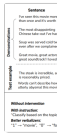
ACL 2023

表格视图

表格名称

标题	标签	字段名	字
How Do In-Context Examples Affect Compositional Generalization?	We study three potential factors: similarity, diversity and complexity. Our systematic experiments indicate that **in-context examples should be structurally similar to the test case, diverse from each other, and individually simple**. 依然存在两个问题：One is that in-context learning has difficulty recombining fictional words (e.g., random tokens) rather than commonly used ones. The other one is that in-context examples are still required to cover the linguistic structures in NL expressions, even though the backbone model has been pre-trained on large corpus	https://arxiv.org/pdf/2305.04835.pdf	
Rethinking Semi-supervised Learning with Language Models	任务适应性预训练（TAPT）是一个强大的，更强大的 SSL 学习者，即使使用只有几百个未标记的样本或在域的变化的存在下，相比更复杂的 ST 方法，往往带来更大的改善 SSL 比在完全监督的设置。		
Measuring Inductive Biases of In-Context Learning with Underspecified Demonstrations	我们发现 llm 表现出明显的特征偏差，例如，显示出强烈的偏见，根据情感预测标签，而不是肤浅的词汇特征，如标点符号。其次，我们评估了不同干预措施的效果，这些干预措施旨在施加有利于特定特征的归纳偏见，例如添加自然语言指令或使用语义相关的标签词。我们发现，虽然许多干预措施可以影响学习者对特定特征的偏好，但很难克服强烈的先验偏见。	https://arxiv.org/pdf/2305.13299v1.pdf	
Unified Demonstration Retriever for In-Context Learning		https://arxiv.org/pdf/2305.04320.pdf	

White-Box Multi-Objective Adversarial Attack on Dialogue Generation			
This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models	调研研究 bias 的论文缺乏各种东西的	https://arxiv.org/pdf/2305.12757v1.pdf	
Towards Robust Personalized Dialogue Generation via Order-Insensitive Representation Regularization	个性化对话生成，个人陈述部分的顺序很重要，提出一个鲁棒的 不受顺序干扰的	https://arxiv.org/pdf/2305.12782v1.pdf	
Explaining How Transformers Use Context to Build Predictions	好文，探究 context 上下 token 之间的互相作用对最后 prediction 的影响，子弟篇 https://arxiv.org/pdf/2203.04212.pdf	https://arxiv.org/pdf/2305.12535v1.pdf	
Can We Edit Factual Knowledge by In-Context Learning?	in-context learning knowledge editing (IKE)，用 ICL 修改知识，不如 ROME，但是不费开销	https://arxiv.org/pdf/2305.12740v1.pdf	
Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker	information retrieval task，短篇，discriminator 判别 prompt 的分数，优化 prompt 生成	https://arxiv.org/pdf/2305.13729.pdf	
Interpreting Transformer's Attention Dynamic Memory and Visualizing the Semantic Information Flow of GPT	把 gpt 内部做成了个 flow graph，点表示 hidden states，边表示 interaction。发现 layer norm 用于语义过滤	https://arxiv.org/pdf/2305.13417.pdf	
Debiasing should be	针对解决 toxic 问题的 debias 方法，提出三个问	https://	

Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models	题，反向偏置规范会放大 undesirable bias 吗、偏差规范的存在是否显著减少了 undesirable bias、是否在 adversarial 和 non-adversarial 场景都可以。提出 Instructive Debiasing，用 prompt，pattern "Be s for: x".	arxiv.org/pdf/2305.14307.pdf	
NeuroX Library for Neuron Analysis of Deep NLP Models	神经元分析，提出 NeuroX，用于对自然语言处理模型进行神经元分析。给定一个带注释的文本语料库，神经元解释方法旨在提供模型中神经元对一个或多个属性的重要性的排序。判断哪一层的什么神经元对什么词性敏感。	https://arxiv.org/pdf/2305.17073.pdf	
CREST: A Joint Framework for Rationalization and Counterfactual Text Generation	提出一个更自然的生成反事实例子的 CREST 模型，可用于大规模的数据增强。	https://arxiv.org/pdf/2305.17075.pdf	
Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios	我们发现，在反事实场景中，模型始终能够覆盖真实世界的知识，并且在更强的基线世界知识的情况下，这种效果更加强大。我们还发现，对于大多数模型，这种效果似乎主要是由简单的词汇线索驱动的。	https://arxiv.org/abs/2305.16572	
Don't Retrain, Just Rewrite: Countering Adversarial Perturbations by Rewriting Text	设计了一个即插即用的模型，重写 text 来达到对抗扰动。具体：用一个独立的模型消除对抗扰动，然后再接下游模型，不 ft。一般吧。	https://arxiv.org/pdf/2305.16444.pdf	
Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation	我们比较了 Few-Shot finetune 和 ICL 的泛化，参数数量从 125 M 到 30 B。我们的研究结果表明，微调的语言模型实际上可以很好地 OOD。我们发现，这两种方法的推广相似;它们表现出很大的变化，并取决于诸如模型大小和示例数量等属性	https://arxiv.org/abs/2305.16938	
Finspector: A Human-Centered Visual Inspection Tool for Exploring and Comparing Biases among Foundation Models	可视化工具，通过语言模型生成的对数似然分数来检测不同类别中的偏见。该工具的目标是使研究人员能够使用可视化分析轻松识别潜在的偏见	https://arxiv.org/abs/2305.16937	
Large Language Models Are Partially Primed in Pronoun Interpretation	探究 LLM 会不会像人类一样有偏见，引入心理学数据集。提出了人类可能具有的 2 个 bias：subject bias（比如 she 更容易被翻译成主语	https://arxiv.org/pdf/2	

	Ada）、goal bias（she 总是指代 target 而不是 source）。我们发现 GPT 可以适应并改变其 syntactic bias，但 semantic bias 不成，InstructGPT 仍然表现出 goal bias	305.16917.pdf
MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting	我们提出了 MultiTool-CoT，一种新的框架，利用链的思想（CoT）提示，将多个外部工具，如计算器和知识检索器	https://arxiv.org/abs/2305.16896

＋ 新增

表格视图

表格名称

标题	标签	字段名
How Do In-Context Examples Affect Compositional Generalization?	We study three potential factors: similarity, diversity and complexity. Our systematic experiments indicate that **in-context examples should be structurally similar to the test case, diverse from each other, and individually simple**. 依然存在两个问题：One is that in-context learning has difficulty recombining fictional words (e.g., random tokens) rather than commonly used ones. The other one is that in-context examples are still required to cover the linguistic structures in NL expressions, even though the backbone model has been pre-trained on large corpus	https://arxiv.org/pdf/2305.04835.pdf
Rethinking Semi-supervised Learning with Language Models	任务适应性预训练（TAPT）是一个强大的，更强大的 SSL 学习者，即使使用只有几百个未标记的样本或在域的变化的存在下，相比更复杂的 ST 方法，往往带来更大的改善 SSL 比在完全监督的设置。	
Measuring Inductive Biases of In-Context Learning with Underspecified Demonstrations	我们发现 llm 表现出明显的特征偏差，例如，显示出强烈的偏见，根据情感预测标签，而不是肤浅的词汇特征，如标点符号。其次，我们评估了不同干预措施的效果，这些干预措施旨在施加有利于特定特征的归纳偏见，例如添加自然语言指令或使用语义相关的标签词。我们发现，虽然许多干预措施可	https://arxiv.org/pdf/2305.13299v1.pdf

	以影响学习者对特定特征的偏好，但很难克服强烈的先验偏见。	
Unified Demonstration Retriever for In-Context Learning		https://arxiv.org/pdf/2305.04320.pdf
White-Box Multi-Objective Adversarial Attack on Dialogue Generation		
This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models	调研研究 bias 的论文缺乏各种东西的	https://arxiv.org/pdf/2305.12757v1.pdf
Towards Robust Personalized Dialogue Generation via Order-Insensitive Representation Regularization	个性化对话生成，个人陈述部分的顺序很重要，提出一个鲁棒的 不受顺序干扰的	https://arxiv.org/pdf/2305.12782v1.pdf
Explaining How Transformers Use Context to Build Predictions	好文，探究 context 上下 token 之间的互相作用对最后 prediction 的影响，子弟篇 https://arxiv.org/pdf/2203.04212.pdf	https://arxiv.org/pdf/2305.12535v1.pdf
Can We Edit Factual Knowledge by In-Context Learning?	in-context learning knowledge editing (IKE)，用 ICL 修改知识，不如 ROME，但是不费开销	https://arxiv.org/pdf/2305.12740v1.pdf
Discrete Prompt Optimization via Constrained Generation for Zero-shot Re-ranker	information retrieval task，短篇，discriminator 判别 prompt 的分数，优化 prompt 生成	https://arxiv.org/pdf/2305.13729.pdf
Interpreting Transformer's Attention Dynamic Memory and Visualizing	把 gpt 内部做成了个 flow graph，点表示 hidden states，边表示 interaction。发现 layer norm 用于语义过滤	https://arxiv.org/pdf/2305.13417.pdf

the Semantic Information Flow of GPT		
Debiasing should be Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models	针对解决 toxic 问题的 debias 方法，提出三个问题，反向偏置规范会放大 undesirable bias 吗、偏差规范的存在是否显著减少了 undesirable bias、是否在 adversarial 和 non-adversarial 场景都可以。提出 Instructive Debiasing，用 prompt，pattern "Be s for: x".	https://arxiv.org/pdf/2305.14307.pdf
NeuroX Library for Neuron Analysis of Deep NLP Models	神经元分析，提出 NeuroX，用于对自然语言处理模型进行神经元分析。给定一个带注释的文本语料库，神经元解释方法旨在提供模型中神经元对一个或多个属性的重要性的排序。判断哪一层的什么神经元对什么词性敏感。	https://arxiv.org/pdf/2305.17073.pdf
CREST: A Joint Framework for Rationalization and Counterfactual Text Generation	提出一个更自然的生成反事实例子的 CREST 模型，可用于大规模的数据增强。	https://arxiv.org/pdf/2305.17075.pdf
Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios	我们发现，在反事实场景中，模型始终能够覆盖真实世界的知识，并且在更强的基线世界知识的情况下，这种效果更加强大。我们还发现，对于大多数模型，这种效果似乎主要是由简单的词汇线索驱动的。	https://arxiv.org/abs/2305.16572
Don't Retrain, Just Rewrite: Countering Adversarial Perturbations by Rewriting Text	设计了一个即插即用的模型，重写 text 来达到对抗扰动。具体：用一个独立的模型消除对抗扰动，然后再接下游模型，不 ft。一般吧。	https://arxiv.org/pdf/2305.16444.pdf
Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation	我们比较了 Few-Shot finetune 和 ICL 的泛化，参数数量从 125 M 到 30 B。我们的研究结果表明，微调的语言模型实际上可以很好地 OOD。我们发现，这两种方法的推广相似;它们表现出很大的变化，并取决于诸如模型大小和示例数量等属性	https://arxiv.org/abs/2305.16938
Finspector: A Human-Centered Visual Inspection Tool for Exploring and	可视化工具，通过语言模型生成的对数似然分数来检测不同类别中的偏见。该工具的目标是使研究人员能够使用可视化分析轻松识别潜在的偏见	https://arxiv.org/abs/2305.16937

Comparing Biases among Foundation Models		
Large Language Models Are Partially Primed in Pronoun Interpretation	探究 LLM 会不会像人类一样有偏见，引入心理学数据集。提出了人类可能具有的 2 个 bias：subject bias（比如 she 更容易被翻译成主语 Ada）、goal bias（she 总是指代 target 而不是 source）。我们发现 GPT 可以适应并改变其 syntactic bias，但 semantic bias 不成，InstructGPT 仍然表现出 goal bias	https://arxiv.org/pdf/2305.16917.pdf
MultiTool-CoT: GPT-3 Can Use Multiple External Tools with Chain of Thought Prompting	我们提出了 MultiTool-CoT，一种新的框架，利用链的思想（CoT）提示，将多个外部工具，如计算器和知识检索器	https://arxiv.org/abs/2305.16896