ChatGPT 信息抽取

- 在 Standard-IE 场景下，性能较低
- 在 OpenIE 场景下，性能优异
- 能对判断原因给出高质量、令人信服的解释
- 过度自信（预测信心与概率不符）
- 判断过程能够围绕着输入进行（忠实度较高）

Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness

评估

- 角度
  - 性能 (Performance)：在多个 IE 任务上的总体性能表现
  - 解释性 (Explainability)：能否对判断给出合理的解释
  - 偏置性 (Calibration)：预测信心与实际概率的偏离程度
  - 忠实性 (Faithfulness)：判断过程是否围绕输入
- 场景
  - Standard-IE：答案从候选标签集合中选择，即指令中包括任务描述、输入文本、模板和标签集合
  - OpenIE：没有候选标签，模型通过理解任务描述、输入文本和模板来生成预测

Keys	Explanation
Performance	
`Open`	Directly ask ChatGPT to predict the class without the label set.
`Standard`	ChatGPT's most likely correct class with a given label set.
`Top3`	The three most likely classes of the given label set from ChatGPT.
`Top5`	The five most likely classes of the given label set from ChatGPT.
`ifOpen_Correct(Manual)`	Manually annotate whether the "`Open`" is reasonable.
Explainability	
`Reason_Open`	The reason why ChatGPT chooses the class in "`Open`".
`Reason_Standard`	The reason why ChatGPT chooses the class in "`Standard`".
`ifR_Open`	Does ChatGPT think that "`Reason_Open`" is reasonable?
`ifR_Standard`	Does ChatGPT think that "`Reason_Standard`" is reasonable?
`ifR_Open(Manual)`	Manually annotate whether the "`Reason_Open`" is reasonable.
`ifR_Standard(Manual)`	Manually annotate whether the "`Reason_Standard`" is reasonable.
Calibration	
`Confidence_Open`	The confidence of ChatGPT in predicting "`Open`".
`Confidence_Standard`	The confidence of ChatGPT in predicting "`Standard`".
Faithfulness	
`FicR_Open(Manual)`	Manually annotate whether the "`Reason_Open`" is fictitious.
`FicR_Standard(Manual)`	Manually annotate whether the "`Reason_Standard`" is fictitious.

Input of Event Detection (ED)

Task Description: Given an input list of words, identify all triggers in the list, and categorize each of them into the predefined set of event types. A trigger is the main word that most clearly expresses the occurrence of an event in the predefined set of event types.

Pre-defined Label Set: The predefined set of event types includes: [Life.Be-Born, Life.Marry, Life.Divorce, Life.Injure, Life.Die, Movement.Transport, Transaction.Transfer-Ownership, Transaction.Transfer-Money, Business.Start-Org, Business.Merge-Org, Business.Declare-Bankruptcy, Business.End-Org, Conflict.Attack, Conflict.Demonstrate, Contact.Meet, Contact. Phone-Write, Personnel.Start-Position, Personnel.End-Position, Personnel.Nominate, Personnel. Elect, Justice.Arrest-Jail, Justice.Release-Parole, Justice.Trial-Hearing, Justice.Charge-Indict, Justice.Sue, Justice.Convict, Justice.Sentence, Justice.Fine, Justice.Execute, Justice.Extradite, Justice.Acquit, Justice.Appeal, Justice.Pardon].

Input and Task Requirement: Perform ED task for the following input list, and print the output: ['Putin', 'concluded', 'his', 'two', 'days', 'of', 'talks', 'in', 'Saint', 'Petersburg', 'with', 'Jacques', 'Chirac', 'of', 'France', 'and', 'German', 'Chancellor', 'Gerhard', 'Schroeder', 'on', 'Saturday', 'still', 'urging', 'for', 'a', 'central', 'role', 'for', 'the', 'United', 'Nations', 'in', 'a', 'post', '-', 'war', 'revival', 'of', 'Iraq', '.'] The output of ED task should be a list of dictionaries following json format. Each dictionary corresponds to the occurrence of an event in the input list and should consists of "trigger", "word_index", "event_type", "top3_event_type", "top5_event_type", "confidence", "if_context_dependent", "reason" and "if_reasonable" nine keys. The value of "word_ index" key is an integer indicating the index (start from zero) of the "trigger" in the input list. The value of "confidence" key is an integer ranging from 0 to 100, indicating how confident you are that the "trigger" expresses the "event_type" event. The value of "if_context_dependent" key is either 0 (indicating the event semantic is primarily expressed by the trigger rather than contexts) or 1 (indicating the event semantic is primarily expressed by contexts rather than the trigger). The value of "reason" key is a string describing the reason why the "trigger" expresses the "event_type", and do not use any " mark in this string. The value of "if_reasonable" key is either 0 (indicating the reason given in the "reason" field is not reasonable) or 1 (indicating the reason given in the "reason" field is reasonable). Note that your answer should only contain the json string and nothing else.

Standard-IE 性能

Task	Dataset	BERT	RoBERTa	SOTA	ChatGPT
Entity Typing(ET)	BBN	80.3	79.8	82.2 (Zuo et al., 2022)	85.6
	OntoNotes 5.0	69.1	68.8	72.1 (Zuo et al., 2022)	73.4
Named Entity Recognition(NER)	CoNLL2003	92.8	92.4	94.6 (Wang et al., 2021)	67.2
	OntoNotes 5.0	89.2	90.9	91.9 (Ye et al., 2022)	51.1
Relation Classification(RC)	TACRED	72.7	74.6	75.6 (Li et al., 2022a)	20.3
	SemEval2010	89.1	89.8	91.3 (Zhao et al., 2021)	42.5
Relation Extraction(RE)	ACE05-R	87.5 | 63.7	88.2 | 65.1	91.1 | 73.0 (Ye et al., 2022)	40.5 | 4.5
	SciERC	65.4 | 43.0	63.6 | 42.0	69.9 | 53.2 (Ye et al., 2022)	25.9 | 5.5
Event Detection(ED)	ACE05-E	71.8	72.9	75.8 (Liu et al., 2022a)	17.1
	ACE05-E+	72.4	72.1	72.8 (Lin et al., 2020)	15.5
Event Argument Extraction(EAE)	ACE05-E	65.3	68.0	73.5 (Hsu et al., 2022)	28.9
	ACE05-E+	64.0	66.5	73.0 (Hsu et al., 2022)	30.9
Event Extraction(EE)	ACE05-E	71.8 | 51.0	72.9 | 51.9	74.7 | 56.8 (Lin et al., 2020)	17.0 | 7.3
	ACE05-E+	72.4 | 52.7	72.1 | 53.4	71.7 | 56.8 (Hsu et al., 2022)	16.6 | 7.8

OpenIE 性能

	Standard-IE	OpenIE
BBN(*ET*)	86.8%	97.2%
CoNLL(*NER*)	69.0%	93.3%
SemEval2010(*RC*)	43.3%	84.3%
ACE05-R(*RE*)	14.9%	23.9%
ACE05-E(*ED*)	12.4%	42.6%
ACE05-E(*EAE*)	17.3%	65.3%
ACE05-E(*EE*)	4.9%	28.8%

top-k 召回率

	top-1	*top-3*	*top-5*
BBN	85.6%	92.7%	94.9% (+9.3%)
SemEval2010	42.5%	62.1%	76.0% (+33.5%)

解释性 & 忠实性

	Stardand Setting			OpenIE Setting		
	Self-check	**Human-check**	**Overlap**	**Self-check**	**Human-check**	**Overlap**
BBN (*ET*)	100.0%	99.2%	99.2%	100.0%	99.5%	99.5%
CoNLL (*NER*)	100.0%	99.3%	99.3%	100.0%	99.7%	99.7%
SemEval (*RC*)	100.0%	100.0%	100.0%	100.0%	99.7%	99.7%
ACE05-R (*RE*)	100.0%	90.0%	90.0%	100.0%	100.0%	100.0%
ACE05-E (*ED*)	100.0%	96.3%	96.3%	100.0%	90.2%	90.2%
ACE05-E (*EAE*)	100.0%	74.1%	74.1%	100.0%	90.4%	90.4%
ACE05-E (*EE*)	100.0%	47.1%	47.1%	94.0%	78.0%	74.0%

	Stardand-IE	**OpenIE**
BBN(*ET*)	98.3%	99.3%
CoNLL(*NER*)	100.0%	98.7%
SemEval(*RC*)	100.0%	99.1%
ACE05-R(*RE*)	90.0%	93.8%
ACE05-E(*ED*)	100.0%	100.0%
ACE05-E(*EAE*)	100.0%	96.5%
ACE05-E(*EE*)	100.0%	97.0%

决策过程围绕着输入原始文本展开

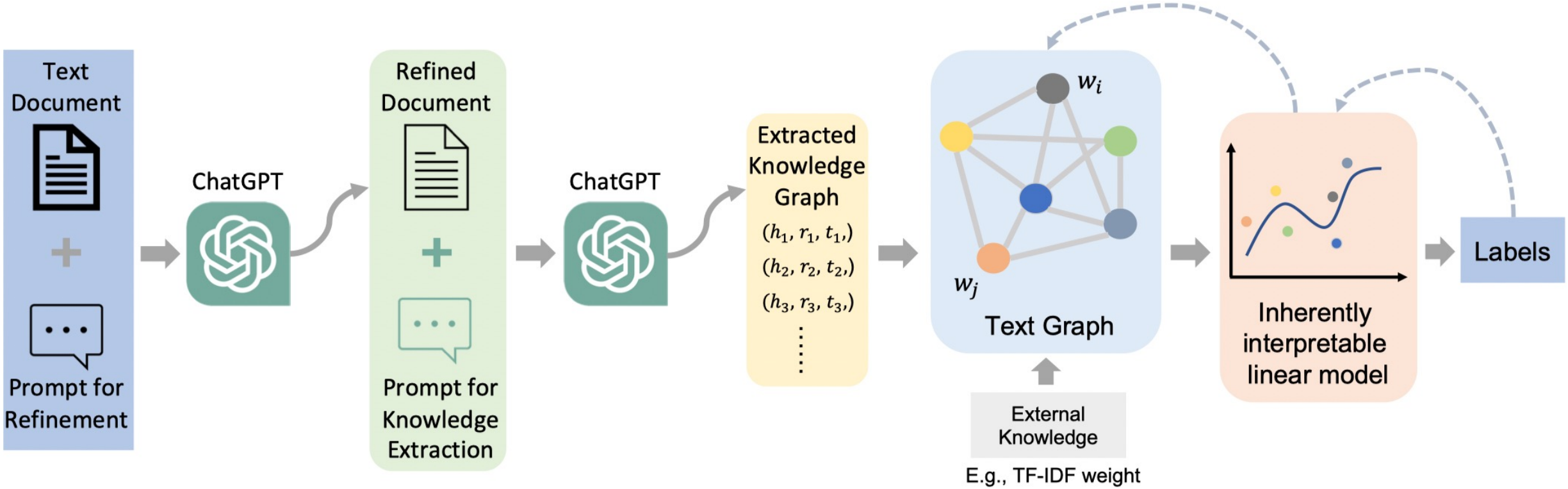
偏置性

	Correct Confidence			Incorrect Confidence		
	BERT	**RoBERTa**	**ChatGPT**	**BERT**	**RoBERTa**	**ChatGPT**
BBN*(ET)*	0.971	0.968	0.888	0.904	0.885	0.828
CoNLL*(NER)*	0.990	0.991	0.864	0.866	0.886	0.785
SemEval*(RC)*	0.983	0.989	0.868	0.871	0.852	0.839
ACE05-R*(RE)*	0.995	0.991	0.760	0.883	0.810	0.764
ACE05-E*(ED)*	0.882	0.944	0.852	0.770	0.871	0.737
ACE05-E*(EAE)*	0.762	0.785	0.956	0.525	0.555	0.910
ACE05-E*(EE)*	0.763	0.782	0.845	0.612	0.628	0.764

	BERT	**RoBERTa**	**ChatGPT**
BBN*(ET)*	0.012	0.012	0.026
CoNLL*(NER)*	0.052	0.044	0.204
SemEval*(RC)*	0.023	0.031	0.460
ACE05-R*(RE)*	0.020	0.014	0.745
ACE05-E*(ED)*	0.161	0.226	0.656
ACE05-E*(EAE)*	0.154	0.168	0.699
ACE05-E*(EE)*	0.211	0.288	0.699

预测信心与真实概率不符（过度自信）

借助 ChatGPT 构建知识图谱



ChatGraph: Interpretable Text Classification by Converting ChatGPT Knowledge to Graphs

步骤一：文本精炼

```
Please generate a refined document of the
    following document. And please ensure
    that the refined document meets the
    following criteria:
1. The refined document should be abstract
    and does not change any original meaning
    of the document.
2. The refined document should retain all the
     important objects, concepts, and
    relationships between them.
3. The refined document should only contain
    information that is from the document.
4. The refined document should be readable
    and easy to understand without any
    abbreviations and misspellings.
Here is the content: [x]
```

修正语法和拼写错误
替换同义词
阐明原文的句子结构

步骤二：知识图谱抽取

```
You are a knowledge graph extractor, and your
     task is to extract and return a
    knowledge graph from a given text.Let's
    extract it step by step:
(1). Identify the entities in the text. An
    entity can be a noun or a noun phrase
    that refers to a real-world object or an
    abstract concept. You can use a named
    entity recognition (NER) tool or a part-
    of-speech (POS) tagger to identify the
    entities.
(2). Identify the relationships between the
    entities. A relationship can be a verb or
     a prepositional phrase that connects two
     entities. You can use dependency parsing
     to identify the relationships.
(3). Summarize each entity and relation as
    short as possible and remove any stop
    words.
(4). Only return the knowledge graph in the
    triplet format: ('head entity', 'relation
    ', 'tail entity').
(5). Most importantly, if you cannot find any
     knowledge, please just output: "None".
Here is the content: [x]
```

Chain Of Thoughts

模板由循序渐进的指令组成

性能

Method	Training Data	20NG	R8	R52	Ohsumed
TF-IDF+LR	Full data	$83.19_{\pm 0.00}$	$93.74_{\pm 0.00}$	$86.95_{\pm 0.00}$	$54.66_{\pm 0.00}$
TextGCN (1 layer)	Full data	$78.85_{\pm 0.10}$	$86.74_{\pm 0.10}$	$73.86_{\pm 0.11}$	$50.25_{\pm 0.08}$
TextGCN (2 layers)	Full data	$86.34_{\pm 0.09}$	$97.07_{\pm 0.10}$	$93.56_{\pm 0.18}$	$68.36_{\pm 0.56}$
ChatGPT	0-shot	$58.70_{\pm 0.00}$	$60.10_{\pm 0.00}$	$75.23_{\pm 0.00}$	$39.93_{\pm 0.00}$
	2-shot	$58.44_{\pm 0.00}$	$72.54_{\pm 0.00}$	$81.68_{\pm 0.00}$	$47.05_{\pm 0.00}$
	5-shot	– [5]	$82.43_{\pm 0.00}$	$90.13_{\pm 0.00}$	$45.39_{\pm 0.00}$
ChatGraph	Full data	$79.15_{\pm 0.08}$	$96.39_{\pm 0.34}$	$92.14_{\pm 0.26}$	$60.79_{\pm 0.14}$
ChatGraph (with TF-IDF)	Full data	$79.68_{\pm 0.37}$	$96.46_{\pm 0.31}$	$93.25_{\pm 0.32}$	$63.63_{\pm 0.33}$