

语言学视角下的 语言模型知识诊断与可解释性探索

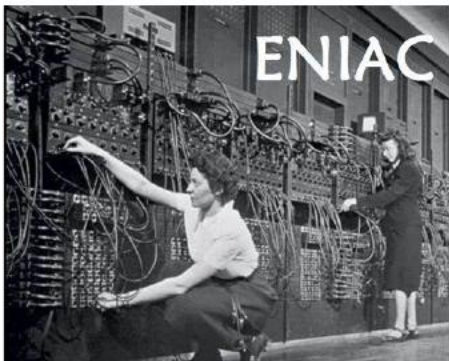
(语言学与AI还能一起愉快地玩耍吗?)

胡韧奋 (北京师范大学)

irishu@bnu.edu.cn

2024年12月4日



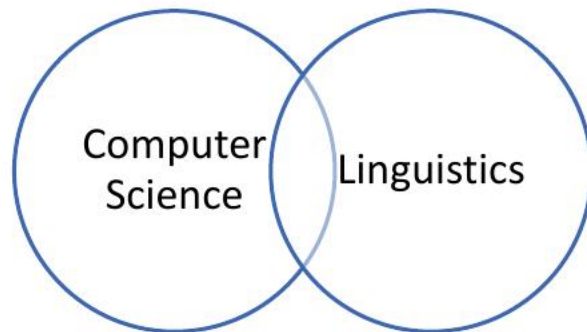


ENIAC (1946)

机器能够理解人类语言，并做出恰当反应吗？



Computational Linguistics
Natural Language Processing

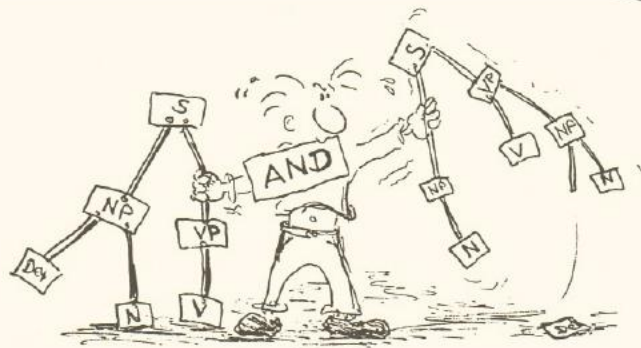


Annual Meeting of the Association for Computational Linguistics (ACL)

1963-2024



神经语言模型能力强、参数量大，它究竟学习到了什么？我们可否利用已有的语言学知识对它进行评估和诊断？

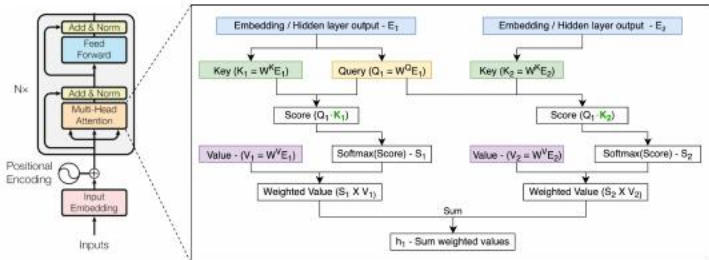


Linguistic Knowledge

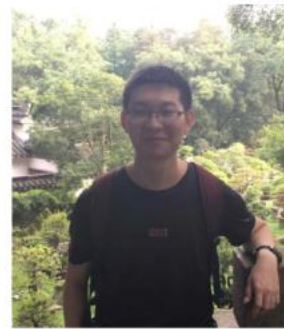
模型诊断评估

模型可解释性

Neural Language Models



如果神经网络模型有效编码了语言学知识，我们是否可以利用它来更好地开展语言研究？



闫真

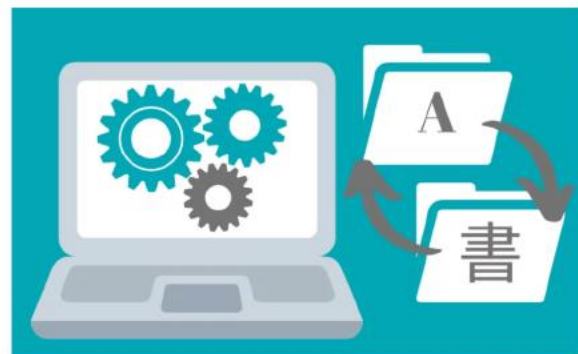


王予沛

语言模型的诞生

例1. 我很喜欢吃四川_____。

例2. 实施摆在眼前，他还是不肯相信。



$$P(S) = P(w_1, w_2, w_3, \dots, w_n)$$

统计语言模型

神经语言模型

统计语言模型

- 借助大规模语料库中的词语共现信息进行统计分析，实现句子概率估计：

$$P(S) = P(w_1) P(w_2|w_1) P(w_3|w_1w_2) \dots P(w_n|w_{1:n-1})$$

- 简化起见，划定窗口大小为N，如N=2时：

$$P(w_k|w_{1:k-1}) \approx P(w_k|w_{k-1})$$

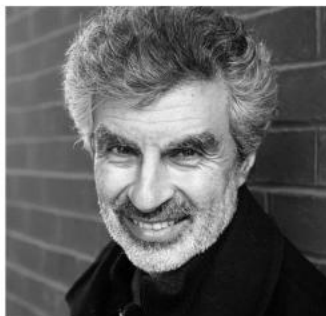
统计语言模型

小猫走进了房间



狗狗跑去了卧室





神经语言模型 (Bengio et al., 2003)

- 利用大规模语料库和神经网络模型，令模型学习预测下一个词，即给定 w_1, w_2, \dots, w_k ，预测 w_{k+1}

$$\hat{y} = P(w_i | w_{1:k}) = LM(w_{1:k}) = \text{softmax}(\mathbf{h} \mathbf{W}^2 + \mathbf{b}^2)$$

$$\mathbf{h} = g(\mathbf{x} \mathbf{W}^1 + \mathbf{b}^1)$$

$$\mathbf{x} = [v(w_1); v(w_2); \dots; v(w_k)]$$

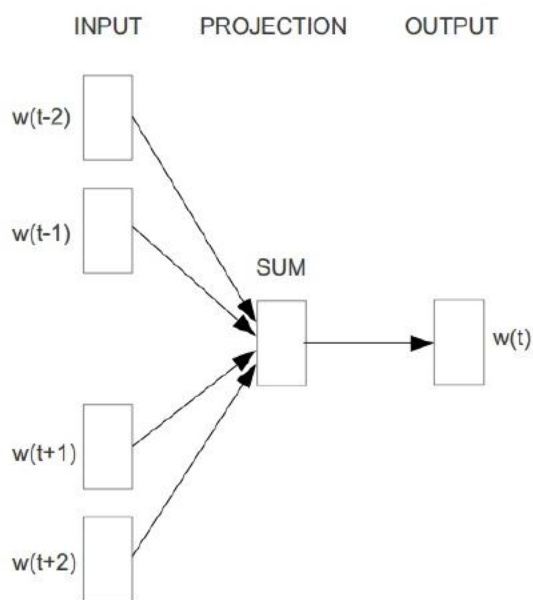
$$v(w) = \mathbf{E}_{[w]}$$

$$w_i \in V \quad \mathbf{E} \in \mathbb{R}^{|V| \times d_w} \quad \mathbf{W}^1 \in \mathbb{R}^{k \cdot d_w \times d_{\text{hid}}} \quad \mathbf{b}^1 \in \mathbb{R}^{d_{\text{hid}}} \quad \mathbf{W}^2 \in \mathbb{R}^{d_{\text{hid}} \times |V|} \quad \mathbf{b}^2 \in \mathbb{R}^{|V|}.$$

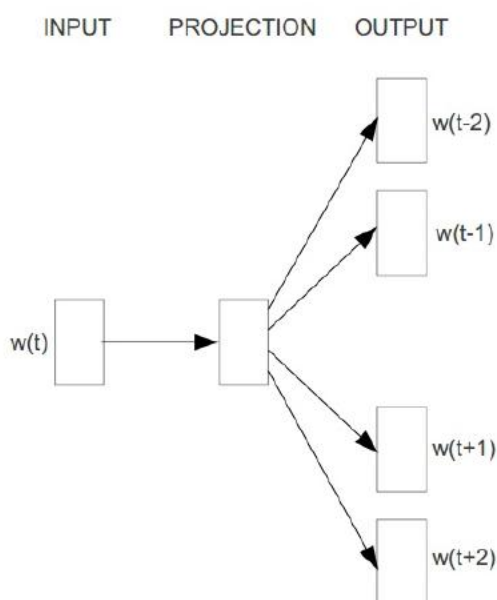
Word2vec (Mikolov et al., 2013)



- **CBOW**: 给定 $w_{k-2}, w_{k-1}, w_{k+1}, w_{k+2}$, 预测 w_k
- **Skip-gram**: 给定 w_k , 预测 $w_{k-2}, w_{k-1}, w_{k+1}, w_{k+2}$



CBOW (a)



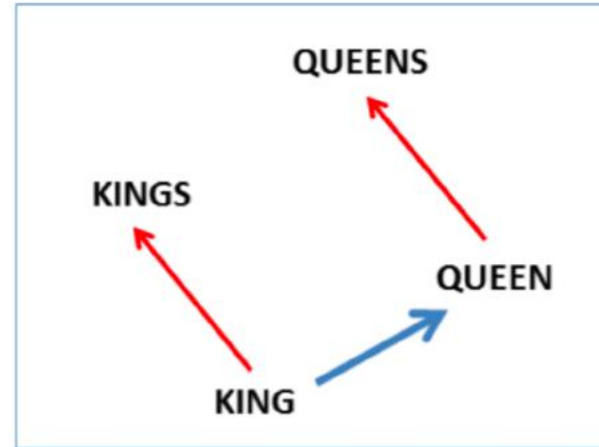
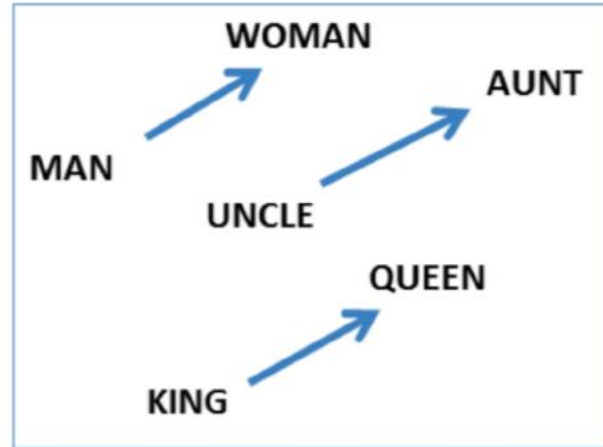
Skip-gram (b)

Linguistic Hypothesis

- Harris (1954): Words that occur in the same contexts tend to have similar meanings.
- Firth (1957): You shall know a word by the company it keeps.

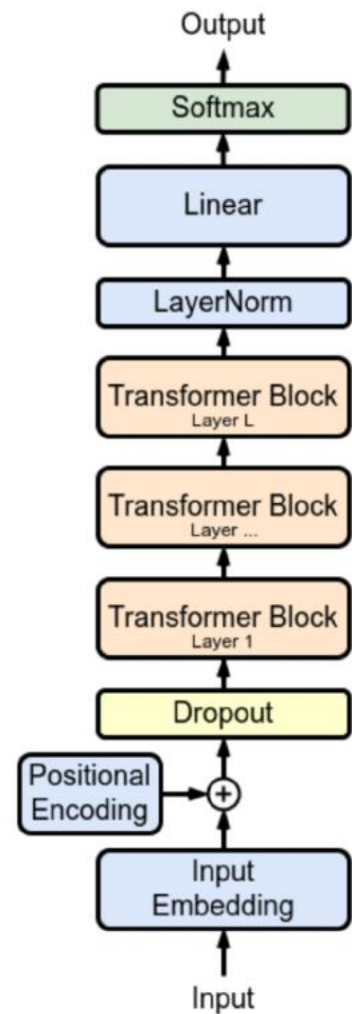
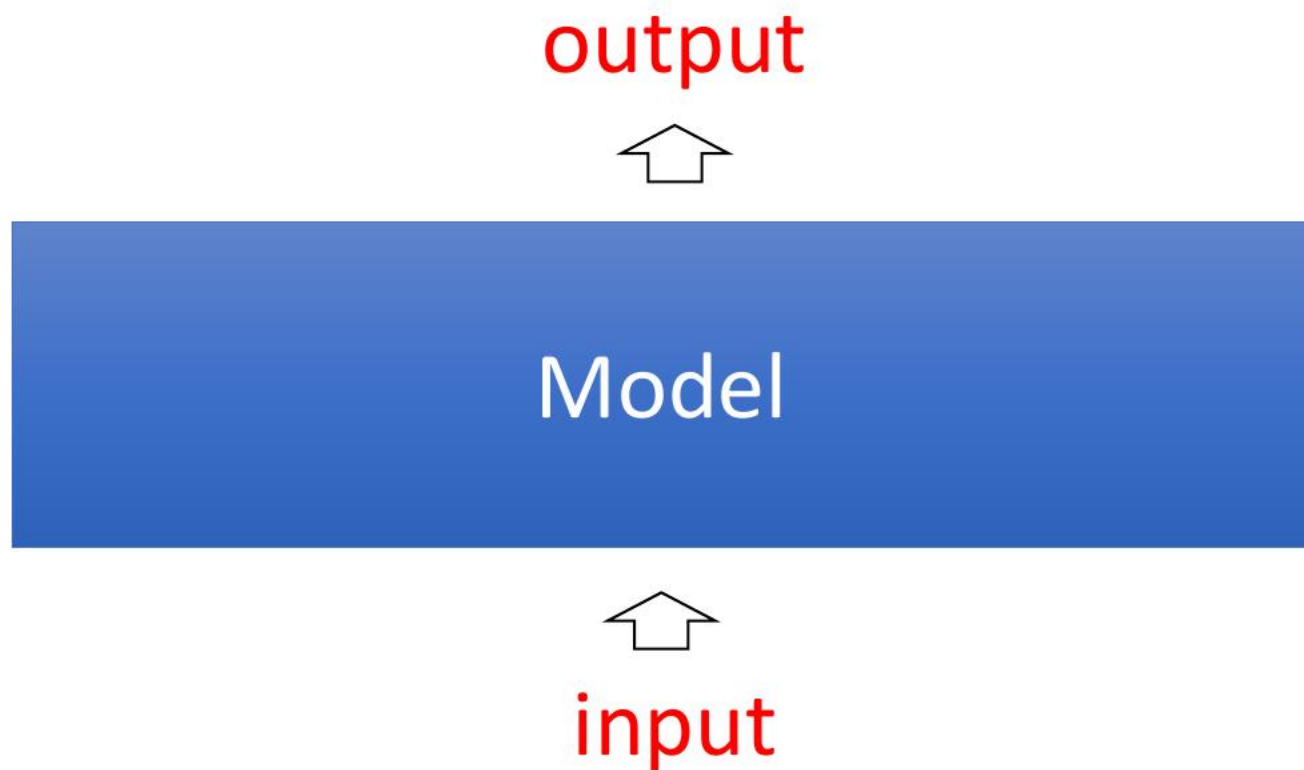
300D

- computer [0.2287 -0.0280 ... 0.2431 -0.26317 -0.1120]
- school [0.0394 -0.28124 ... 0.17091 0.22821 -0.074532]
- computing [0.1426 0.0074 ... 0.2210 -0.2819 -0.1143]



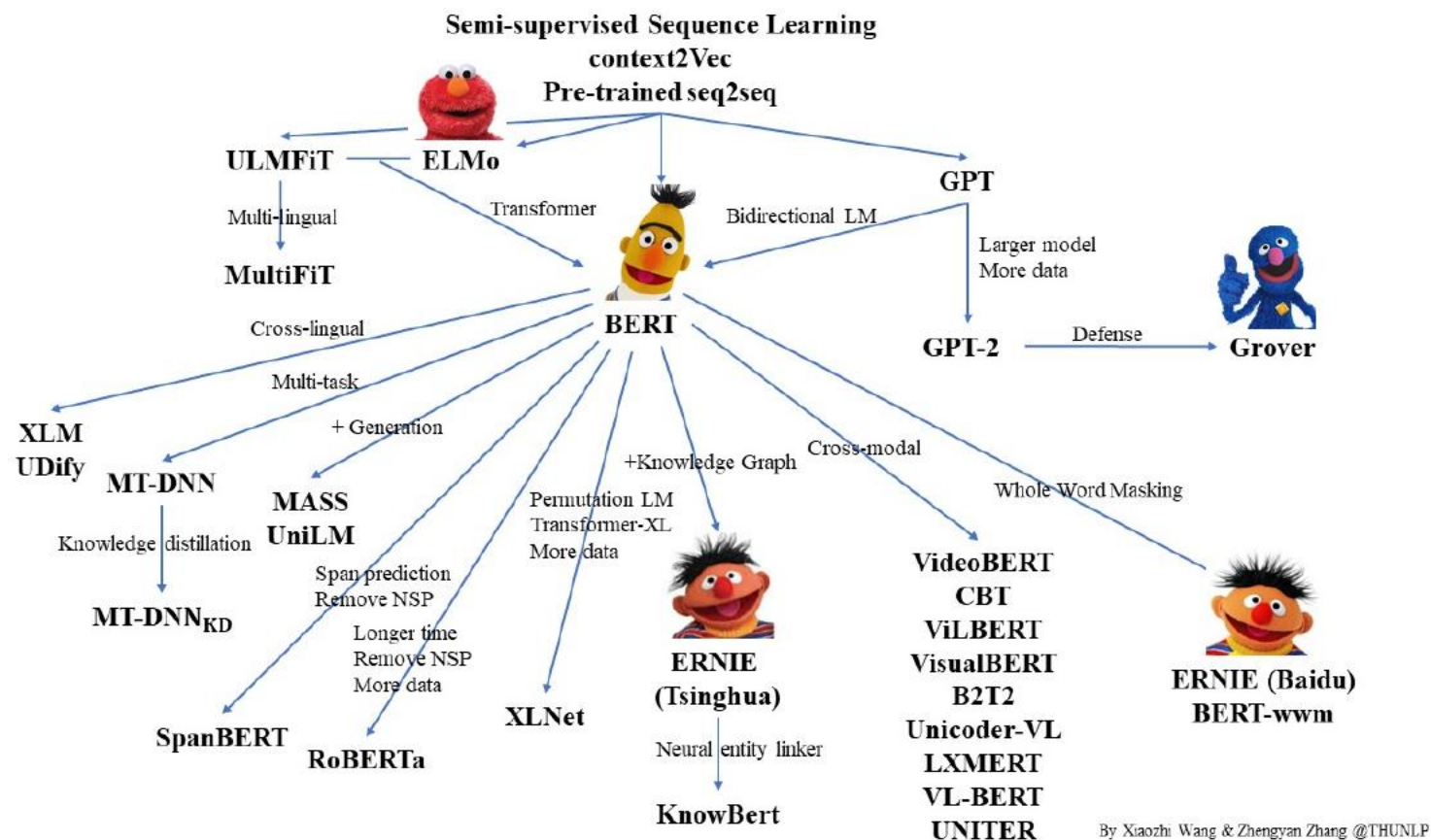
- **Semantic analogy**
 - man – woman \approx uncle – aunt \approx king - queen
- **Grammatical analogy**
 - kings – king \approx queens – queen

神经语言模型 (2003-2024)



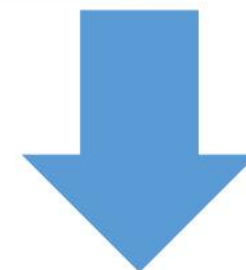
Transformers (Vaswani et al., 2017)

预训练语言模型 (Pre-trained Language Models)



以BERT系列为代表

双向预训练
语言模型



单向预训练
语言模型

以GPT系列为代表

<https://github.com/thunlp/PLMpapers>

Xiaozhi Wang and Zhengyan Zhang @THUNLP

BERT (Devlin et al., 2018)

- 训练语料: BooksCorpus + Wikipedia
- 训练任务:
 - Masked LM (randomly select 15% of the words)
 - Next sentence prediction (negative sampling)

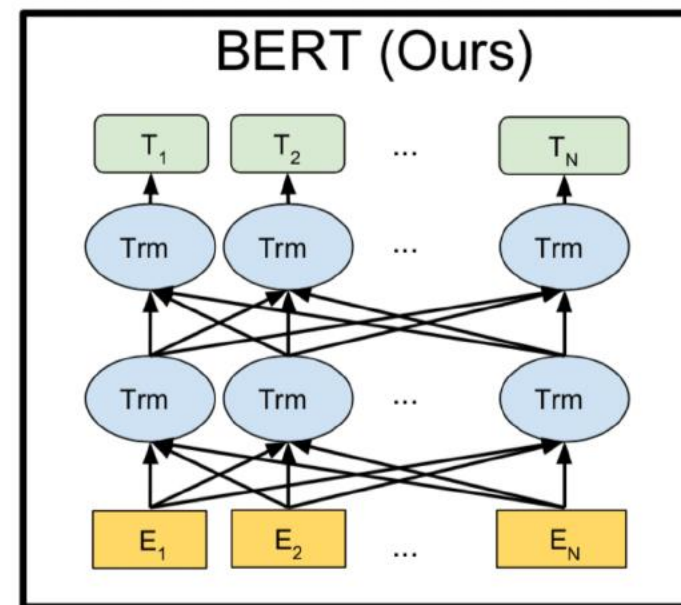
- 80% of the time: Replace the word with the [MASK] token, e.g., my dog is hairy → my dog is [MASK]
- 10% of the time: Replace the word with a random word, e.g., my dog is hairy → my dog is apple
- 10% of the time: Keep the word unchanged, e.g., my dog is hairy → my dog is hairy. The purpose of this is to bias the representation towards the actual observed word.

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]

Label = IsNext

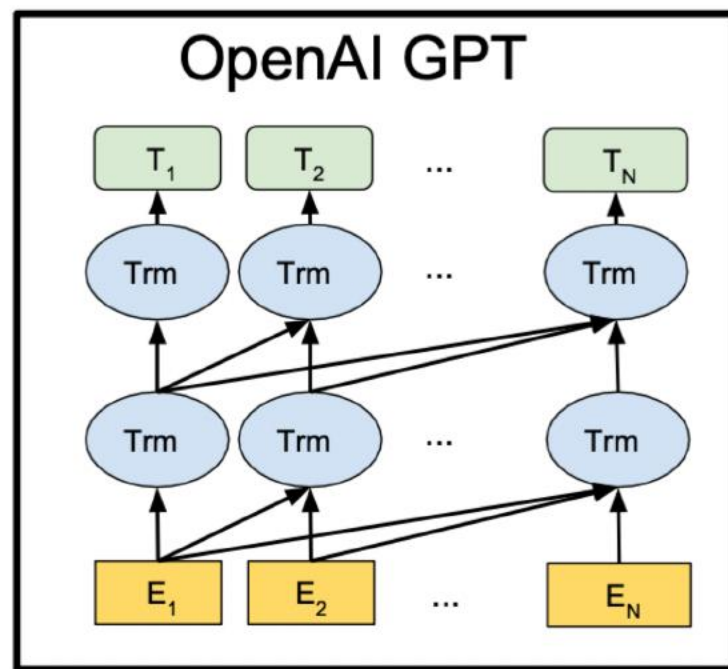
Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]

Label = NotNext



Generative Pre-trained Transformers (Radford et al., 2018)

- 预训练：预测下一个词
 - 学？
 - 学而？
 - 学而时？
 - 学而时习？
 - 学而时习之？
 - 学而时习之， ？
 - 学而时习之， 不？
 - 学而时习之， 不亦？
 - 学而时习之， 不亦说？
 - 学而时习之， 不亦说乎？



预训练语言模型引发的问题

- 预训练+微调效果好，但模型结构复杂、参数量大、训练和决策过程不透明。



评估基准：主要看模型微调效果



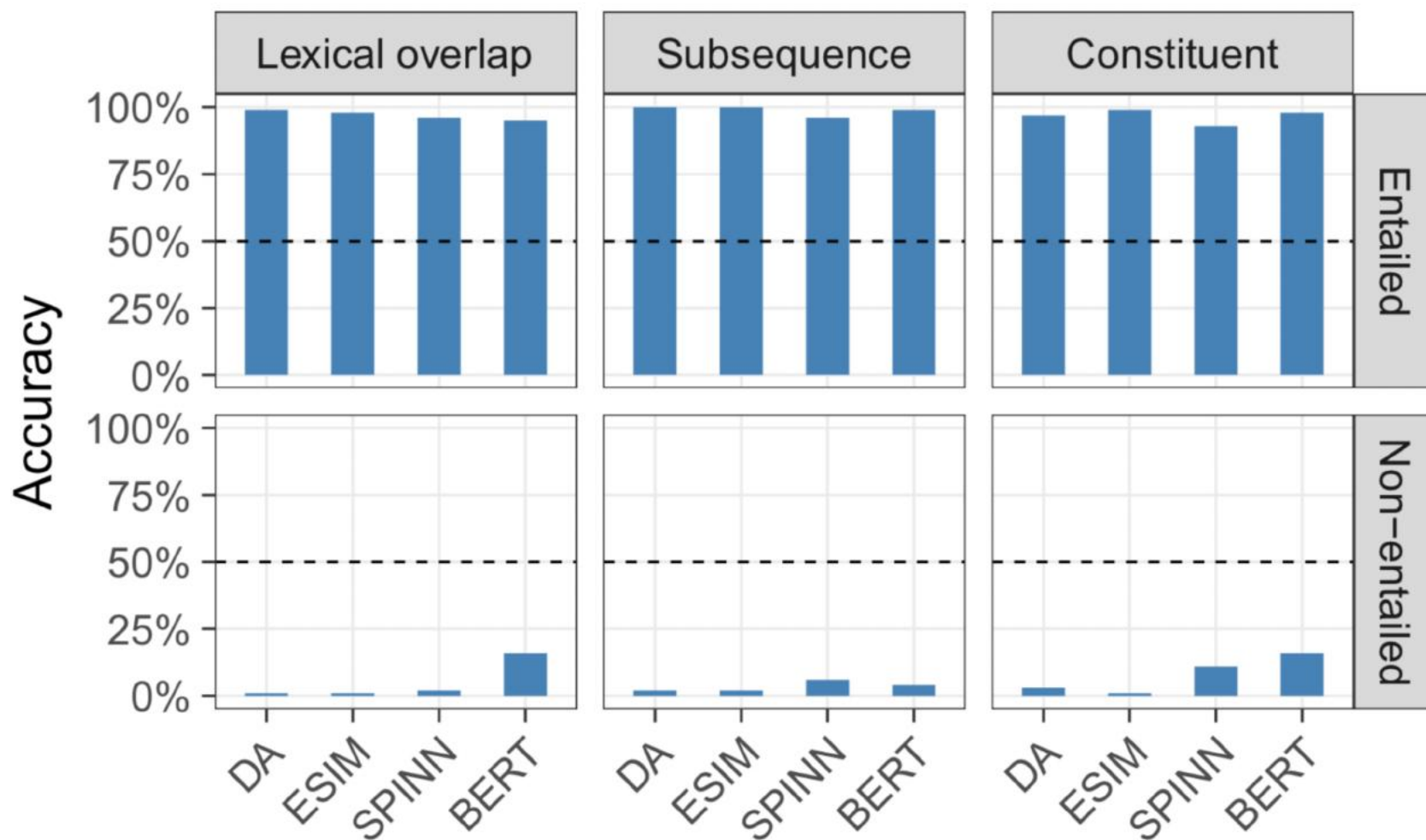
- 模型可能依赖浅层启发式规则进行预测，虽然在各项任务中表现优异，但未必真正理解语言。

Right for the Wrong Reasons (McCoy et al., 2019)

- Statistical learners such as standard neural network architectures are prone to adopting **shallow heuristics** that succeed for the majority of training examples, instead of learning the underlying generalizations that they are intended to capture.

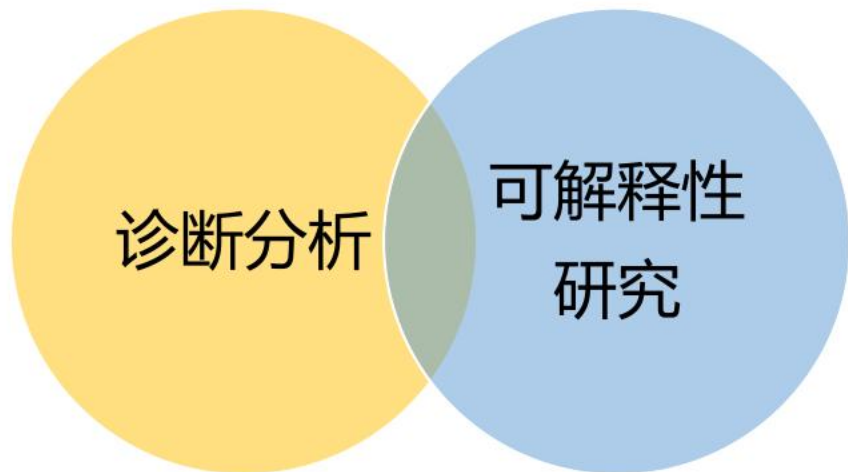
Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	The doctor was paid by the actor . —————→ The doctor paid the actor. WRONG
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near the actor danced . —————→ The actor danced. WRONG
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If the artist slept , the actor ran. —————→ The artist slept. WRONG

Right for the Wrong Reasons [\(McCoy et al., 2019\)](#)



预训练语言模型引发的问题

- 预训练+微调效果好，但模型结构复杂、参数量大、训练和决策过程不透明。
 - 模型在预训练和微调阶段分别学习到了什么知识？
 - 模型如何从输入文本中提取特征，并根据这些特征做出预测？
 - 模型的决策过程能否被人类理解和信任？



构建可信可控模型(高利害性场景)

改进优化模型(知识漏洞偏见/风险)

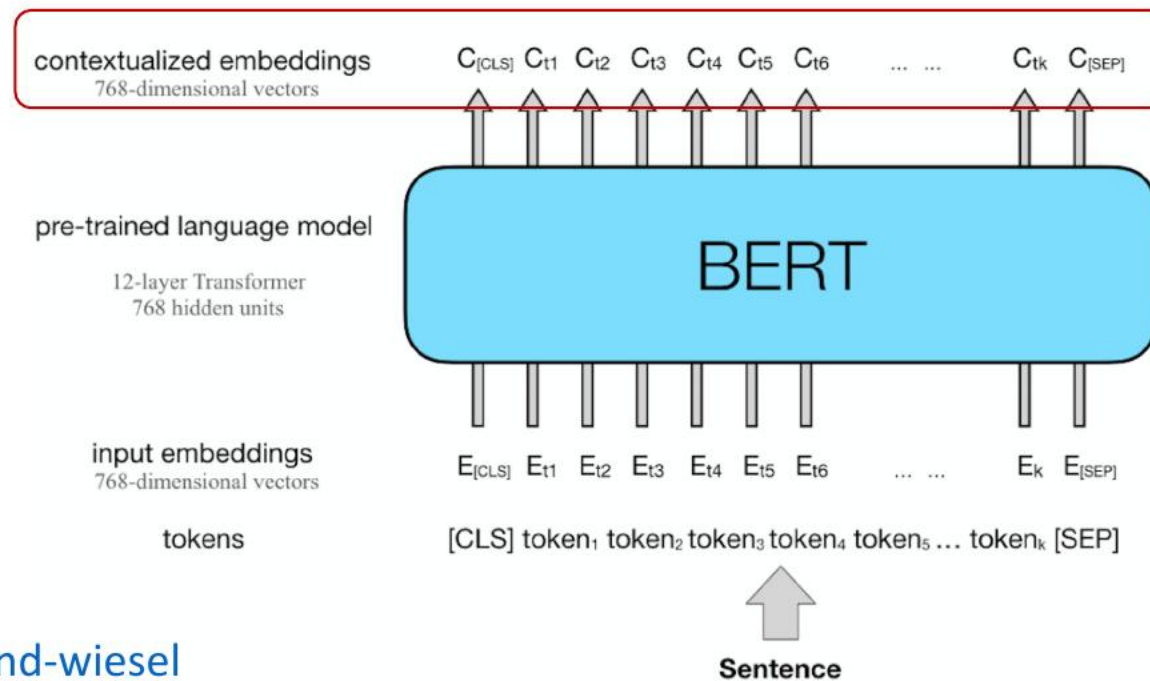
理论建构(底层原理/训练动力学)

“探针” (Probing) 任务：从神经科学中借鉴诊断分析手段



Neural Probes in Understanding Brain Activity

<http://braintour.harvard.edu/archives/portfolio-items/hubel-and-wiesel>



探针任务 (Conneau et al., 2018)

*Probing task: a **classification** problem that focuses on simple **linguistic properties** of sentences.*

- Surface information
 - SentLen: 6-way classification
 - WC(word content): 1000-way classification
- Syntactic information
 - The bigram shift (BShift): binary classification
 - TreeDepth: 8-way classification (5-12)
 - TopConst: 20-way classification
- Semantic
 - Tense classification
 - SubjNum classification
 - ObjNum classification
 - SOMO (randomly replace a word)
 - Coordination inversion

BShift, SOMO and CoordInv manipulations can accidentally generate acceptable sentences.

Jawahar et al. (2019) What does BERT learn about the structure of language?

(1) 10 probing tasks created by Conneau et al. (2018)

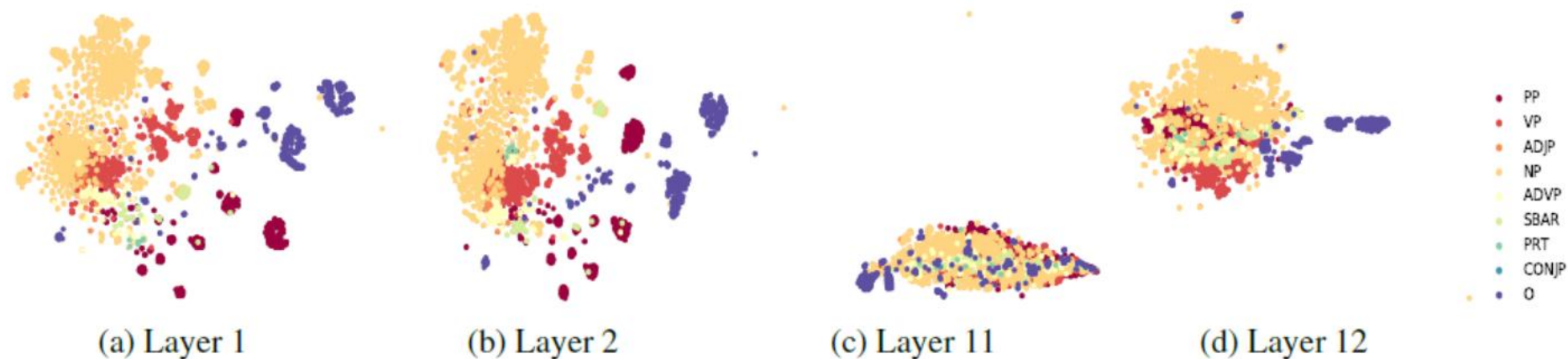
Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	96.2 (3.9)	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	69.8 (69.6)	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	41.3 (13.0)	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	88.1 (21.9)	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	84.1 (39.5)	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	82.2 (21.1)	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	87.0 (37.1)	90.0 (28.0)	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	78.7 (28.9)
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	65.2 (15.3)	74.9 (25.4)

Table 2: Probing task performance for each BERT layer. The value within the parentheses corresponds to the difference in performance of trained vs. untrained BERT.

BERT embeds a rich hierarchy of linguistic signals: surface information at the **bottom**, syntactic information in the **middle**, semantic information at the **top**.

Jawahar et al. (2019) What does BERT learn about the structure of language?

(2) phrasal syntax with span embeddings



BERT mostly captures phrase-level information in the **lower layers** and that this information gets gradually diluted in higher layers.

Hewitt and Manning (2019). A Structural Probe for Finding Syntax in Word Representations

- whether syntax trees are embedded in a linear transformation of a neural network's word representation space?

distance between two nodes

the depth of each node

embedding?

squared L2 distance & the distance between words in the parse tree
squared L2 norm & the depth in the parse tree.

(ignore the direction of the relation)

Hewitt and Manning (2019). A Structural Probe for Finding Syntax in Word Representations

Method	Distance		Depth	
	UUAS	DSpr.	Root%	NSpr.
LINEAR	48.9	0.58	2.9	0.27
ELMo0	26.8	0.44	54.3	0.56
DECAY0	51.7	0.61	54.3	0.56
PROJ0	59.8	0.73	64.4	0.75
ELMo1	77.0	0.83	86.5	0.87
BERTBASE7	79.8	0.85	88.0	0.87
BERTLARGE15	82.5	0.86	89.4	0.88
BERTLARGE16	81.7	0.87	90.1	0.89

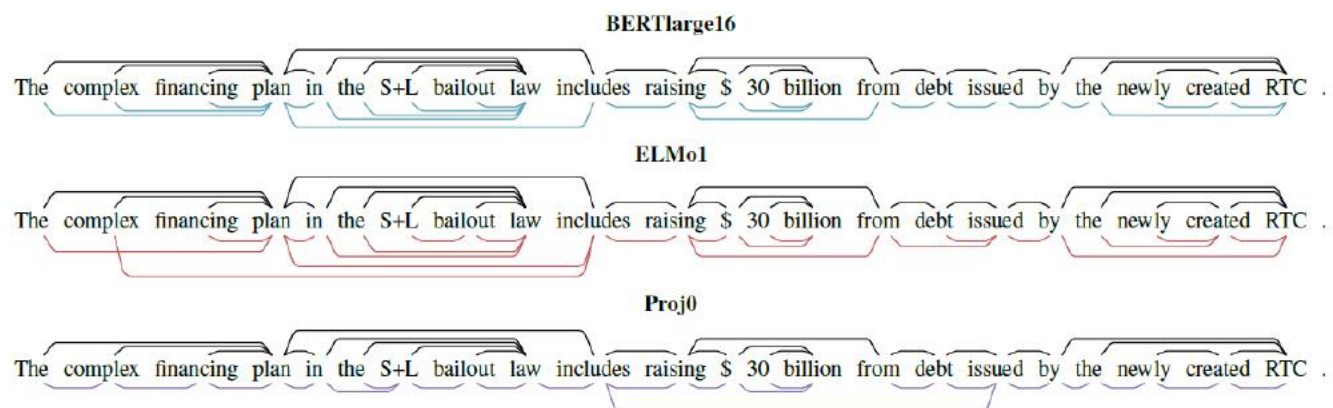
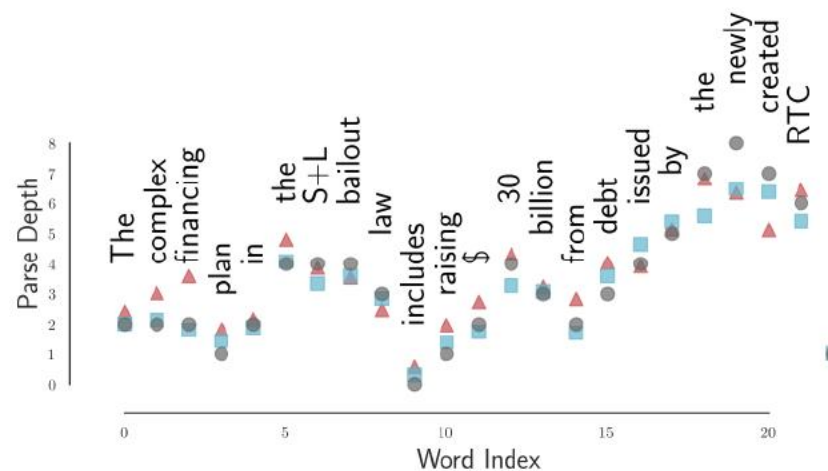
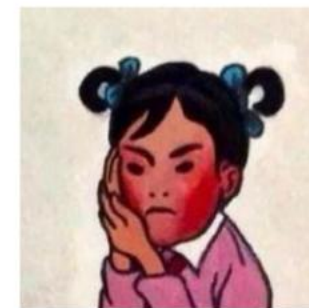
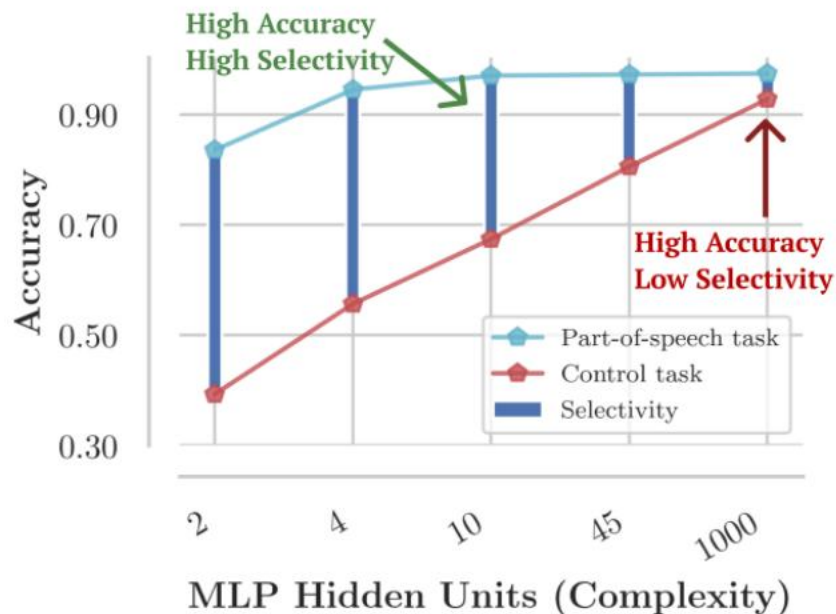


Figure 2: Minimum spanning trees resultant from predicted squared distances on BERTLARGE16 and ELMo1 compared to the best baseline, PROJ0. Black edges are the gold parse, above each sentence; blue are BERTLARGE16, red are ELMo1, and purple are PROJ0.

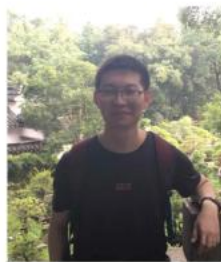
Hewwit and Liang (2019). Designing and Interpreting Probes with Control Tasks



- 预训练语言模型在语探针任务上取得了很好的表现，但这些探针任务依赖进一步训练**额外的探针分类器**。究竟是预训练模型中编码了语言学知识，还是探针分类器学习到了知识？



Control tasks associate word types with random outputs.



闫真

“开箱即用”的探针任务

1. 构建探针任务及数据集，**无需更新模型和训练分类器**，即可实现知识评测，探究：预训练语言模型本身是否具备语法知识和语义知识？模型是否学习到了事实知识，以及掌握常识推理能力？
2. 对比**微调前后**的模型知识差异：是否出现“**被动遗忘**”？
3. 扰动模型，令其“**主动遗忘**”知识，是否影响其他任务表现？

“开箱即用”的探针任务

- 根据预训练语言模型“本能”设计探针任务，无需额外训练
- 覆盖语法、语义、事实知识、常识推理四个维度。

Question: Mike [MASK] the toy car yesterday.

Choices: A: break. B: breaks. C: broke. D: broken. E: breaking.

(a) syntactic

a: He deposited 500 dollars in the *bank*.

↕ similarity (a, b)

b: The *bank* will accommodate him with a loan. similarity (a, c)

↕ similarity (b, c)

c: The river broke its *bank* during the flood.

(b) semantic

Elvis Presley was born in *Tupelo*.

Head entity Relation Tail entity



Elvis Presley was born in [MASK].

(c) relation

The trophy doesn't fit into the brown suitcase because [MASK] is too large.

Choices: A: trophy. B: suitcase.

(d) reasoning

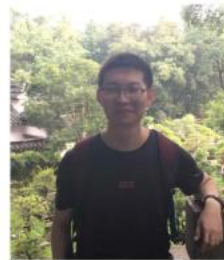
“开箱即用”的探针任务

利用模型“本能”设计任务

完形填空 + 语境向量相似度

任务	问题	候选项/答案
语法	By 1980, they had two [MASK] , and Leavell's career was soaring.	daughter / -s
	But her body 's long, long in the mold, and her [MASK] soul in Heaven.	bless / -ed / -ing
	She [MASK] the screen door, and I stepped inside.	open / -ed / -ing / -s
语义 - 相似度	a: There was an urgent <i>appeal</i> for blood donors to treat victims of the crash.	similarity(a, b)
	b: Yesterday, villagers added their voice to the <i>appeal</i> for urgent action.	similarity(b, c)
	c: Our site is now for 14-year-old, we're going to <i>appeal</i> directly to their interests.	similarity(a, c)
语义-义项排序	<i>Sentence</i> : He is too hard up to buy [MASK] clothes. <i>Target</i> : fancy	
	<i>Sense a</i> : luxurious, expensive, lavish, decorative, elaborate	Rank(a)
	<i>Sense b</i> : imagination, creativity, conception, insight, invention	Rank(b)
事实	<i>To Build a Fire</i> is a short story by [MASK] author Jack London.	American
推理	The trophy doesn't fit into the brown suitcase because [MASK] is too large.	trophy / suitcase
	The trophy doesn't fit into the brown suitcase because [MASK] is too small.	trophy / suitcase

“开箱即用”的探针任务

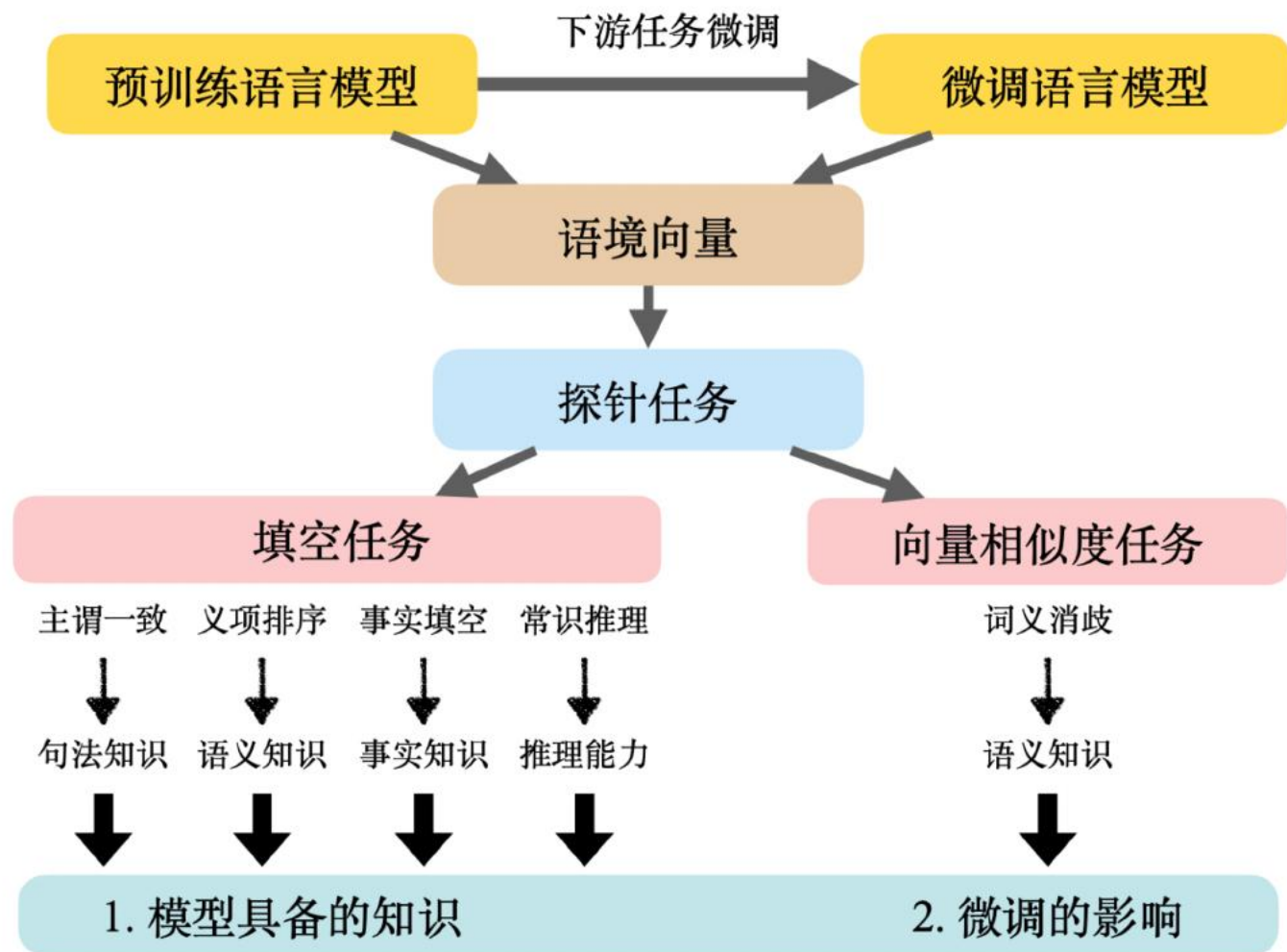


闫真

知识类型	任务	形式	候选项个数	数据规模	数据来源
语法	单词形态预测	多选	2, 3, 4	31,860	COHA
语义	义项排序	多选	N/A	4,036	牛津同义词词典
语义	词义相似度	多选	3	35,735	牛津词典
事实	实体预测	填空	N/A	12,932	KBP37, FewRel
推理	推理	多选	2	3,248	Winograd

实验设计

- BERT
- RoBERTa
- DistilBERT



微调数据集

- CoLA: 语法可接受性
- SST-2: 情感分类
- MRPC: 句对语义等价
- STS-B: 句对语义相似性
- MNLI: 句对语义蕴含
- SQuAD2.0: 阅读理解

实验结果 (填空任务)

		微调前	在各项任务微调后的表现					微调后准确率下降均值	
		Baseline	CoLA	MNLI	MRPC	SQuAD	SST-2	STS-B	Avg
BERT	Syntactic	90.53	-0.37	-3.06	-0.68	-2.90	-2.23	-0.75	-1.67
	Semantic-Mean	69.87	-0.25	-1.39	-0.92	-2.30	-2.13	0.13	-1.14
	Semantic-LogMean	74.40	-0.19	-1.38	-0.81	-2.37	-1.90	0.63	-1.01
	Fact-Fewrel	51.09	-2.52	-6.04	-2.75	-17.19	-3.18	-3.03	-5.79
	Fact-KBP37	39.80	-2.41	-6.61	-3.45	-18.93	-2.11	-4.53	-6.34
	Reasoning	52.00	0.86	0.49	0.67	-0.31	0.43	-0.37	0.30
	Avg	62.95	-0.81	-3.00	-1.32	-7.33	-1.85	-1.32	
RoBERTa	Syntactic	90.55	-0.90	-14.23	-6.37	-9.90	-9.50	-8.42	-8.22
	Semantic-Mean	66.87	-2.72	-13.23	-6.24	-12.19	-7.88	-10.73	-8.83
	Semantic-LogMean	74.08	-2.97	-17.07	-7.53	-14.34	-10.38	-13.55	-10.97
	Fact-Fewrel	46.37	-8.18	-26.23	-18.04	-34.15	-12.15	-33.00	-21.96
	Fact-KBP37	39.72	-6.11	-22.72	-17.75	-28.63	-12.84	-29.41	-19.58
	Reasoning	52.30	-0.27	-0.86	-1.29	-1.38	-0.12	-1.47	-0.90
	Avg	61.65	-3.53	-15.72	-9.54	-16.77	-8.81	-16.10	
DistilBERT	Syntactic	91.03	-1.68	-6.15	-4.58	-11.88	-3.08	-4.21	-5.26
	Semantic-Mean	69.20	-1.31	-2.85	-2.55	-6.59	-2.40	-1.83	-2.92
	Semantic-LogMean	75.12	-1.88	-4.48	-3.09	-8.20	-3.94	-3.74	-4.22
	Fact-Fewrel	47.47	-0.80	-8.50	-4.35	-28.89	-0.52	-8.60	-8.61
	Fact-KBP37	36.91	-1.92	-11.21	-7.33	-30.99	-1.13	-10.16	-10.46
	Reasoning	51.75	0.55	0.03	-0.59	-0.59	0.12	-0.22	-0.12
	Avg	61.91	-1.17	-5.53	-3.75	-14.52	-1.82	-4.79	

实验结果（填空任务）

		Baseline
BERT	Syntactic	90.53
	Semantic-Mean	69.87
	Semantic-LogMean	74.40
	Fact-Fewrel	51.09
	Fact-KBP37	39.80
	Reasoning	52.00
	Avg	62.95
RoBERTa	Syntactic	90.55
	Semantic-Mean	66.87
	Semantic-LogMean	74.08
	Fact-Fewrel	46.37
	Fact-KBP37	39.72
	Reasoning	52.30
	Avg	61.65
DistilBERT	Syntactic	91.03
	Semantic-Mean	69.20
	Semantic-LogMean	75.12
	Fact-Fewrel	47.47
	Fact-KBP37	36.91
	Reasoning	51.75
	Avg	61.91

微调前：

- 语法知识 > 90%
- 语义知识 > 65%
- 事实知识 35%-52%
- 推理能力 50%左右

三个模型都可以较好地编码语法和语义知识、部分事实型知识，但不具备推理能力。

实验结果（填空任务）

		Baseline	CoLA	MNLI	MRPC	SQuAD	SST-2	STS-B	Avg
BERT	Syntactic	90.53	-0.37	-3.06	-0.68	-2.90	-2.23	-0.75	-1.67
	Semantic-Mean	69.87	-0.25	-1.39	-0.92	-2.30	-2.13	0.13	-1.14
	Semantic-LogMean	74.40	-0.19					0.63	-1.01
	Fact-Fewrel	51.09	-2.52					-3.03	-5.79
	Fact-KBP37	39.80	-2.41					-4.53	-6.34
	Reasoning	52.00	0.86					-0.37	0.30
	Avg	62.95	-0.81					-1.32	
RoBERTa	Syntactic	90.55	-0.90					-8.42	-8.22
	Semantic-Mean	66.87	-2.72					-10.73	-8.83
	Semantic-LogMean	74.08	-2.97					-13.55	-10.97
	Fact-Fewrel	46.37	-8.18					-33.00	-21.96
	Fact-KBP37	39.72	-6.11					-29.41	-19.58
	Reasoning	52.30	-0.27					-1.47	-0.90
	Avg	61.65	-3.53					-16.10	
DistilBERT	Syntactic	91.03	-1.68					-4.21	-5.26
	Semantic-Mean	69.20	-1.31					-1.83	-2.92
	Semantic-LogMean	75.12	-1.88					-3.74	-4.22
	Fact-Fewrel	47.47	-0.80					-8.60	-8.61
	Fact-KBP37	36.91	-1.92					-10.16	-10.46
	Reasoning	51.75	0.55	0.03	-0.59	-0.59	0.12	-0.22	-0.12
	Avg	61.91	-1.17	-5.53	-3.75	-14.52	-1.82	-4.79	

微调后都出现了一定程度的遗忘:

语法知识下降 5.56%

语义知识
- 平均排名下降6.26%
- 平均对数排名7.24%

事实知识
- FewRel下降 25.08%
- KBP37 下降 31.24%

模型差异

	Baseline	CoLA	MNLI	MRPC	SQuAD	SST-2	STS-B	Avg	
BERT	Syntactic	90.53	-0.37	-3.06	-0.68	-2.90	-2.23	-0.75	-1.67
	Semantic-Mean	69.87	-0.25	-1.39	-0.92	-2.30	-2.13	0.13	-1.14
	Semantic-LogMean	74.40	-0.19	-1.38	-0.81	-2.37	-1.90	0.63	-1.01
	Fact-Fewrel	51.09	-2.52	-6.04	-2.75	-17.19	-3.18	-3.03	-5.79
	Fact-KBP37	39.80	-2.41	-6.61	-3.45	-18.93	-2.11	-4.53	-6.34
	Reasoning	52.00	0.86	0.49	0.67	-0.31	0.43	-0.37	0.30
	Avg	62.95	-0.81	-3.00	-1.32	-7.33	-1.85	-1.32	
RoBERTa	Syntactic	90.55	-0.90	-14.23	-6.37	-9.90	-9.50	-8.42	-8.22
	Semantic-Mean	66.87	-2.72	-13.23	-6.24	-12.19	-7.88	-10.73	-8.83
	Semantic-LogMean	74.08	-2.97	-17.07	-7.53	-14.34	-10.38	-13.55	-10.97
	Fact-Fewrel	46.37	-8.18	-26.23	-18.04	-34.15	-12.15	-33.00	-21.96
	Fact-KBP37	39.72	-6.11	-22.72	-17.75	-28.63	-12.84	-29.41	-19.58
	Reasoning	52.30	-0.27	-0.86	-1.29	-1.38	-0.12	-1.47	-0.90
	Avg	61.65	-3.53	-15.72	-9.54	-16.77	-8.81	-16.10	
DistilBERT	Syntactic	91.03	-1.68	-6.15	-4.58	-11.88	-3.08	-4.21	-5.26
	Semantic-Mean	69.20	-1.31	-2.85	-2.55	-6.59	-2.40	-1.83	-2.92
	Semantic-LogMean	75.12	-1.88	-4.48	-3.09	-8.20	-3.94	-3.74	-4.22
	Fact-Fewrel	47.47	-0.80	-8.50	-4.35	-28.89	-0.52	-8.60	-8.61
	Fact-KBP37	36.91	-1.92	-11.21	-7.33	-30.99	-1.13	-10.16	-10.46
	Reasoning	51.75	0.55	0.03	-0.59	-0.59	0.12	-0.22	-0.12
	Avg	61.91	-1.17	-5.53	-3.75	-14.52	-1.82	-4.79	

RoBERTa在探针任务上下降最多，但是在下游任务上表现却最好

下游任务差异

CoLA表现最好

SQuAD表现最差

		Baseline	CoLA	MNLI	MRPC	SQuAD	SST-2	STS-B	Avg
BERT	Syntactic	90.53	-0.37	-3.06	-0.68	-2.90	-2.23	-0.75	-1.67
	Semantic-Mean	69.87	-0.25	-1.39	-0.92	-2.30	-2.13	0.13	-1.14
	Semantic-LogMean	74.40	-0.19	-1.38	-0.81	-2.37	-1.90	0.63	-1.01
	Fact-Fewrel	51.09	-2.52	-6.04	-2.75	-17.19	-3.18	-3.03	-5.79
	Fact-KBP37	39.80	-2.41	-6.61	-3.45	-18.93	-2.11	-4.53	-6.34
	Reasoning	52.00	0.86	0.49	0.67	-0.31	0.43	-0.37	0.30
	Avg	62.95	-0.81	-3.00	-1.32	-7.33	-1.85	-1.32	
RoBERTa	Syntactic	90.55	-0.90	-14.23	-6.37	-9.90	-9.50	-8.42	-8.22
	Semantic-Mean	66.87	-2.72	-13.23	-6.24	-12.19	-7.88	-10.73	-8.83
	Semantic-LogMean	74.08	-2.97	-17.07	-7.53	-14.34	-10.38	-13.55	-10.97
	Fact-Fewrel	46.37	-8.18	-26.23	-18.04	-34.15	-12.15	-33.00	-21.96
	Fact-KBP37	39.72	-6.11	-22.72	-17.75	-28.63	-12.84	-29.41	-19.58
	Reasoning	52.30	-0.27	-0.86	-1.29	-1.38	-0.12	-1.47	-0.90
	Avg	61.65	-3.53	-15.72	-9.54	-16.77	-8.81	-16.10	
DistilBERT	Syntactic	91.03	-1.68	-6.15	-4.58	-11.88	-3.08	-4.21	-5.26
	Semantic-Mean	69.20	-1.31	-2.85	-2.55	-6.59	-2.40	-1.83	-2.92
	Semantic-LogMean	75.12	-1.88	-4.48	-3.09	-8.20	-3.94	-3.74	-4.22
	Fact-Fewrel	47.47	-0.80	-8.50	-4.35	-28.89	-0.52	-8.60	-8.61
	Fact-KBP37	36.91	-1.92	-11.21	-7.33	-30.99	-1.13	-10.16	-10.46
	Reasoning	51.75	0.55	0.03	-0.59	-0.59	0.12	-0.22	-0.12
	Avg	61.91	-1.17	-5.53	-3.75	-14.52	-1.82	-4.79	

实验结果（向量相似度任务）

		Baseline	CoLA	MNLI	MRPC	SQuAD	SST-2	STS-B	Avg
BERT	Last	82.99	-2.44	-4.41	-1.87	-9.23	-17.29	-2.53	-6.30
	Best	83.36	-0.69	-1.26	-0.85	-1.86	-0.51	-0.9	-1.01
	Average	84.18	-0.37	-1.59	-0.7	-1.78	-0.6	-0.8	-0.97
	Avg	83.51	-1.17	-2.42	-1.14	-4.29	-6.13	-1.41	
RoBERTa	Last	80.8	-0.71	-32.07	-7.54	-23.06	-29.89	-10.71	-17.33
	Best	82.91	0.61	-1.46	-0.6	-2.19	-0.33	-0.71	-0.78
	Average	84.12	0.63	-2.58	-1.01	-4.75	-3.88	-1.65	-2.21
	Avg	82.61	0.18	-12.04	-3.05	-10.00	-11.37	-4.36	
DistilBERT	Last	83.09	-4.25	-10.54	-7.77	-16.13	-12.58	-9.69	-10.16
	Best	83.09	-2.26	-5.57	-4.49	-8.22	-2.95	-4.31	-4.63
	Average	82.4	-0.07	-3.78	-1.68	-5.9	-0.41	-3.32	-2.53
	Avg	82.86	-2.19	-6.63	-4.65	-10.08	-5.31	-5.77	

微调前：

- 准确率均 > 80%

微调后：

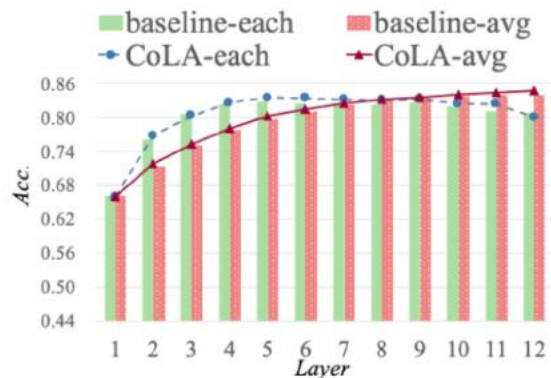
- 最后一层下降 11.26%
- 最佳层下降 2.14%
- 所有层取平均下降 1.90%



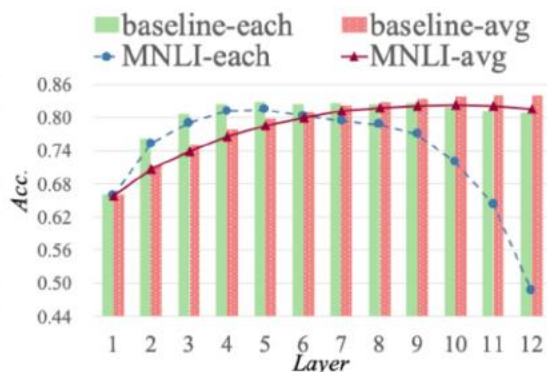
遗忘与两方面因素相关

- 不同的预训练模型
- 不同的微调任务

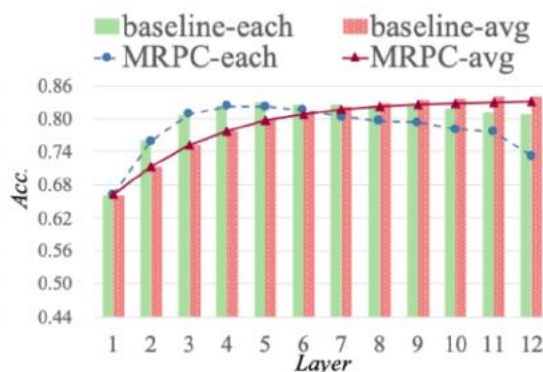
逐层分析 (向量相似度任务)



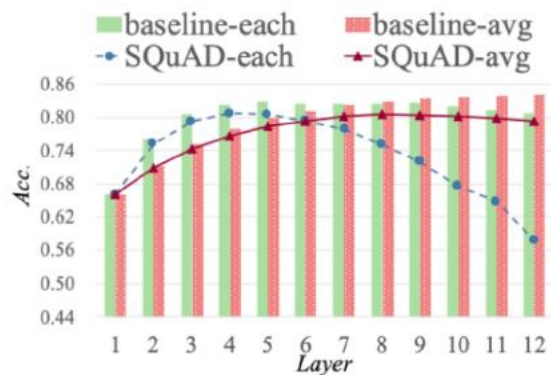
(a) baseline vs. CoLA



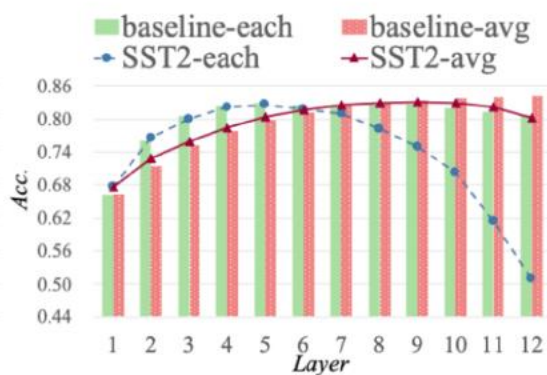
(b) baseline vs. MNLI



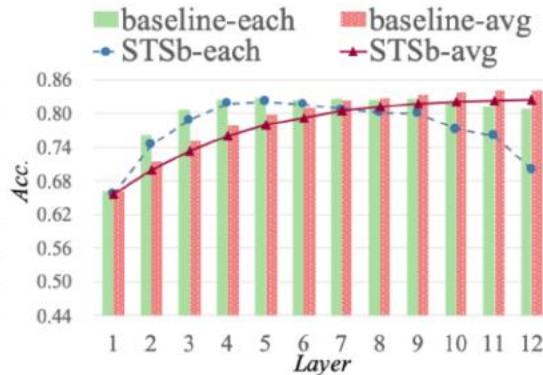
(c) baseline vs. MRPC



(d) baseline vs. SQuAD



(e) baseline vs. SST-2



(f) baseline vs. STS-b

微调前:

- 每一层的准确率先上升后下降
- 逐层平均准确率持续上升

微调后:

- 每一层的准确率仍然是先上升后下降，但是最后一层的准确率骤降。
- 逐层平均的准确率仍然持续上升

添加“扰动”，实现可控遗忘

- 扰动是否会影响模型在特定知识上的编码能力？
- 特定知识遗忘后，模型对其他类型知识的编码以及在下游任务的表现是否会受到影响？

语法扰动

She [MASK] the screen door, and I stepped inside.

*open, **opened**, opening, opens*



***open**, opened, opening, opens*

语义扰动

- *There was an urgent appeal for blood donors to treat victims of the crash.*
- *Yesterday, villagers added their voice to the appeal for urgent action.*

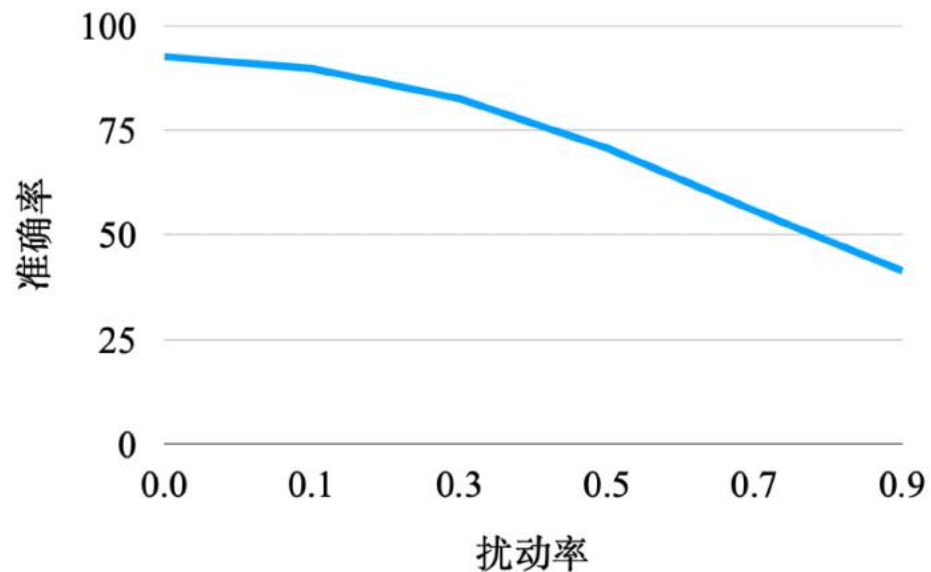
标签 1 (义项相同)



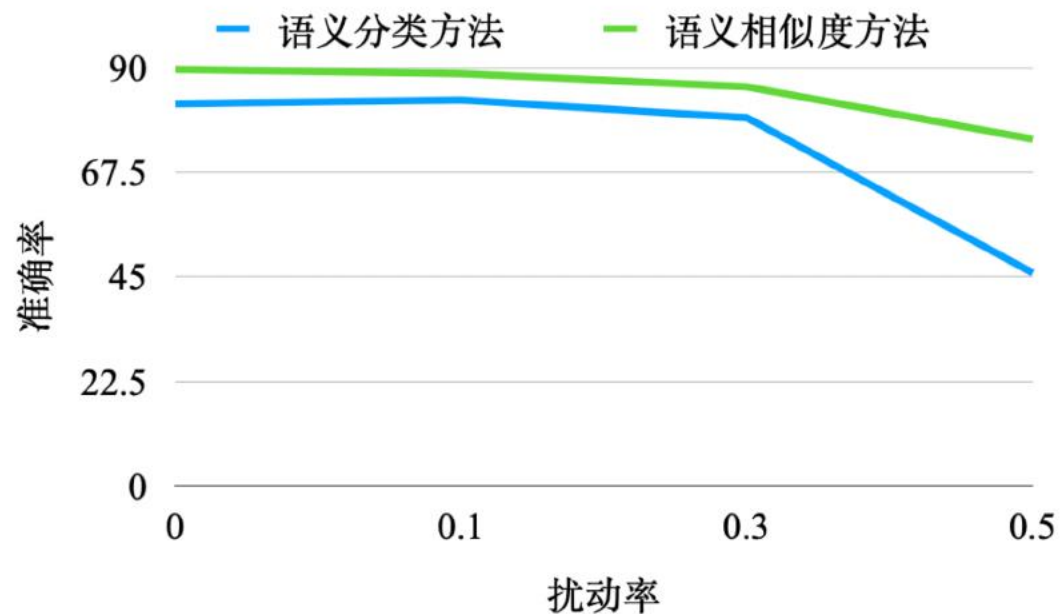
标签 0 (义项不同)

扰动影响

加入扰动后，语法任务准确率变化情况



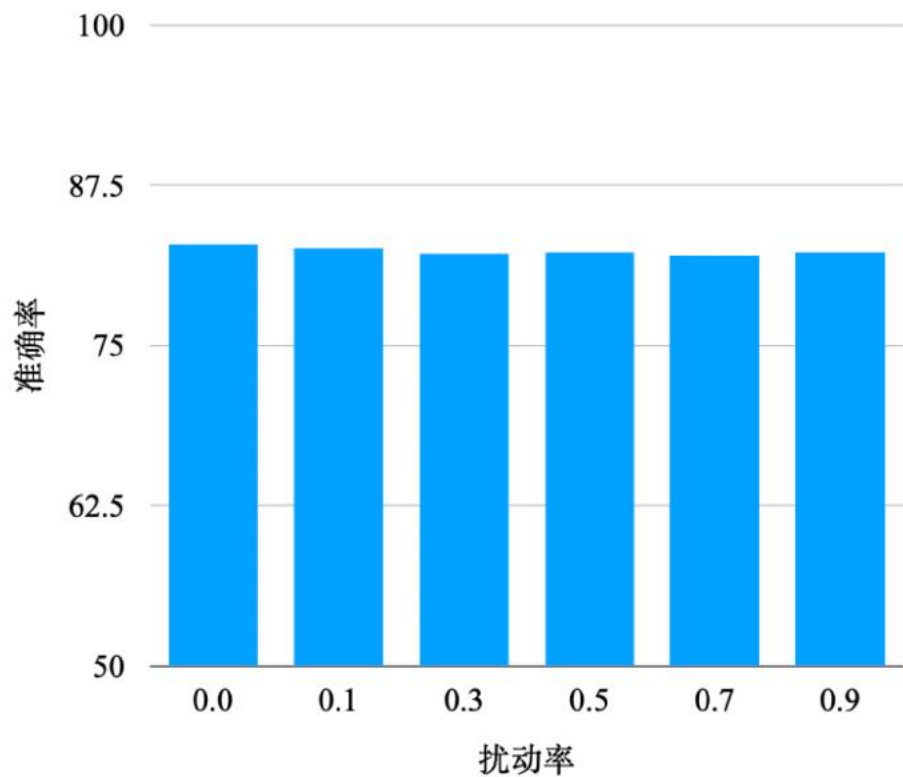
加入扰动后，语义任务准确率变化情况



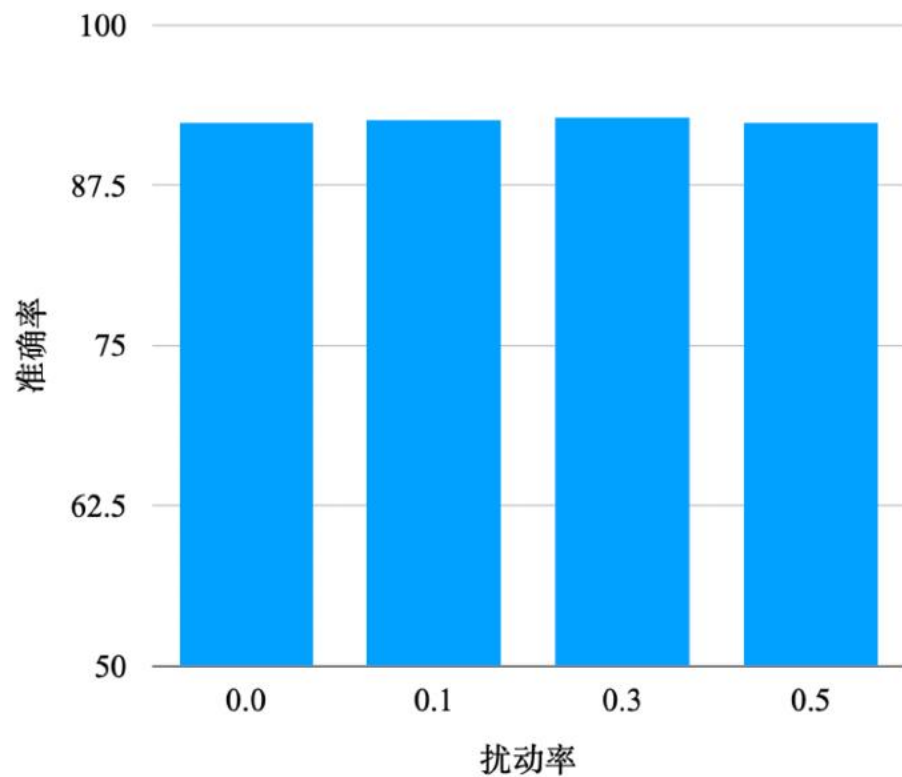
扰动影响

这两个指定维度的句法和语义知识在模型编码中相对独立

语法知识被破坏后
模型在语义任务上的表现

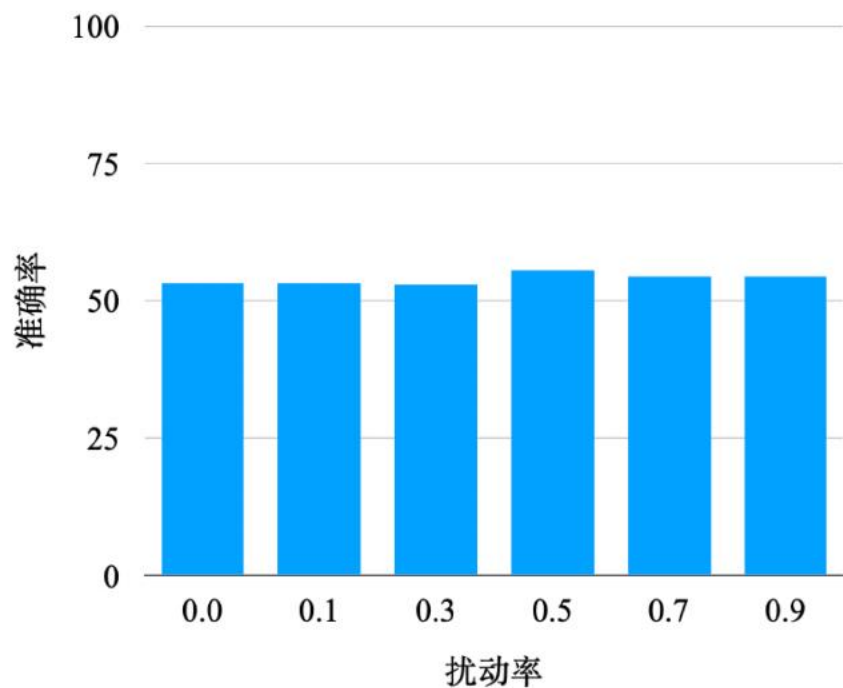


语义知识被破坏后
模型在句法任务上的表现

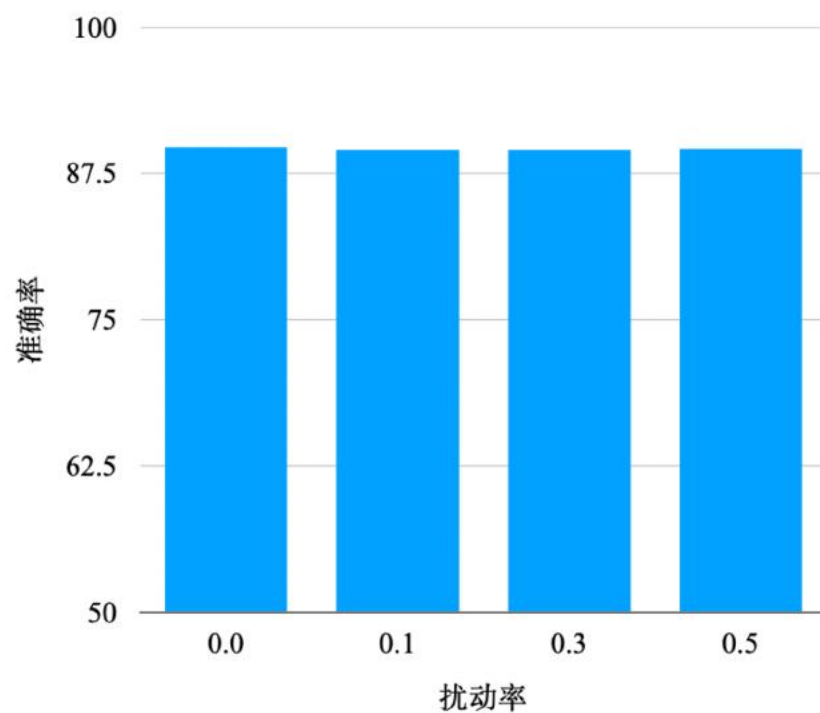


扰动影响

语法知识被破坏后
模型在 CoLA 任务上的表现



语义知识被破坏后
模型在 MRPC 任务上的表现



预训练模型在下游任务微调过程中并未很好地利用语言知识进行决策，而是在拟合任务数据集中一些简单特征和模式 (Gururangan et al., 2018; McCoy et al., 2019)。

研究问题

基于“开箱即用”的探针任务

1. 构建探针任务及数据集，**无需更新模型和训练分类器**，即可实现知识评测，探究：预训练语言模型本身是否具备句法知识和语义知识？模型是否学习到了事实知识，以及掌握常识推理能力？
2. 对比**微调前后**的模型知识差异：是否出现“**被动遗忘**”？
3. 扰动模型，令其“**主动遗忘**”知识，是否影响其他任务表现？

总结 基于“开箱即用”的探针任务

- 1. 预训练语言模型评测**：BERT类模型能够较好地编码语法和语义知识、部分事实知识，但不具备推理能力。
- 2. 微调影响**：微调后模型的语言学知识和事实知识出现明显“遗忘”，事实知识“遗忘”程度更甚，知识“遗忘”程度也与不同的模型和下游任务相关，例如在下游任务上表现最好的RoBERTa模型，反而“遗忘”了更多的知识。
- 3. 扰动影响**：探针任务所界定的语法和语义知识在模型中保持相对独立的编码状态；同时，微调过程中，模型对下游任务的学习对预训练模型本身编码的语言学知识并无明显依赖。

思考

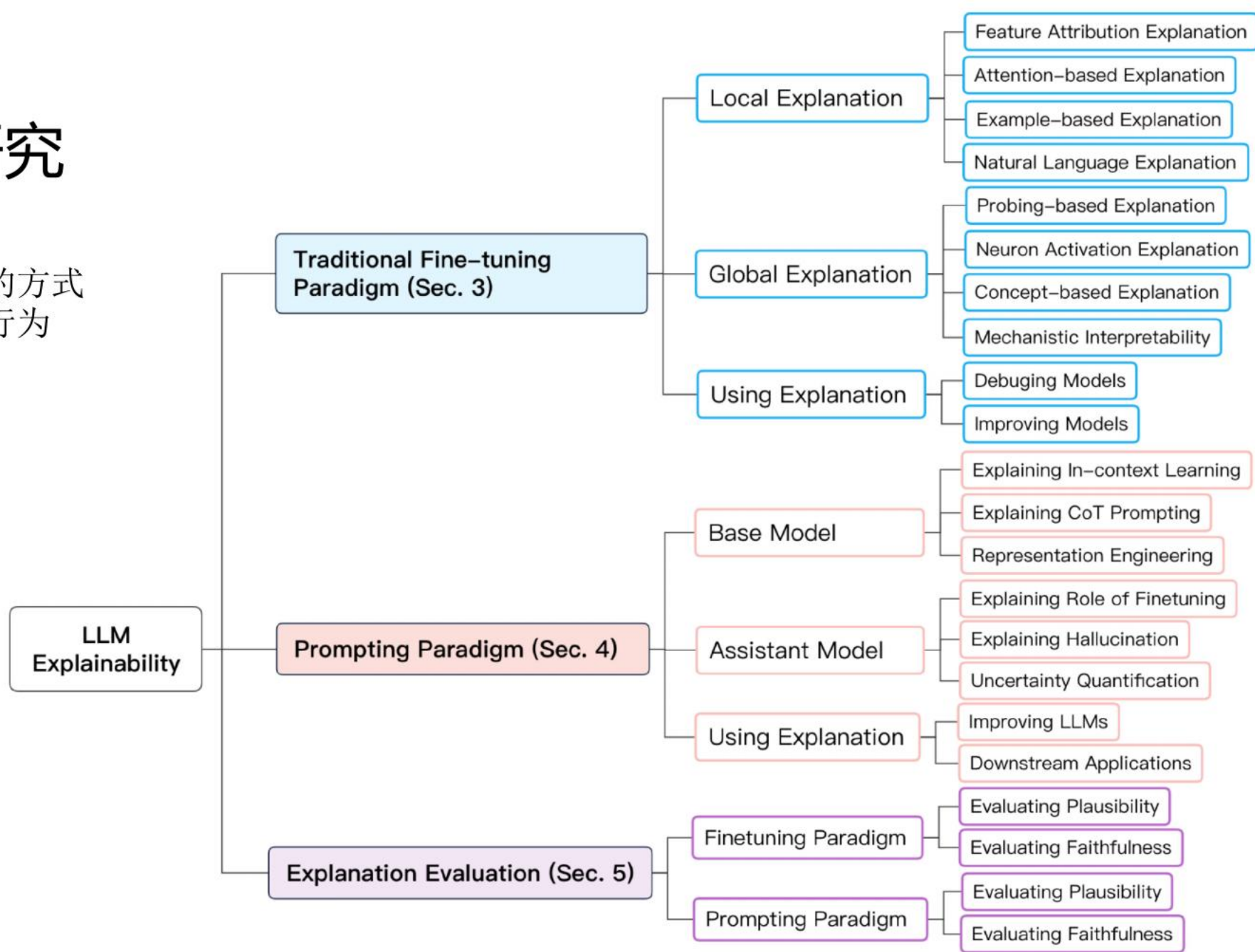
- 大语言模型的“推理”能力由何而来？
- 大语言模型多任务联合学习的优势
- 神经网络中可控遗忘的机制和作用
 - LLM Unlearning (Yao et al., 2024): (1) removing harmful responses, (2) erasing copyright-protected content as requested, and (3) reducing hallucinations.

Outline

- 背景
- 预训练语言模型的知识诊断
- 基于反事实样本的可解释性探索

可解释性研究

- 用人类可以理解的方式解释或呈现模型行为
 - 面向用户
 - 面向开发者



Why?

Should You Fine-Tune BERT for Automated Essay Scoring?

Elijah Mayfield and **Alan W Black**

Language Technologies Institute

Carnegie Mellon University

Table 1: Performance on each of ASAP datasets 2-6, in QWK. The final row shows the gap in QWK between the best neural model and the N-gram baseline.

Model	2	3	4	5	6
Human IRR	.80	.77	.85	.74	.74
Hewlett	.74	.75	.82	.83	.78
Taghipour	.69	.69	.81	.81	.82
Woods	.71	.71	.81	.82	.83
Cozma	.73	.68	.83	.83	.83
Rodriguez (BERT)	.68	.72	.80	.81	.81
Rodriguez (best)	.70	.72	.82	.82	.82
N-Grams	.71	.71	.78	.80	.79
Embeddings	.42	.41	.60	.49	.36
BERT-CLR	.66	.70	.80	.80	.79
BERT-1CYC	.64	.71	.82	.81	.79
BERT Features	.61	.59	.75	.75	.74
DistilBERT	.65	.70	.82	.81	.79
N-Gram Gap	-.05	.00	.04	.01	.00

Why?

Multi-Stage Pre-training for Automated Chinese Essay Scoring

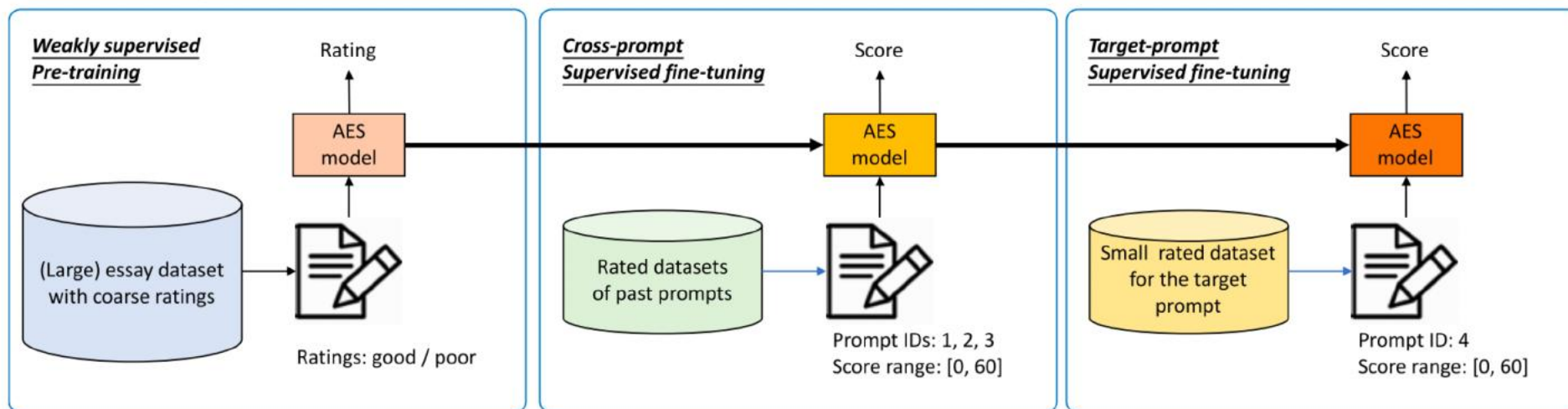
Wei Song¹, Kai Zhang¹, Ruiji Fu^{2,3}, Lizhen Liu¹, Ting Liu⁴, Miaomiao Cheng¹

¹College of Information Engineering and Academy for Multidisciplinary Studies,
Capital Normal University, Beijing, China

²State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China

³iFLYTEK AI Research (Hebei), Langfang, China

⁴Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China





王予沛

Background

- **L2C-Rater (Wang & Hu, 2021)**

- We built an L2 Chinese essay scoring system based on Ordinal Logistic Regression model with **90 linguistic features** with consideration of both language complexity and correctness.
- Web Tool: <https://l2c.shenshen.wiki>

- **Combination of Linguistic features and BERT pre-trained language model (2022.10-2023.04)**

- BERT performs better than linguistic-feature-based machine learning model.
- The integration of linguistic features yielded no performance enhancement.

What features **beyond linguistic patterns** does BERT capture???

Background

“柳暗花明又是坑” T_T



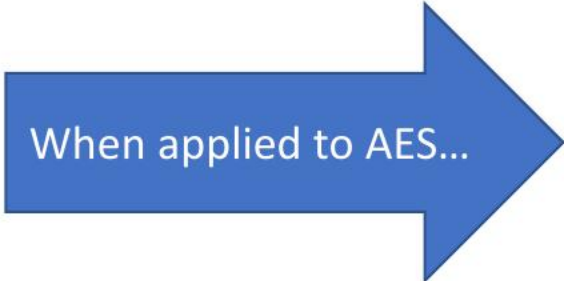
王予沛

• Interpretability of black-box neural models

- integrated gradients (2023.07-2023.08)

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad



- Occlusion based on token masking (2023.09-2023.12)
-

转换视角：模型决策是否遵循人类评分标准？

- 主流英语考试的评分标准：IELTS Writing, TOEFL iBT Independent Writing, TOEIC Writing, PTE Academic Writing and the ELLIPSE dataset.

A.1 IELTS Writing

- Task achievement: fully satisfies all the requirements of the task; clearly presents a **fully developed response**.
- Coherence and cohesion: uses **cohesion** in such a way that it attracts no attention; skillfully **manages paragraphing**.
- Lexical resource: uses **a wide range of vocabulary with very natural and sophisticated control of lexical features**; rare **minor errors occur only as 'slips'**.
- Grammatical range and accuracy: uses **a wide range of structures with full flexibility and accuracy**; rare **minor errors occur only as 'slips'**.

A.2 TOEFL Independent Writing

- Effectively addresses the topic and task.
- Is **well organized and well developed**, using clearly appropriate explanations, exemplifications and/or details.
- Displays **unity, progression and coherence**.
- Displays consistent facility in the use of language, demonstrating **syntactic variety, appropriate word choice and idiomaticity**, though it may have **minor lexical or grammatical errors**.

(1) **conventions and accuracy**

(2) **language complexity**

(3) **organization and development**

Counterfactual Analysis

- 反事实干预
 - 利用特定干预来生成反事实样本，从而分析给定情况下的因果效应 (Feder et al., 2022)。
- 现有的反事实样本生成方法
 - 关键词替换 (Garg et al., 2019)
 - 句子重写 (Ross et al., 2022; Wu et al., 2021)
 - 人工编辑 (Gardner et al., 2020)

通常局限于简单的局部干预，或需要耗费大量人力进行标注，阻碍了NLP模型中高层概念因果效应的实际评估。

LLMs can help!
(Gat et al., 2023)

Generation of Counterfactual Samples

Concept	Intervention	Description
Conventions & Accuracy	Error Correction	Prompt GPT-4 Turbo to correct spelling, punctuation, and grammar errors.
	Spelling Errors Introduction	Use nlpaug (Ma, 2019) to misspell 30% of words in 50% of sentences.
	Agreement Errors Introduction	Introduce subject-verb agreement (SVA) errors in 50% of sentences.
Language Complexity	Word Order Swapping (WOS)	Use nlpaug to swap 30% of words in 50% of sentences.
	Complexification	Prompt GPT-4 Turbo to enhance vocabulary and sentence structure.
Organization & Development	Simplification	Prompt GPT-4 Turbo to simplify vocabulary and sentence structure.
	Intra-paragraph Shuffling	Shuffle sentence order <i>within paragraphs</i> to disrupt <i>local</i> cohesion.
	Inter-text Shuffling	Shuffle sentence order <i>across the entire essay</i> to disrupt <i>global</i> cohesion.

Table 1: Overview of **positive** and **negative** counterfactual intervention methods used.

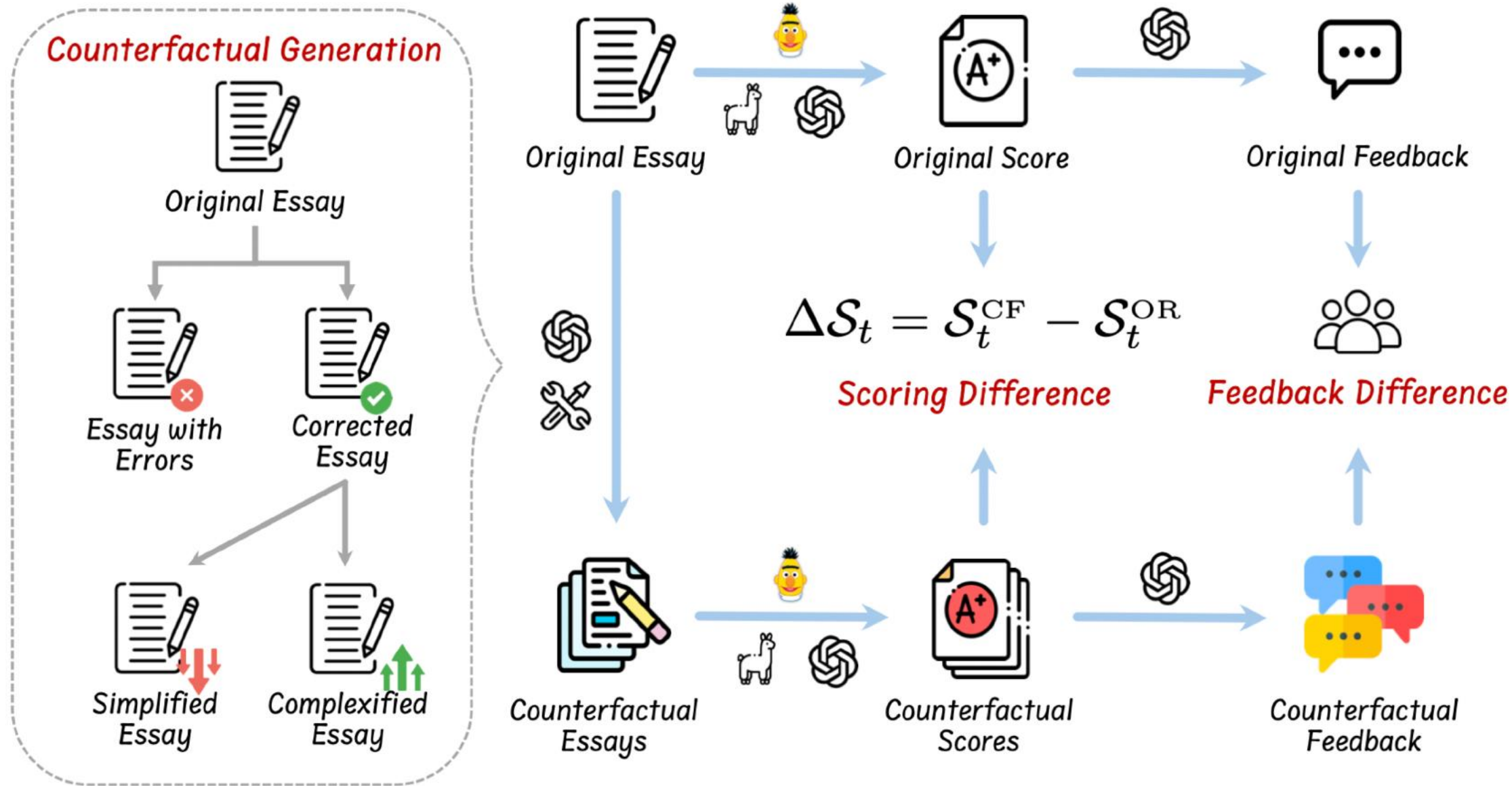
Original: In the **up coming** years **i wanna** save up and work hard to **get** money, so that **i** can open my own personal store.. \n\n **in** the store **i going to be** selling high **branded** shoes, clothes, and **outer collectives**. and **Im** going to **learning** how to invest my money as **i** sell things before opening up my dream store. \n\n **i wanna** do this as a job because **ive awlays had a passhine** for shoes and high **branded** things. **and if i** can pull this off **i think ill be veary succes full** in life **i will be getting and veary good salary ill be able to pay off anythig and have to money** to support my family.

Error Correction: "In the **upcoming** years, **I want to** save up and work hard to **earn** money, so that **I** can open my own personal store. \n\n **In** the store, **I am going to sell** high-branded shoes, clothes, and **other collectibles**. **I'm also** going to **learn** how to invest my money as **I** sell things before opening up my dream store. \n\n **I want to** do this as a job because **I've always had a passion** for shoes and high-branded **items**. **If I** can pull this off, **I think I'll be very successful** in life. **I will be earning a very good salary, I'll be able to pay off anything, and have the money** to support my family."

Complexification: In the forthcoming years, my ambition is to accumulate sufficient capital and diligently strive to generate income, thereby enabling me to inaugurate my own boutique. Within this emporium, I intend to offer an array of premium-branded footwear, apparel, and assorted collectibles. Concurrently, I shall endeavor to acquire financial acumen, particularly in investment strategies, as I engage in commerce prior to the launch of my envisioned enterprise. Pursuing this career path is fueled by an enduring zeal for exclusive footwear and luxury goods. Should I succeed in actualizing this venture, I am confident that it will herald substantial success in my life. Not only will I secure a lucrative income, but I will also possess the financial capacity to discharge any obligations and provide for my family's needs.

Simplification: In the next few years, I plan to save money and work hard. I want to make enough money to open my own store. In my store, I will sell popular shoes, clothes, and other special items. I will also learn how to use my money well as I sell things. I want to open my dream store one day. I have always loved shoes and popular brands. If I can do this, I think I will do very well in life. I will make a lot of money. I will be able to pay for everything and have enough to take care of my family.

Table 3: Example of a 3-point ELLIPSE essay and its counterfactual counterparts generated by GPT-4 Turbo.



Wang et al. (2024) Diagnosing the Rationale Alignment of AES Methods based on Linguistically-informed Counterfactuals

Datasets

	TOEFL11	ELLIPSE
Total Size	12,100 essays	6,482 essays
Data Split	Train 9,900	Train 3,914
	Val 1,100	Val 2,568
	Test 1,100	Test 2,568
Source	2006-2007 TOEFL exams	8th-12th grade English learners
Rating Scale	Low/Medium/High	1-5 scale (0.5 increments)
Eval Metrics	Weighted F1, QWK	RMSE, QWK

Scoring Performance on Original Essays

	Setting	TOEFL11		ELLIPSE	
		F1 ↑	QWK ↑	RMSE ↓	QWK ↑
Small models	BERT	0.783	0.736	0.437	0.680
	ROBERTA	0.795	0.739	0.430	0.695
	DEBERTA	0.790	0.741	0.422	0.720
LLMs	GPT-3.5-ZSL	0.599	0.408	0.701	0.399
	GPT-3.5-FSL	0.546	0.314	<u>0.570</u>	0.378
	GPT-3.5-SFT-100	0.710	0.592	0.550	0.629
	GPT-4-ZSL	0.368	0.380	0.960	0.261
	GPT-4-FSL	0.490	0.477	0.680	0.466
	LLAMA-3-8B-ZSL	0.558	0.297	0.628	0.345
	LLAMA-3-8B-FSL	0.435	0.441	1.039	0.054
	LLAMA-3-70B-ZSL	0.524	0.390	0.903	0.182
	LLAMA-3-70B-FSL	<u>0.609</u>	<u>0.562</u>	0.589	<u>0.503</u>

Table 5: The scoring agreement performance on both test sets: **best setting** in bold, **fine-tuned GPT-3.5** with a green shadow, **best off-the-shelf LLMs** underlined. Metrics with ↑ indicate that higher values are better, while the one with ↓ indicates that lower values are better.

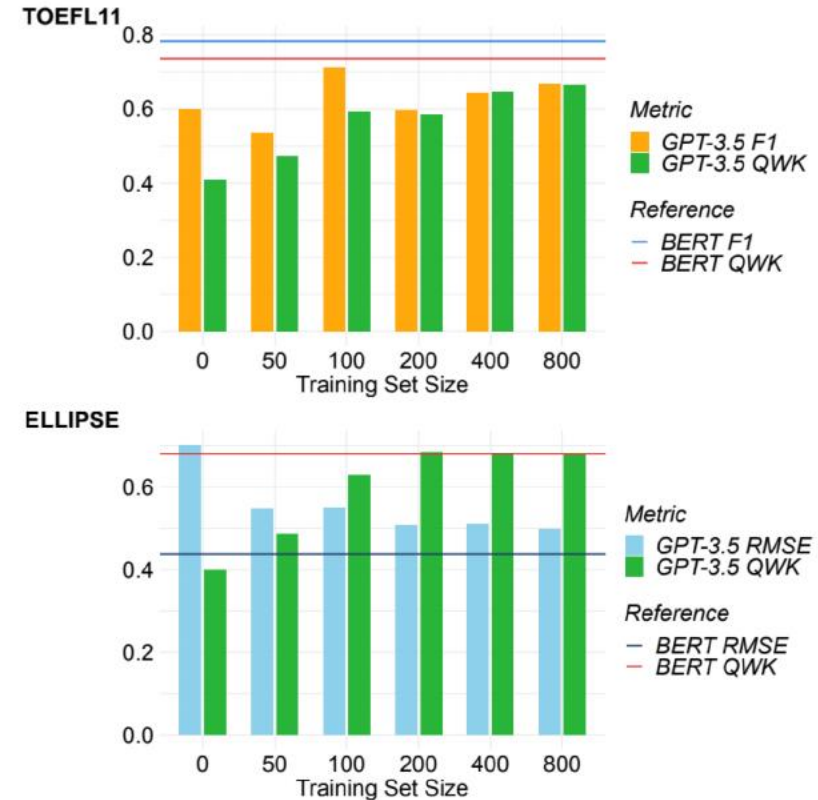


Figure 3: Scoring performance of GPT-3.5 SFT models with varying size of training data. The models' performance improves as the number of training samples increases, reaching comparable or equivalent levels to BERT-like models.

Mean score differences between original and counterfactual groups

Dataset	Setting	Conventions & Accuracy			Language Complexity		Organization & Development		
		Error Correction (+)	Error Introduction (-)		Complexification (+)	Simplification (-)	InParaShuffle (-)	InTextShuffle (-)	
			-	Spelling					SVA
TOEFL11	BERT	1.03 ^{+.043} _{-.041}	-0.92 ^{+.032} _{-.033}	-0.22 ^{+.013} _{-.014}	-1.26 ^{+.033} _{-.032}	0.42 ^{+.035} _{-.035}	-0.69 ^{+.033} _{-.033}	-0.01 ^{+.006} _{-.006}	-0.01 ^{+.006} _{-.006}
	ROBERTA	0.99 ^{+.043} _{-.044}	-0.79 ^{+.033} _{-.032}	-0.45 ^{+.021} _{-.021}	-1.13 ^{+.033} _{-.033}	0.24 ^{+.032} _{-.031}	-0.35 ^{+.025} _{-.025}	-0.19 ^{+.010} _{-.011}	-0.02 ^{+.005} _{-.005}
	DEBERTA	1.19 ^{+.045} _{-.046}	-0.92 ^{+.031} _{-.031}	-0.35 ^{+.016} _{-.016}	-1.24 ^{+.033} _{-.032}	0.33 ^{+.034} _{-.032}	-0.27 ^{+.027} _{-.026}	-0.06 ^{+.005} _{-.005}	-0.06 ^{+.005} _{-.005}
	GPT-3.5-ZSL	0.64 ^{+.032} _{-.031}	-0.76 ^{+.033} _{-.034}	-0.20 ^{+.026} _{-.026}	-0.59 ^{+.032} _{-.030}	0.27 ^{+.025} _{-.024}	0.01 ^{+.019} _{-.020}	-0.31 ^{+.030} _{-.030}	-0.42 ^{+.032} _{-.032}
	GPT-4-ZSL	0.92 ^{+.025} _{-.025}	-0.80 ^{+.025} _{-.025}	-0.35 ^{+.021} _{-.021}	-0.80 ^{+.026} _{-.026}	0.66 ^{+.025} _{-.025}	-0.24 ^{+.021} _{-.021}	-0.24 ^{+.018} _{-.017}	-0.29 ^{+.019} _{-.019}
ELLIPSE	LLAMA-3-8B-ZSL	0.58 ^{+.027} _{-.026}	-0.37 ^{+.029} _{-.029}	-0.07 ^{+.018} _{-.018}	-0.17 ^{+.023} _{-.024}	0.57 ^{+.026} _{-.026}	-0.11 ^{+.023} _{-.023}	-0.15 ^{+.024} _{-.024}	-0.23 ^{+.026} _{-.026}
	LLAMA-3-70B-ZSL	0.64 ^{+.026} _{-.025}	-0.56 ^{+.025} _{-.025}	-0.24 ^{+.021} _{-.022}	-0.41 ^{+.023} _{-.023}	1.19 ^{+.032} _{-.032}	-0.17 ^{+.024} _{-.024}	-0.15 ^{+.019} _{-.019}	-0.19 ^{+.021} _{-.021}
	BERT	0.84 ^{+.014} _{-.014}	-0.57 ^{+.011} _{-.011}	-0.09 ^{+.003} _{-.003}	-0.57 ^{+.011} _{-.011}	0.31 ^{+.009} _{-.009}	-0.11 ^{+.008} _{-.008}	-0.01 ^{+.002} _{-.002}	-0.02 ^{+.002} _{-.003}
	ROBERTA	0.92 ^{+.014} _{-.015}	-0.50 ^{+.009} _{-.009}	-0.11 ^{+.003} _{-.003}	-0.54 ^{+.009} _{-.009}	0.25 ^{+.008} _{-.007}	-0.05 ^{+.007} _{-.007}	-0.01 ^{+.002} _{-.002}	-0.10 ^{+.003} _{-.003}
	DEBERTA	1.06 ^{+.016} _{-.016}	-0.64 ^{+.013} _{-.013}	-0.20 ^{+.006} _{-.006}	-0.64 ^{+.013} _{-.013}	-0.08 ^{+.007} _{-.007}	0.01 ^{+.005} _{-.005}	-0.02 ^{+.001} _{-.001}	-0.07 ^{+.002} _{-.002}
	GPT-3.5-ZSL	0.77 ^{+.019} _{-.018}	-0.60 ^{+.019} _{-.018}	-0.19 ^{+.015} _{-.015}	-0.35 ^{+.018} _{-.018}	0.48 ^{+.016} _{-.016}	0.08 ^{+.014} _{-.014}	-0.15 ^{+.015} _{-.014}	-0.18 ^{+.016} _{-.017}
	GPT-3.5-FSL	0.35 ^{+.014} _{-.014}	-0.46 ^{+.015} _{-.015}	-0.15 ^{+.012} _{-.012}	-0.31 ^{+.014} _{-.014}	0.36 ^{+.014} _{-.014}	-0.04 ^{+.012} _{-.012}	-0.11 ^{+.013} _{-.012}	-0.16 ^{+.014} _{-.014}
	GPT-4-ZSL*	0.87 ^{+.060} _{-.058}	-0.64 ^{+.047} _{-.047}	-0.30 ^{+.045} _{-.045}	-0.56 ^{+.045} _{-.045}	0.96 ^{+.065} _{-.065}	-0.05 ^{+.058} _{-.057}	-0.10 ^{+.033} _{-.035}	-0.19 ^{+.037} _{-.040}
	GPT-4-FSL*	0.61 ^{+.052} _{-.048}	-0.71 ^{+.060} _{-.060}	-0.27 ^{+.050} _{-.050}	-0.56 ^{+.048} _{-.050}	0.67 ^{+.055} _{-.052}	-0.09 ^{+.045} _{-.043}	-0.14 ^{+.032} _{-.035}	-0.23 ^{+.042} _{-.045}
	LLAMA-3-8B-ZSL	0.32 ^{+.017} _{-.016}	-0.31 ^{+.018} _{-.018}	-0.06 ^{+.011} _{-.011}	-0.11 ^{+.013} _{-.014}	0.70 ^{+.013} _{-.013}	0.01 ^{+.009} _{-.010}	-0.06 ^{+.011} _{-.012}	-0.10 ^{+.014} _{-.014}
	LLAMA-3-8B-FSL	0.06 ^{+.011} _{-.011}	-0.11 ^{+.016} _{-.016}	-0.02 ^{+.008} _{-.008}	-0.06 ^{+.011} _{-.011}	0.07 ^{+.016} _{-.016}	-0.00 ^{+.007} _{-.007}	-0.02 ^{+.010} _{-.010}	-0.02 ^{+.012} _{-.011}
	LLAMA-3-70B-ZSL*	0.51 ^{+.018} _{-.018}	-0.41 ^{+.011} _{-.011}	-0.11 ^{+.009} _{-.009}	-0.19 ^{+.010} _{-.010}	1.63 ^{+.019} _{-.019}	0.03 ^{+.018} _{-.018}	-0.03 ^{+.007} _{-.007}	-0.06 ^{+.008} _{-.008}
	LLAMA-3-70B-FSL*	0.51 ^{+.070} _{-.068}	-0.54 ^{+.065} _{-.065}	-0.12 ^{+.033} _{-.035}	-0.24 ^{+.050} _{-.052}	1.08 ^{+.055} _{-.055}	-0.04 ^{+.040} _{-.040}	-0.11 ^{+.040} _{-.042}	-0.13 ^{+.043} _{-.045}
	GPT-3.5-SFT-50*	0.83 ^{+.075} _{-.072}	-0.64 ^{+.077} _{-.080}	-0.14 ^{+.045} _{-.050}	-0.34 ^{+.065} _{-.068}	0.96 ^{+.060} _{-.062}	0.08 ^{+.055} _{-.052}	-0.09 ^{+.045} _{-.045}	-0.10 ^{+.047} _{-.050}
GPT-3.5-SFT-100*	1.12 ^{+.080} _{-.080}	-0.95 ^{+.080} _{-.080}	-0.26 ^{+.052} _{-.052}	-0.58 ^{+.057} _{-.055}	0.88 ^{+.055} _{-.057}	0.05 ^{+.050} _{-.048}	-0.18 ^{+.047} _{-.050}	-0.19 ^{+.048} _{-.050}	
GPT-3.5-SFT-200*	1.03 ^{+.092} _{-.090}	-0.57 ^{+.087} _{-.090}	-0.01 ^{+.068} _{-.070}	-0.32 ^{+.072} _{-.070}	0.79 ^{+.052} _{-.055}	-0.02 ^{+.037} _{-.037}	0.06 ^{+.060} _{-.060}	0.02 ^{+.062} _{-.062}	
GPT-3.5-SFT-400*	1.11 ^{+.087} _{-.090}	-0.95 ^{+.075} _{-.075}	-0.30 ^{+.060} _{-.060}	-0.66 ^{+.068} _{-.065}	0.76 ^{+.055} _{-.057}	-0.03 ^{+.045} _{-.042}	-0.18 ^{+.052} _{-.052}	-0.23 ^{+.050} _{-.052}	
GPT-3.5-SFT-800*	1.02 ^{+.085} _{-.085}	-0.83 ^{+.080} _{-.080}	-0.23 ^{+.065} _{-.067}	-0.55 ^{+.070} _{-.070}	0.94 ^{+.055} _{-.055}	-0.03 ^{+.048} _{-.050}	-0.14 ^{+.052} _{-.055}	-0.23 ^{+.060} _{-.062}	

Feedback Generation and Evaluation

200 samples from ELLIPSE, GPT4-turbo, three annotators

Session 1: Essay Scoring

User: Read and evaluate the essay: ...

Assistant: {'score': 3.0}

Session 2: Providing Feedback

User: Please provide balanced and constructive feedback on the following aspects of the essay you have just rated (not the example essay):

1. Organization: ...
2. Language Use: ...
3. Conventions: ...

Your response should be a structured JSON object with the following keys:

```
```  
{
 "organization_feedback": "",
 "language_use_feedback": "",
 "conventions_feedback": ""
}
```
```

If possible, include direct citations from the essay to substantiate your feedback.

Which feedback (I or II) more clearly or frequently points out spelling errors?

- (a) Feedback I
- (b) Feedback II
- (c) Both are similar
- (d) Uncertain

Figure 4: An Example of Feedback Generation

Feedback for the Original

****organization_feedback****

The essay begins with a clear topic sentence that introduces the importance of parks for improving mood, but the development of ideas throughout the body is somewhat disjointed. The writer provides personal anecdotes and general observations to support the argument, but transitions between ideas are abrupt, and there is a lack of clear logical flow. The conclusion does not effectively summarize the main points or provide a strong final thought on the topic. The essay would benefit from a more structured approach, with each paragraph focusing on a single aspect of the importance of parks and using transitions to guide the reader through the argument.

****language_use_feedback****

Morphology: There are several errors in verb tense consistency ('i use to just go outside' should be 'I used to just go outside') and subject-verb agreement ('kids finds parks activity enjoyable' should be 'kids find park activities enjoyable'). Syntax: Sentence structures are often simplistic and repetitive, lacking variety and complexity. There are also instances of awkward phrasing ('my mom always straggle with tell my brothers to go outside' should be 'my mom always struggles to tell my brothers to go outside'). Vocabulary: The vocabulary is generally appropriate but lacks diversity and sophistication. There are instances of imprecise language ('be open minded' could be 'become open-minded', 'enjoy life' could be 'enjoy the outdoors').

****conventions_feedback****

The essay contains several spelling errors ('straggle' should be 'struggle', 'rob' should be 'rope'), incorrect capitalization ('iPod', 'iPhone', 'I', and the first word of sentences should be capitalized), and punctuation mistakes (missing commas in compound sentences, incorrect spacing around punctuation marks). Attention to these details would improve the readability and professionalism of the essay.

Feedback for the Counterfactual

****organization_feedback****

The essay's organization needs improvement. Ideas about the importance of parks and their impact on mood and childhood experiences are presented, but the logical flow is disjointed. Transitions between ideas are abrupt or missing, making it difficult to follow the argument. The introduction lacks a clear thesis statement, and the body paragraphs do not consistently develop the initial idea. The conclusion is not clearly defined, and the essay ends abruptly without summarizing the main points or providing a final thought on the importance of parks for children.

****language_use_feedback****

Morphology: There are several errors in verb tenses and subject-verb agreement (e.g., 'my mm takes him ty the park and ha play' should be 'my mom takes him to the park and he plays'). Syntax: Sentence structures are often incomplete or incoherent, which disrupts the flow of information (e.g., 'Know i barely even go outside i go ti school and then i come back is just' is difficult to understand). Vocabulary: Word choice is sometimes inaccurate or too simplistic, and there is a lack of diversity in vocabulary. More precise language could enhance the clarity of the essay (e.g., 'bouring' should be 'boring', 'activity enjoyable' should be 'activities enjoyable').

****conventions_feedback****

The essay contains numerous spelling errors (e.g., 'iMy', 'btother', 'ty', 'ha play', 'lf', 'alls', 'jn', 'tie', 'suing', 'activity', 'mint', 'whi offen', 'leke', 'ext', 'litter', 'bi on m'i phone', 'outside', 'tou', 'weill', 'offen mindend', 'thingh', 'dayl', 'field', 'wak ta giong tte park', 'nown', 'mer', 'estay', 'gou'). Capitalization is inconsistent, and punctuation is often missing or used incorrectly, making it challenging to understand the intended meaning of sentences.

Feedback Generation and Evaluation

200 samples from ELLIPSE, GPT4-turbo, three annotators

Category	Counterfactual Type	Detection Rate%
Conventions	Error Correction	72
	Spelling	68
	SVA	48
	WOS	80
Language Complexity	Complexification	100
	Simplification	32
Organization	InParaShuffle	40
	InTextShuffle	20

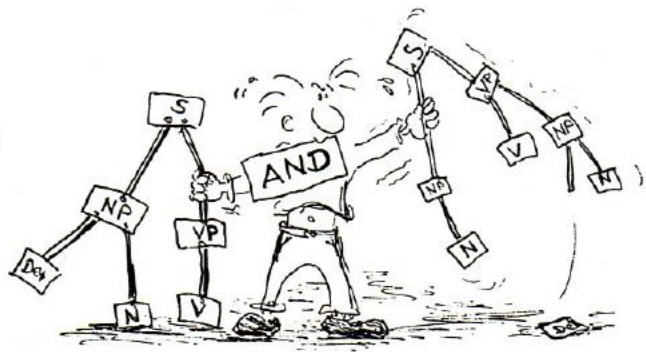
Table 7: Voting-Based Detection Rates of Original vs. Counterfactual Feedback.

Conclusions

- **BERT-like models** show higher scoring agreement with human raters than LLMs, and they are highly sensitive to interventions on **conventions & accuracy and language complexity**. However, they struggle to distinguish interventions on **organization and development**.
- **LLMs** respond more comprehensively to all the interventions than BERT-like models do, indicating **a closer alignment** with the criteria specified in scoring rubrics.
- Further analysis shows that while **zero-shot prompted LLMs show lower scoring agreement** compared to BERT-like models, a major reason for this is that **they tend to be conservative or strict when evaluating the essay**. FSL and SFT could significantly increase the agreement level with annotated essays serve as anchors to neutralize the conservatism.

多模态、多智能体、具身智能的技术发展催生新的需求：

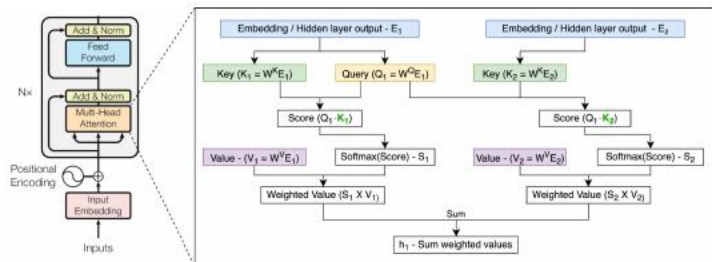
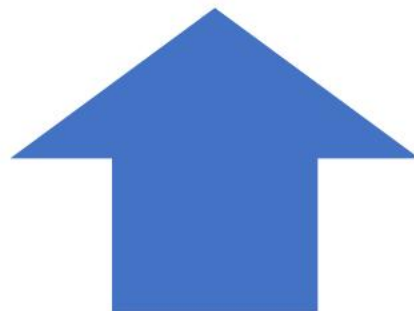
- 视听模态能否促进文本模态的语言理解和推理能力？
- 话语互动视角的训练数据构建和模型评估
- 语言演化模拟
-



Linguistic Knowledge



Neural Language Models



第一、巧妙利用模型内在“本能”（填空、向量、困惑度等）。

第二、通过会话合理引导或注入知识，使之产出符合研究者所需的“标注数据”。



00100
00111
10

00100
00111
110



敬请各位师友批评指正！

