$$

- Chain-of-Thought Reasoning (Explanations)
  - 显式展示推理过程，给出判断理由（解释）

$$\hat{a} = \arg\max_{a} \sum_{e} p_M(a, e \mid q, C) \qquad C = \{(q_1, e_1, a_1) \ldots (q_m, e_m, a_m)\}$$

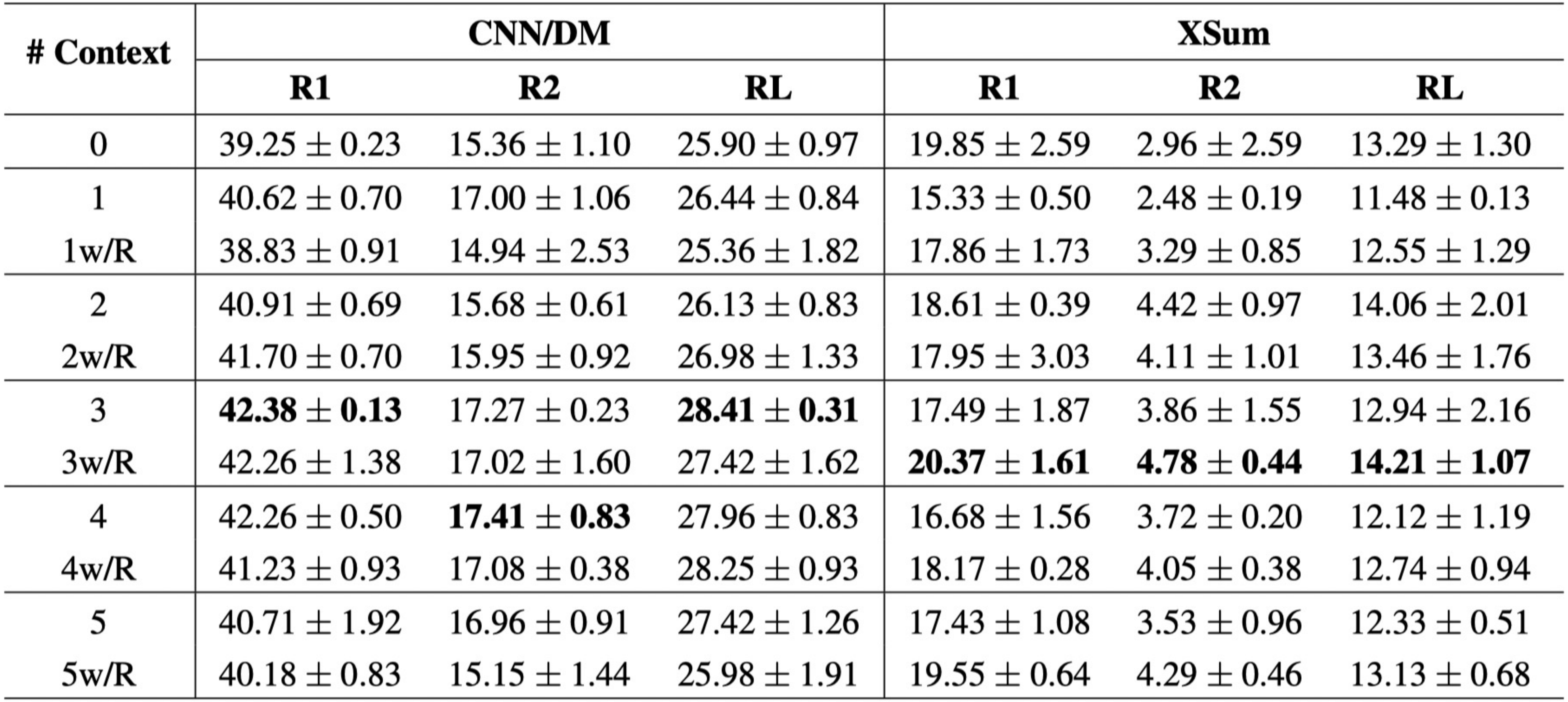
Setting	Prompt
Extractive	**System:** You are an extractive summarizer that follows the output pattern. **User:** Please extract sentences as the summary. The summary should contain *m* sentences. Document: [*Test Document*] [*Format Instruction*].
Abstractive	**System:** You are an abstractive summarize that follows the output pattern. **User:** Please write a summary for the document. Document: [*Test Document*] [*Format Instruction*]
In-context	**System:** You are an extractive summarizer that follows the output pattern. **User:** The following examples are successful extractive summarization instances: [*n Document-Summary Pair*s]. Please summarize the following document. Document: [*Test Document*]. The summary should contain *m* sentences. [*Format Instruction*].
Explanation	**System:** You are an extractive summarizer that follows the output pattern. **User:** The following examples are successful extractive summarization instances: [*n Document-Summary-Reason Triads*]. Please summarize the following document and give the reason. Document: [*Test Document*]. The summary should contain *m* sentences. [*Format Instruction*].
Extract-abstract	**System:** You are an abstractive summarizer that follows the output pattern. **User:** Please revise the extracted summary based on the document. The revised summary should include the information in the extracted summary. Document: [*Test Docuemnt*] Extractive Summary: [*Extractive Summary*] [*Format Instruction*].

性能低于传统微调模型



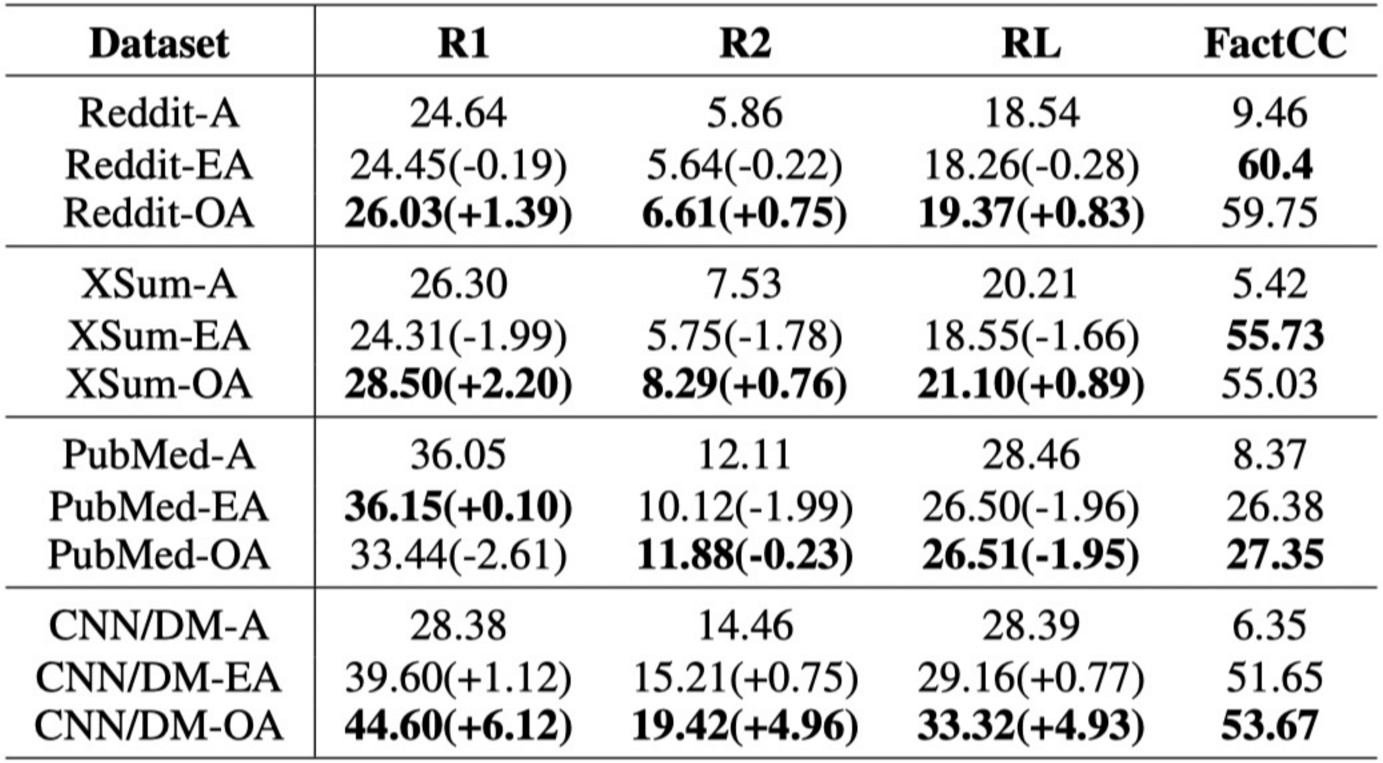
Dataset	Ext-SOTA			Ext-GPT			Abs-SOTA			Abs-GPT		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**	**R1**	**R2**	**RL**
Reddit	25.09	6.17	20.13	21.40	4.69	14.62	32.03	11.13	25.51	24.64	5.86	18.54
XSum	24.86	4.66	18.41	19.85	2.96	13.29	48.12	24.95	40.00	26.30	7.53	20.21
PubMed	41.21	14.91	36.75	36.15	11.94	25.30	-	-	-	36.05	12.11	28.46
CNN/DM	44.41	20.86	40.55	39.25	17.09	25.64	47.16	22.55	43.87	38.48	14.46	28.39

In-context Learning



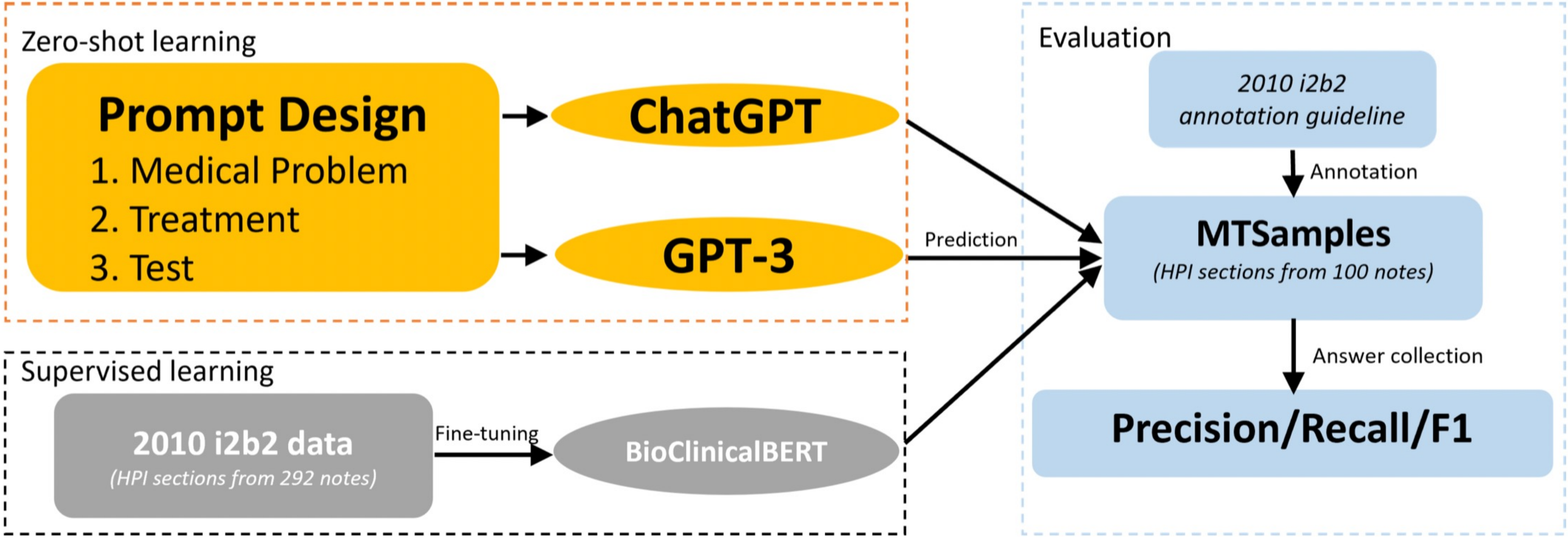
# Context	CNN/DM			XSum		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**
0	39.25 ± 0.23	15.36 ± 1.10	25.90 ± 0.97	19.85 ± 2.59	2.96 ± 2.59	13.29 ± 1.30
1	40.62 ± 0.70	17.00 ± 1.06	26.44 ± 0.84	15.33 ± 0.50	2.48 ± 0.19	11.48 ± 0.13
1w/R	38.83 ± 0.91	14.94 ± 2.53	25.36 ± 1.82	17.86 ± 1.73	3.29 ± 0.85	12.55 ± 1.29
2	40.91 ± 0.69	15.68 ± 0.61	26.13 ± 0.83	18.61 ± 0.39	4.42 ± 0.97	14.06 ± 2.01
2w/R	41.70 ± 0.70	15.95 ± 0.92	26.98 ± 1.33	17.95 ± 3.03	4.11 ± 1.01	13.46 ± 1.76
3	**42.38 ± 0.13**	17.27 ± 0.23	**28.41 ± 0.31**	17.49 ± 1.87	3.86 ± 1.55	12.94 ± 2.16
3w/R	42.26 ± 1.38	17.02 ± 1.60	27.42 ± 1.62	**20.37 ± 1.61**	**4.78 ± 0.44**	**14.21 ± 1.07**
4	42.26 ± 0.50	**17.41 ± 0.83**	27.96 ± 0.83	16.68 ± 1.56	3.72 ± 0.20	12.12 ± 1.19
4w/R	41.23 ± 0.93	17.08 ± 0.38	28.25 ± 0.93	18.17 ± 0.28	4.05 ± 0.38	12.74 ± 0.94
5	40.71 ± 1.92	16.96 ± 0.91	27.42 ± 1.26	17.43 ± 1.08	3.53 ± 0.96	12.33 ± 0.51
5w/R	40.18 ± 0.83	15.15 ± 1.44	25.98 ± 1.91	19.55 ± 0.64	4.29 ± 0.46	13.13 ± 0.68

Extract-then-Generate 框架



Dataset	R1	R2	RL	FactCC
Reddit-A	24.64	5.86	18.54	9.46
Reddit-EA	24.45(-0.19)	5.64(-0.22)	18.26(-0.28)	**60.4**
Reddit-OA	**26.03(+1.39)**	**6.61(+0.75)**	**19.37(+0.83)**	59.75
XSum-A	26.30	7.53	20.21	5.42
XSum-EA	24.31(-1.99)	5.75(-1.78)	18.55(-1.66)	**55.73**
XSum-OA	**28.50(+2.20)**	**8.29(+0.76)**	**21.10(+0.89)**	55.03
PubMed-A	36.05	12.11	28.46	8.37
PubMed-EA	**36.15(+0.10)**	10.12(-1.99)	26.50(-1.96)	26.38
PubMed-OA	33.44(-2.61)	**11.88(-0.23)**	**26.51(-1.95)**	**27.35**
CNN/DM-A	28.38	14.46	28.39	6.35
CNN/DM-EA	39.60(+1.12)	15.21(+0.75)	29.16(+0.77)	51.65
CNN/DM-OA	**44.60(+6.12)**	**19.42(+4.96)**	**33.32(+4.93)**	**53.67**

Zero-shot Clinical Entity Recognition using ChatGPT



Zero-shot learning

Prompt Design

1. Medical Problem
2. Treatment
3. Test

ChatGPT

GPT-3

Supervised learning

2010 i2b2 data

(HPI sections from 292 notes)

Fine-tuning

BioClinicalBERT

Prediction

Evaluation

2010 i2b2 annotation guideline

Annotation

MTSamples

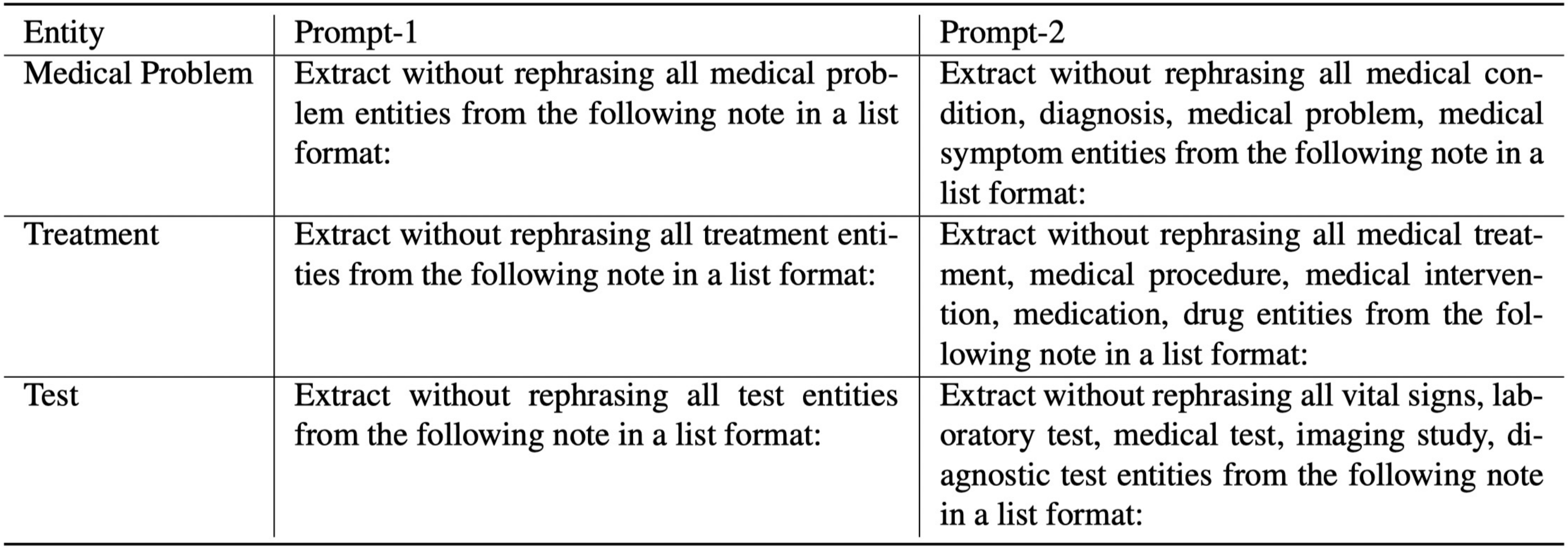
(HPI sections from 100 notes)

Answer collection

Precision/Recall/F1

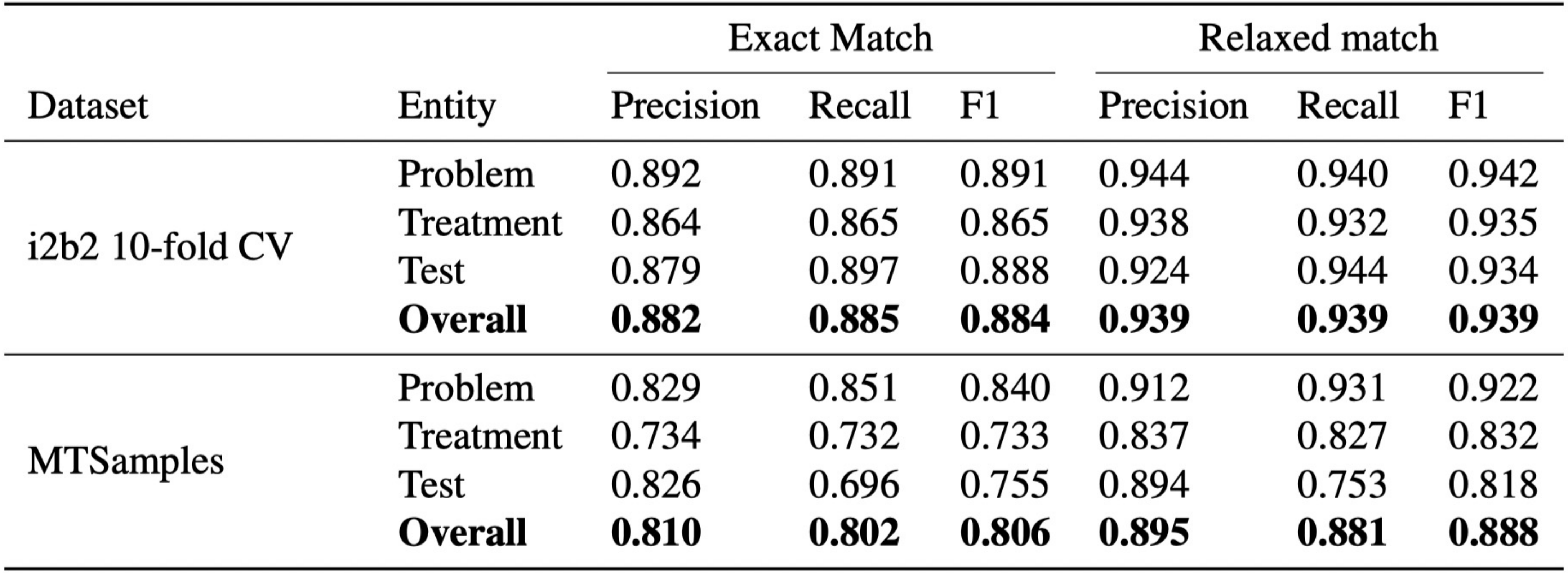
Prompt Design

- 只包含 entity types
- 添加实体类型信息



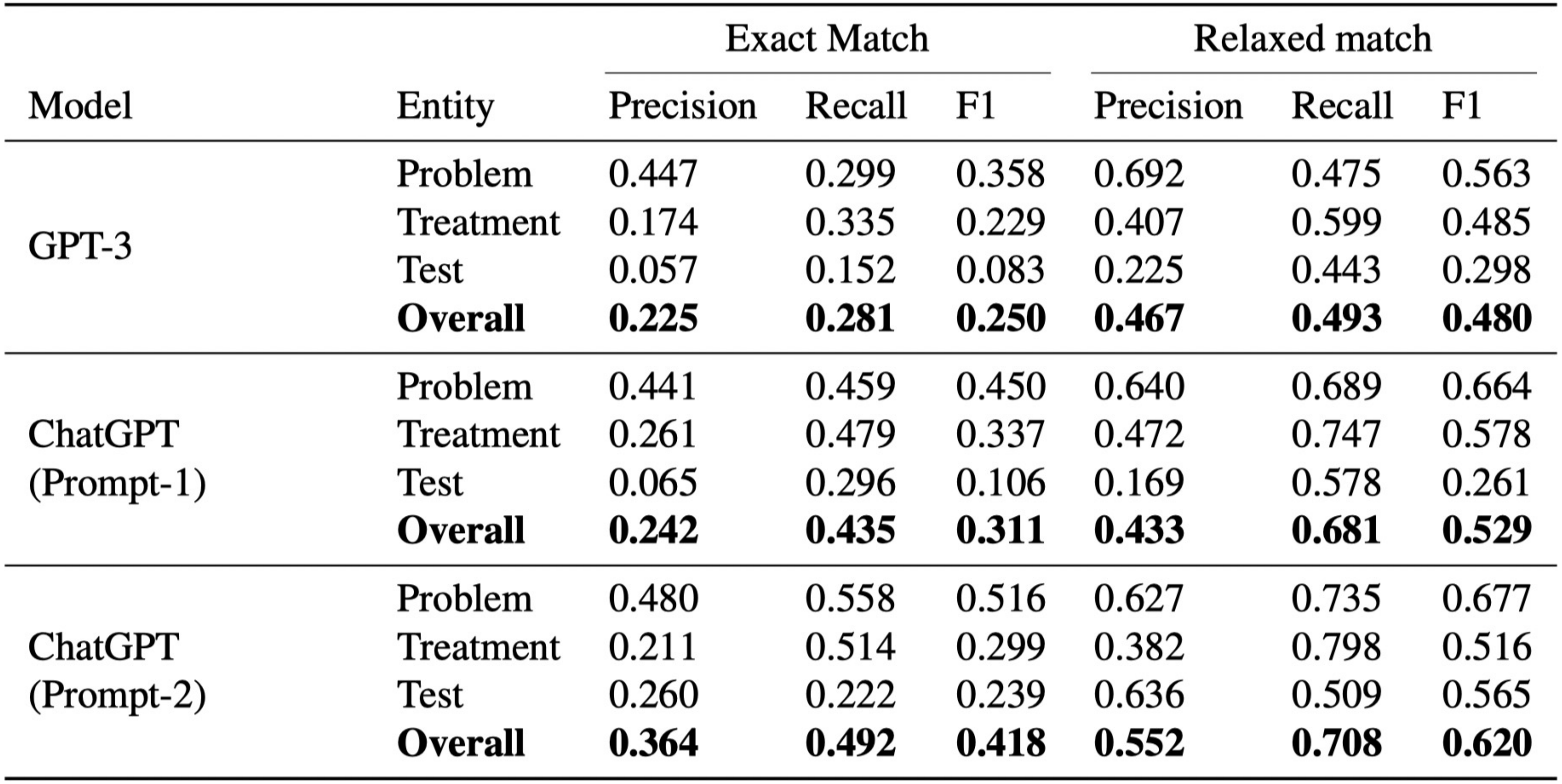
Entity	Prompt-1	Prompt-2
Medical Problem	Extract without rephrasing all medical problem entities from the following note in a list format:	Extract without rephrasing all medical condition, diagnosis, medical problem, medical symptom entities from the following note in a list format:
Treatment	Extract without rephrasing all treatment entities from the following note in a list format:	Extract without rephrasing all medical treatment, medical procedure, medical intervention, medication, drug entities from the following note in a list format:
Test	Extract without rephrasing all test entities from the following note in a list format:	Extract without rephrasing all vital signs, laboratory test, medical test, imaging study, diagnostic test entities from the following note in a list format:

Supervised learning



Dataset	Entity	Exact Match			Relaxed match		
		Precision	Recall	F1	Precision	Recall	F1
i2b2 10-fold CV	Problem	0.892	0.891	0.891	0.944	0.940	0.942
	Treatment	0.864	0.865	0.865	0.938	0.932	0.935
	Test	0.879	0.897	0.888	0.924	0.944	0.934
	Overall	**0.882**	**0.885**	**0.884**	**0.939**	**0.939**	**0.939**
MTSamples	Problem	0.829	0.851	0.840	0.912	0.931	0.922
	Treatment	0.734	0.732	0.733	0.837	0.827	0.832
	Test	0.826	0.696	0.755	0.894	0.753	0.818
	Overall	**0.810**	**0.802**	**0.806**	**0.895**	**0.881**	**0.888**

Zero-shot learning



Model	Entity	Exact Match			Relaxed match		
		Precision	Recall	F1	Precision	Recall	F1
GPT-3	Problem	0.447	0.299	0.358	0.692	0.475	0.563
	Treatment	0.174	0.335	0.229	0.407	0.599	0.485
	Test	0.057	0.152	0.083	0.225	0.443	0.298
	Overall	**0.225**	**0.281**	**0.250**	**0.467**	**0.493**	**0.480**
ChatGPT (Prompt-1)	Problem	0.441	0.459	0.450	0.640	0.689	0.664
	Treatment	0.261	0.479	0.337	0.472	0.747	0.578
	Test	0.065	0.296	0.106	0.169	0.578	0.261
	Overall	**0.242**	**0.435**	**0.311**	**0.433**	**0.681**	**0.529**
ChatGPT (Prompt-2)	Problem	0.480	0.558	0.516	0.627	0.735	0.677
	Treatment	0.211	0.514	0.299	0.382	0.798	0.516
	Test	0.260	0.222	0.239	0.636	0.509	0.565
	Overall	**0.364**	**0.492**	**0.418**	**0.552**	**0.708**	**0.620**

Discussion

- 错误主要出现在识别出一些无意义实体（FP 80.3%）
  - “Dr. X” or “Mr .ABC” 识别为 test entities
- FP中，42.45% 为类别判断错误
  - CT scan 误判为 medical problem
- 自由发挥（9.43% of total FP, 19.23% of total FN）
  - 改变措辞，导致匹配失败
  - 输出基于文本语义推测出的信息
  - 摘要文本信息
- 无法识别出同指实体
  - ”her medications” or ”her symptoms”