自然语言处理
Natural Language Processing

第 11 章　自然语言处理系统评价

授课教师：　黄河燕

授课时间：2024.12

内 容 提 要

11.1 概述

11.2 典型评价指标

11.3 评价方法

11.4 基于大模型的自动评价

第一节

概述

常见的自然语言处理任务

- 分类任务：
  - 情感分析、文本分类、命名实体识别.....
- 生成任务：
  - 机器翻译、文本摘要、对话生成.....

自然语言处理系统评价

- 相比于视觉、语音等领域任务
  - 自然语言处理任务输出形式差异大：
    - 除了分类任务会输出类别之外，生成任务还会输出具有开放性和主观性的文本，很难用简单的数值衡量。
  - 实际应用中评价维度差异大：
    - 分类系统：文本分类是否正确？
    - 翻译系统：翻译文本是否流畅、语义准确？
    - 对话系统：是否连贯、流畅且满足用户需求？
    - 文本生成：是否内容丰富且符合语境？

自然语言处理系统评价

- 怎样合理评价不同类型的系统？

- 情感分析任务
  - 模型根据输入（如用户的评论内容），判断其属于正面、负面还是中性内容。

情感分析

- “这家餐厅的服务太差了，等了半小时还没上菜。”--模型输出：负面评论
- “菜品非常美味，环境也很优雅，下次还会再来！”--模型输出：正面评论
- “菜品还不错，但价格偏高，服务一般。”--模型输出：中性评论

- 我们对该模型的评价：“还不错”
  - 评价标准：根据经验，输出结果符合预期
  - 这种评价是否科学合理？

自然语言处理系统评价

- 自然语言处理系统评价：评估衡量生成结果质量的过程，通过一定的方法和指标，判断系统是否完成任务并满足目标需求。
- 核心任务：判断模型的输出是否准确以及符合任务需求。
- 目的：
  - 验证模型性能：帮助研究者了解模型的生成能力是否达标。
  - 优化模型迭代：发现模型不足，为改进提供方向。
  - 选择最佳模型：在候选模型中确定最优方案。

自然语言处理系统评价的核心问题

- 核心问题：
  - 如何衡量系统输出的准确度？
  - 如何衡量系统生成文本的质量？（如流畅性、语义一致性、事实性）
  - 评价结果是否具有一致性和可复现性？
  - 评价方法能否捕捉深层的语义信息？
  - 开放式任务可能具有多种正确答案，在考虑多样性的基础上，怎样进行合理的评价？

自然语言处理系统评价的演变

- 分类任务
  - 长期使用精确率、召回率、F1 Score
- 生成任务的历史演变：
  - 早期：基于规则的评价（BLEU、ROUGE）。
  - 深度学习时代：使用预训练模型，引入语义理解（BERTScore、COMET）。
  - 大模型时代：基于 ChatGPT 等大模型的开放式和灵活评价方法逐步成为趋势。

主要授课内容

- 典型评价指标
  - 常用评价指标（如 BLEU、ROUGE）。
- 评价方法
  - 典型评价指标的改进方法。
  - 人工评价的优缺点及应用场景。
- 基于大模型的自动评价
  - 如何应用大模型自动评价自然语言处理系统。
  - 大模型自动评价的局限性。

第二节

典型评价指标

典型评价指标

- 针对不同自然语言处理任务，有不同的评价指标。
  - 文本分类，命名实体识别（分类任务）：一般使用精确率、召回率、F1 Score
  - 机器翻译，文本摘要（生成任务）：一般使用BLEU，ROUGE

F1 Score

- 文本分类任务的常用评价指标是F1 Score
  - F1 Score 是一种综合了精确率（Precision）和召回率（Recall）的指标，用于评估分类模型的性能。
  - F1 Score的值介于0和1之间，值越大表示模型性能越好。
- 精确率是预测为正类的样本中实际为正类的比例
- 召回率是实际为正类的样本中被正确预测为正类的比例

F1 Score的计算公式

- 精确率（Precision） $$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$
  - 其中，TP（True Positive）是被正确预测为正类的样本数，FP（False Positive）是被错误预测为正类的样本数。

- 召回率（Recall） $$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
  - 其中，FN（False Negative）是实际为正类但被错误预测为负类的样本数。

- F1 Score $$F1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$$

F1 Score 计算示例

- 假设我们有一个文本分类任务，需要将文本分为“垃圾邮件”（正类）和“非垃圾邮件”（负类）。测试集中有以下数据：

实际类别	预测类别	
正类（垃圾邮件）	正类（垃圾邮件）	TP
正类（垃圾邮件）	负类（非垃圾邮件）	FN
负类（非垃圾邮件）	正类（垃圾邮件）	FP
负类（非垃圾邮件）	负类（非垃圾邮件）	TN
负类（非垃圾邮件）	正类（垃圾邮件）	FP

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{1}{1+2} = 0.333$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{1}{1+1} = 0.5$$

$$F1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} = 2 \times \frac{0.333 \times 0.5}{0.333 + 0.5} \approx 0.4$$

多分类任务的F1 Score

- 已经介绍了二分类任务的F1 Score计算方法
- 对于多分类任务，可以通过以下方法计算得到系统的F1 Score：
  - 1. **宏平均（Macro Average）**：对每个类别的F1 Score 求平均，不考虑类别样本数量差异。
  - 2. **微平均（Micro Average）**：汇总每个类别的TP、FP、FN之后，计算全局的精确率、召回率和F1 Score。
  - 3. **加权平均（Weighted Average）**：根据每个类别的样本数量，对其F1 Score加权平均。

BLEU

- BLEU (Bilingual Evaluation Understudy) 是评价自然语言处理系统生成的文本的指标，最初用于评估机器翻译和一个或多个参考翻译之间的相似度[1]。
- BLEU的值介于0和1之间（在很多学术论文中会将计算结果乘100），分数越高代表机器翻译结果与参考翻译相似度越高，机器翻译系统的性能越好。
- BLEU还可以用于评价文本摘要、图像描述等任务。

[1] BLEU: a Method for Automatic Evaluation of Machine Translation

BLEU

- 在提出Transformer模型的论文“Attention Is All You Need”中汇报了模型在EN-DE（英-德）和EN-FR（英-法）翻译语向的BLEU。

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

N-gram

- 在介绍BLEU的计算方法前，首先介绍N-gram的概念。N-gram是自然语言处理中重要的概念，表示一组连续的N个词，例如
  - *The cat is on the mat*
  - 1-gram: “The”, “cat”, “is”, “on”, “the”, “mat”
  - 2-gram: “The cat”, “cat is”, “is on”, “on the”, “the mat”
  - 3-gram: “The cat is”, “cat is on”, “is on the”, “on the mat”
  - 4-gram: “The cat is on”, “cat is on the”, “is on the mat”

BLEU 计算公式

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log p_n\right)$$

- BP（Brevity Penalty，长度惩罚）：用来惩罚翻译句子过短的情况，定义为

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \le r \end{cases}$$

  - c 是机器翻译的长度， r 是参考翻译的长度。
- pn：第n-gram 的精确率，定义为：

$$p_n = \frac{\text{机器翻译中与参考翻译匹配的 } n\text{-gram 个数}}{\text{机器翻译中的总 } n\text{-gram 个数}}$$

- N：表示使用的最大 n-gram 数，通常 N=4。

BLEU 计算步骤

- 1. 分词：将机器翻译和参考翻译的句子分成单词或短语。
- 2. n-gram 匹配：统计机器翻译中的 n-gram 与参考翻译中 n-gram 的匹配个数。
- 3. 计算每个 n-gram 的精确率 pn 。
- 4. 计算长度惩罚（BP）。
- 5. 结合公式，计算 BLEU 值。

BLEU 计算示例

- 输入示例：
  - **机器翻译：** *The cat is on the mat*
  - **参考翻译：** *The cat sat on the mat*
- 分词：
  - 机器翻译：["The", "cat", "is", "on", "the", "mat"]
  - 参考翻译：["The", "cat", "sat", "on", "the", "mat"]

BLEU 计算示例

- 1-gram 匹配：
  - [“The”, “cat”, “on”, “the”, “mat”]（共 5 个），计算得到 p1=5/6
- 2-gram 匹配：
  - [“The cat”, “on the”, “the mat”]（共 3 个），计算得到 p2=3/5
- 3-gram 匹配：
  - [“on the mat”]（共 1 个），计算得到 p3=1/4
- 4-gram 匹配：
  - 无匹配，计算得到 p4=0/3= 0

BLEU 计算示例

- 计算BP
  - 机器翻译长度 c = 6，参考翻译长度 r = 6。
  - BP=1 （因为 c = r ）。
- 计算BLEU

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\frac{1}{4}(\log p_1 + \log p_2 + \log p_3 + \log p_4)\right)$$

$$\mathrm{BLEU} = 1 \cdot \exp\left(\frac{1}{4}(\log \frac{5}{6} + \log \frac{3}{5} + \log \frac{1}{4} + \log 0)\right)$$

BLEU 计算示例

- 由于4-gram匹配为0，上式出现了log0的情况，导致最终BLEU为0。
  - 这种情况并不合理，一般需要使用平滑方法。
- 使用Sacrebleu工具包[1]中默认的exp方法
  - 将p4替换为 $\frac{1}{2 \times \text{num}_{\text{n-gram}}} = \frac{1}{2 \times 3} = 0.167$

$$\begin{aligned}\text{BLEU} &= 1 \cdot \exp\left(\frac{1}{4}(\log\frac{5}{6} + \log\frac{3}{5} + \log\frac{1}{4} + \log\frac{1}{6})\right) \\ &= 0.38\end{aligned}$$

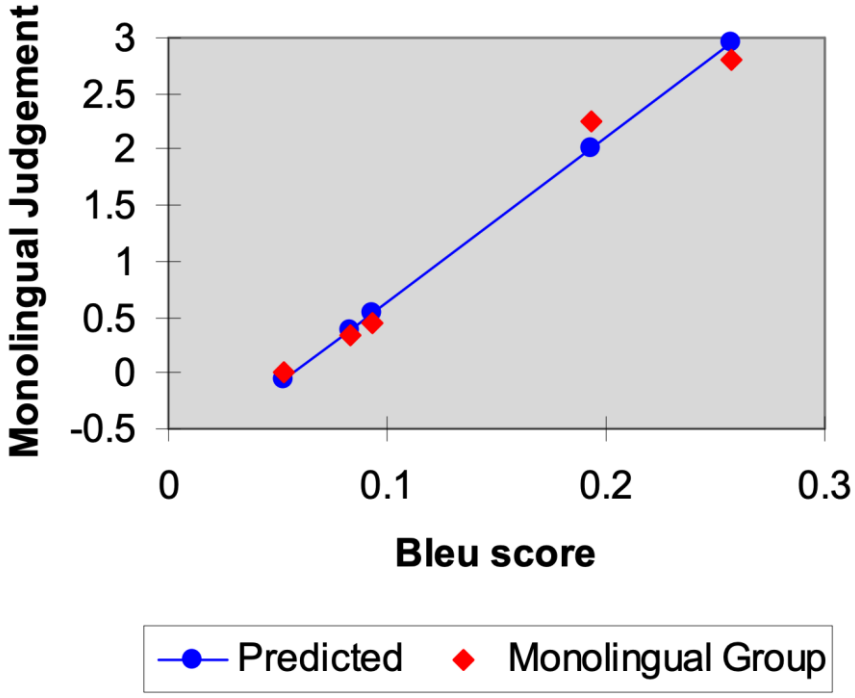
[1] Sacrebleu是一个BLEU评测工具包，常用于学术论文和国际评测。

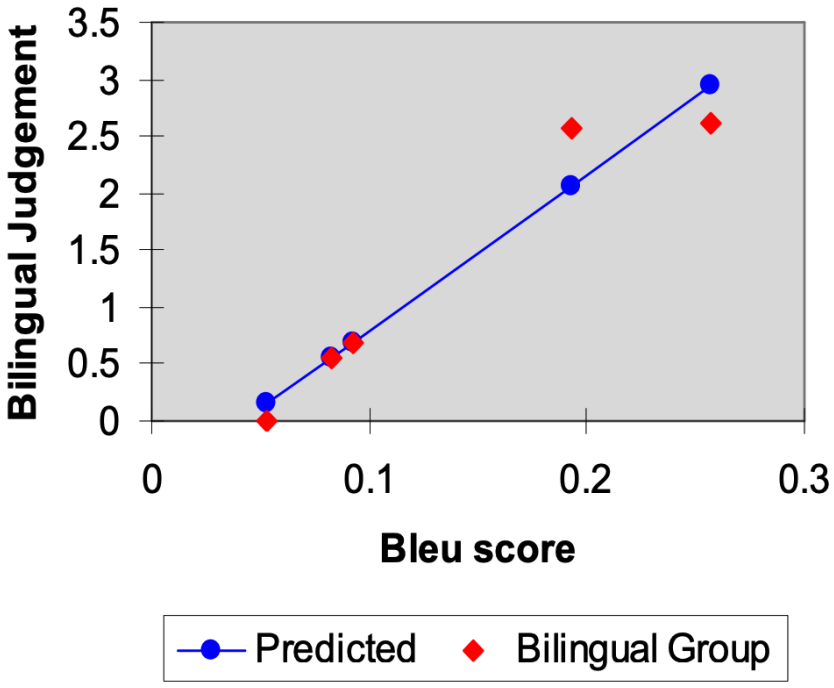
BLEU 为什么需要计算BP

- 为什么需要引入长度惩罚（BP）？
  - **过短的翻译容易匹配较多的** n-grams
  - 如果没有长度惩罚，机器翻译系统可能倾向于生成非常短的翻译，甚至仅生成几个高频词（如“the”或“is”）。可能在 n-gram 匹配上得分较高，但这些翻译通常无法传达完整的语义。例如：
    - **参考翻译：** The cat is on the mat.
    - **机器翻译：** The cat.
    - 1-gram 匹配率： p1 = 2/2 = 1 （完全匹配）。
    - 但这种翻译过于简略，无法准确表达原文意思。

BLEU为什么可以评价机器翻译系统

- 如何说明一个评价指标效果“好”？
  - 怎样评价一个评价指标？
  - 计算评价指标和人类评价的相关性。





ROUGE

- ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是用于评估自动文本摘要、机器翻译和文本生成任务的评价指标，它主要用于衡量机器生成文本与参考文本的相似程度。
- 与 BLEU 偏向精确率不同，ROUGE 更加关注 **召回率**，因此它特别适合摘要生成任务，因为摘要通常希望涵盖更多参考文本的关键信息。

ROUGE计算

- ROUGE-N
  - ROUGE-N 是基于 n-gram 的指标，计算生成文本与参考文本之间共有的 n-grams 的比例。
  - n 通常取 1（ROUGE-1，单词级匹配）和 2（ROUGE-2，二元短语匹配）。
- ROUGE-L (Longest Common Subsequence, LCS)
  - ROUGE-L 基于最长公共子序列（LCS）的概念，衡量生成文本与参考文本之间的语序相似性。

ROUGE 计算示例

- **参考文本**: *The cat is on the mat.*
- **生成文本**: *The cat sat on the mat.*
  - ROUGE-1 (**单词级匹配**):
  - 参考文本的单词集合：{The, cat, is, on, the, mat}
  - 生成文本的单词集合：{The, cat, sat, on, the, mat}
  - 匹配的单词数：5（The, cat, on, the, mat）

$$\text{ROUGE-1} = \frac{\text{匹配单词数}}{\text{参考文本单词总数}} = \frac{5}{6} \approx 0.833$$

ROUGE 计算示例

- **参考文本**: *The cat is on the mat.*
- **生成文本**: *The cat sat on the mat.*
  - ROUGE-L（**最长公共子序列**）
  - 最长公共子序列（LCS）：The cat on the mat（长度为 5）。

$$\text{ROUGE-L (Recall)} = \frac{\text{LCS 长度}}{\text{参考文本长度}} = \frac{5}{6} \approx 0.833$$

第三节

评价方法

评价方法

- F1 Score、BLEU、ROUGE都是基于参考答案的评价方法
  - Reference based，根据给定的参考答案比对模型的生成结果。
  - 这种评价方法有什么问题？

BLEU 指标的问题

- **1. 缺乏语义理解**
  - BLEU 本质上是基于 n-gram 的精确匹配，它只关注生成文本与参考文本在表面词序上的相似性。
  - 它无法捕捉语义相似性，即即使生成的文本与参考文本意义完全相同，但用词不同，BLEU 分数可能仍然很低。

BLEU 指标的问题

- **2. 过于依赖参考答案**
  - BLEU 分数依赖于预先标注的参考文本。如果参考答案过少或不够多样化，模型的表现可能被低估。
  - 在开放生成任务中（如对话或故事生成），可能存在多种正确答案，但 BLEU 难以体现这种灵活性。

BLEU指标的问题

- **3. 不区分重要性**
  - BLEU 对每个 n-gram 给予了相同的权重，无法区分哪些词语对句子意义更为关键。
  - 例如，对于描述性文本，缺少关键动词可能比缺少一个次要形容词更严重，但 BLEU 无法反映这一点。

BLEU指标的问题

- **4. 对多样性和创新的惩罚**
  - 生成任务中，模型可能输出创造性内容，但如果与参考文本没有显式重叠，分数会被低估。
  - 这在对话、创意写作或开放性问题回答中尤为突出。
- **5. 缺乏对语义和人类偏好的反映**
  - 无法完全反映人类的主观评价，例如流畅性、连贯性、信息性等关键指标。

BLEU指标的问题

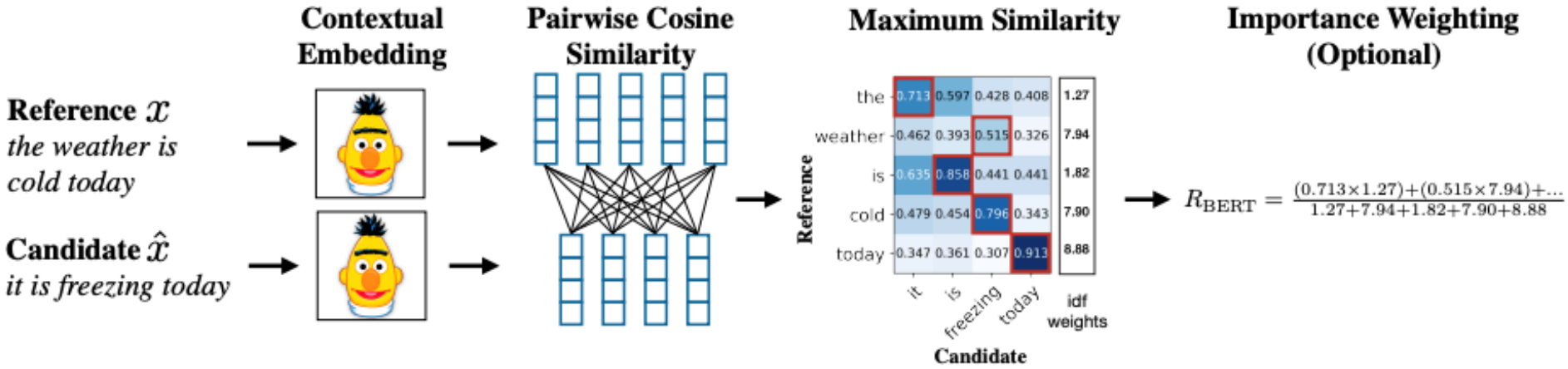
- 导致BLEU存在上述问题的关键之处在于：

BLEU的计算只是最基础的字符串匹配。

  - 怎样改进？
  - 如何引入语义信息？
  - 不使用参考译文进行比对？

引入语义信息

- BERTScore[1]
  - 计算两个句子使用预训练模型（如BERT，RoBERTa）得到的嵌入向量的余弦相似度。
  - 实验结果表明BERTScore和人类评估结果的相关性更高。



[1] BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

典型评价指标的改进方法

- BLEURT[1]
  - 输入机器翻译结果，参考译文至BERT。取BERT的[CLS]位置作为表示向量

$$\boldsymbol{v}_{[\mathrm{CLS}]}, \boldsymbol{v}_{x_1}, \ldots, \boldsymbol{v}_{x_r}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\tilde{x}_p} = \mathrm{BERT}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

  - 在该向量后连接一个线性层，用于输出分数

$$\hat{y} = \boldsymbol{f}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \boldsymbol{W}\tilde{\boldsymbol{v}}_{[\mathrm{CLS}]} + \boldsymbol{b}$$

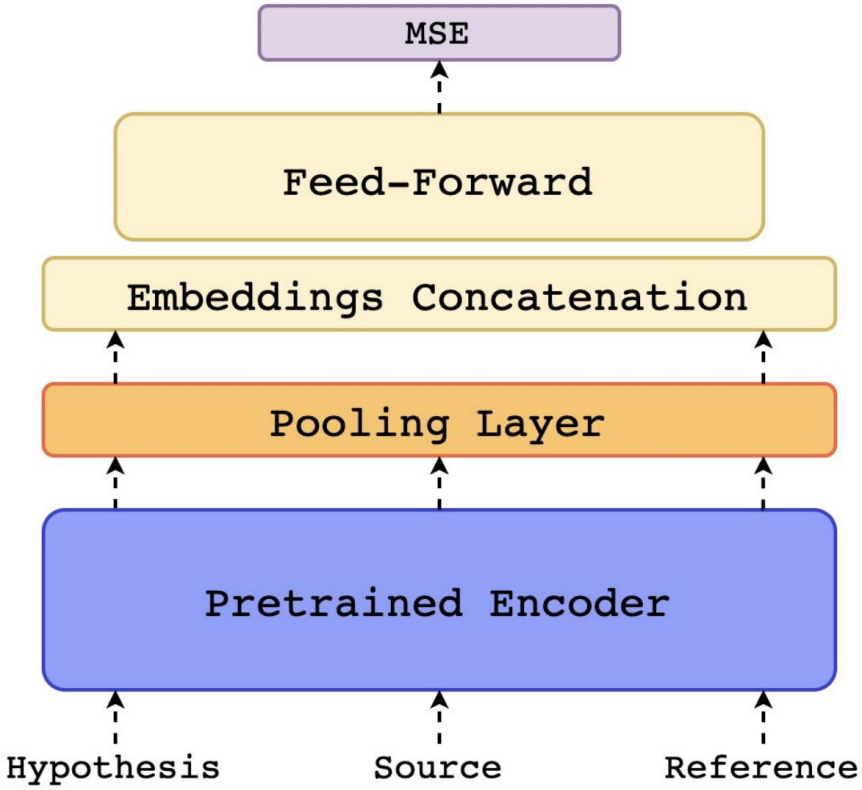
  - 使用人工评估机器翻译结果的分数微调BERT

$$\ell_{\text{supervised}} = \frac{1}{N}\sum_{n=1}^{N} \|y_i - \hat{y}\|^2$$

[1] BLEURT: Learning Robust Metrics for Text Generation

典型评价指标的改进方法

- COMET[1]
  - 输入机器翻译结果、原文本、参考译文，输出评价分数。
  - 使用预训练模型（BERT, XLM, XLM-RoBERTa）不同层的表示。
  - 需要带有翻译结果评估分数的数据集进行微调。



[1] COMET: A Neural Framework for MT Evaluation

基于参考答案评价的局限性

- 引入语义信息后，上述方法仍然需要基于参考答案进行评价。
  - 依赖参考文本比对语义相似性，对于更复杂的目标（流畅性、事实性、风格等）难以评估。
  - 开放领域任务，可能存在多种合理答案，这些指标可能无法考虑生成内容的创新性，不利于评估模型的多样性和创造性。
- 不基于参考答案的评价方法
  - Reference free，不需要给定参考答案，就可以对模型生成结果进行打分。

不基于参考答案的评价方法

- 使用预训练的语言模型计算生成文本的困惑度（Perplexity，PPL）。
  - 评估生成文本的流畅性。困惑度越低，生成文本越流畅。
  - 无法直接评估语义准确性或任务相关性。
- 计算模型生成的所有文本之间的BLEU（self-BLEU）。
  - 评估生成文本的多样性。分数越低，生成文本的多样性越高。
  - 无法直接评估语义准确性或任务相关性。

不基于参考答案的评价方法

- 不基于参考答案对自然语言处理系统进行评价比较困难。
  - 开放式任务很难给出标准的参考答案，如何不基于参考答案对这类任务进行评价？
- 引入人工对自然语言处理系统进行评价。
  - 人工评价
  - 通过人工根据不同维度（如流畅性、连贯性、事实性等）进行评分。

人工评价方法

- 适用场景：
  - 需要高准确度评价时，作为自动评价的补充。
  - 开放式任务（如开放对话、故事生成）难以用规则量化时。
- 评价维度：
  - **流畅性**：语言表达是否自然，句法是否正确。
  - **连贯性**：上下文是否逻辑通顺、衔接流畅。
  - **信息性**：是否提供足够的信息量。
  - **事实性**：生成的内容是否准确。
  - **风格适配性**：生成内容是否符合目标风格。

人工评价方法

- 人工评价方法的优点：
  - 贴近真实用户体验。
  - 可对生成文本进行多维度、细粒度的评价。
  - 针对不同任务和需求定制化评价方式。
- 人工评价方法的缺点：
  - 费时费力，成本高。
  - 主观性强，不同评审者之间可能存在偏差。

缓解人工评价方法的缺点

- 人工评价成本高且耗时：
  - 利用自动化工具或模型生成初步结果，缩小需要人工验证的范围。
  - 先用自动评价方法筛选待评价文本，再用人工重点检查质量较高或问题突出的文本。
- 人工评价主观性强：
  - 增加评估者人数，缓解个体偏差。
  - 制定详细的评分标准和评价维度，确保评估者对评价内容有一致的理解。
  - 使用标准化的评分表格或评价模板，减少因评估者理解差异造成的偏差。

人工评价

- 对于自动评价指标难以评估的任务或系统，尝试引入人工评价。
  - 费时费力，成本高
  - 主观性强

- 由于大模型（如ChatGPT）有很强的理解能力，可以引入大模型对系统进行评价。

自然语言处理系统评价示例

- 机器翻译任务
  - 原文：Transformer models have revolutionized NLP by enabling efficient parallel computation and contextual understanding.
  - 参考译文：Transformer模型通过高效的并行计算和上下文理解，彻底改变了自然语言处理领域。
  - 系统A：Transformer模型通过并行计算和上下文分析，改变了自然语言处理。
  - 系统B：Transformer模型已经通过高效的平行计算和上下文理解革新了自然语言处理。

自然语言处理系统评价示例

- 分析
  - 系统A：Transformer模型通过并行计算和上下文分析，改变了自然语言处理。
    - **优势**：翻译句式流畅，较符合一般中文表达习惯。
    - **劣势**：术语不够专业（如“上下文分析”不完全等同于“上下文理解”），导致语义信息有丢失。
  - 系统B：Transformer模型已经通过高效的平行计算和上下文理解革新了自然语言处理。
    - **优势**：术语准确，完整保留了原文中的关键信息（如“高效的平行计算”和“上下文理解”）。
    - **劣势**：翻译风格偏直译，句子显得僵硬，少了自然流畅感。

自然语言处理系统评价示例

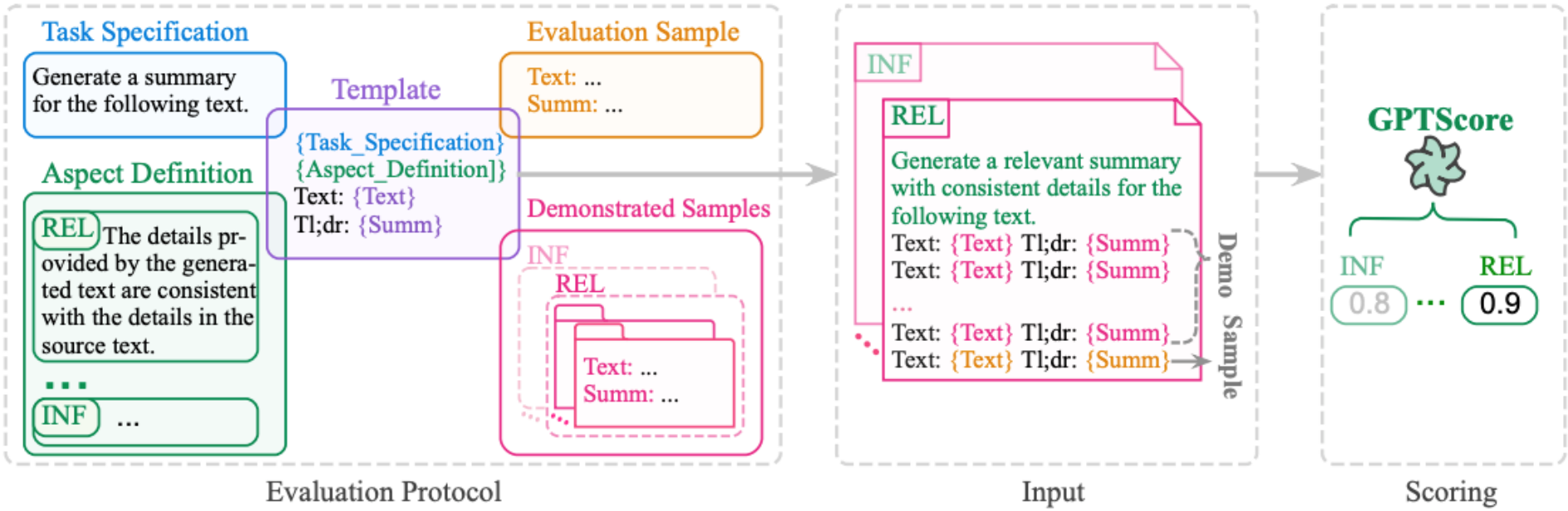
- 自动评价（BLEU）
  - **系统A**：54.8。**系统B**：55.9。
- 人工评价
  - **系统 A**：术语准确度一般，“上下文分析”含义偏弱但流畅，得分 7/10。**系统 B**：术语精准且语义完整，但句式生硬，得分 6/10。
- 大模型评价
  - **系统 A**：翻译流畅但术语不够专业，得分 7.5/10。**系统 B**：语义完整，术语准确但不够自然，得分 8/10。

第四节

基于大模型的自动评价

使用大模型的自动评价

- GPTScore [1]



[1] GPTScore: Evaluate as You Desire

使用大模型的自动评价

- 大模型评价有更全面的语义理解
  - 由于大模型训练数据多样化、规模庞大，无需领域内数据微调。可以直接用于评估生成答案是否合理，符合语义。
- 对生成多样性容忍度更高
  - 对于开放式任务，大模型不会因为答案形式与参考文本不同而低估质量。
- 可以从多个角度进行综合评价
  - 针对流畅性、连贯性、事实性、风格等角度进行评价

使用大模型的自动评价

- 大模型具有丰富的世界知识，可以基于背景信息评估生成内容
  - 可以检测出事实性错误
- 支持更灵活的评价方式
  - 大模型具有很强的理解能力，可以基于任务需求定制化评价
- 可交互的评价方式
  - 研究人员在使用大模型评价后，可以让其解释为什么某个生成内容得分低，并提出改进的建议

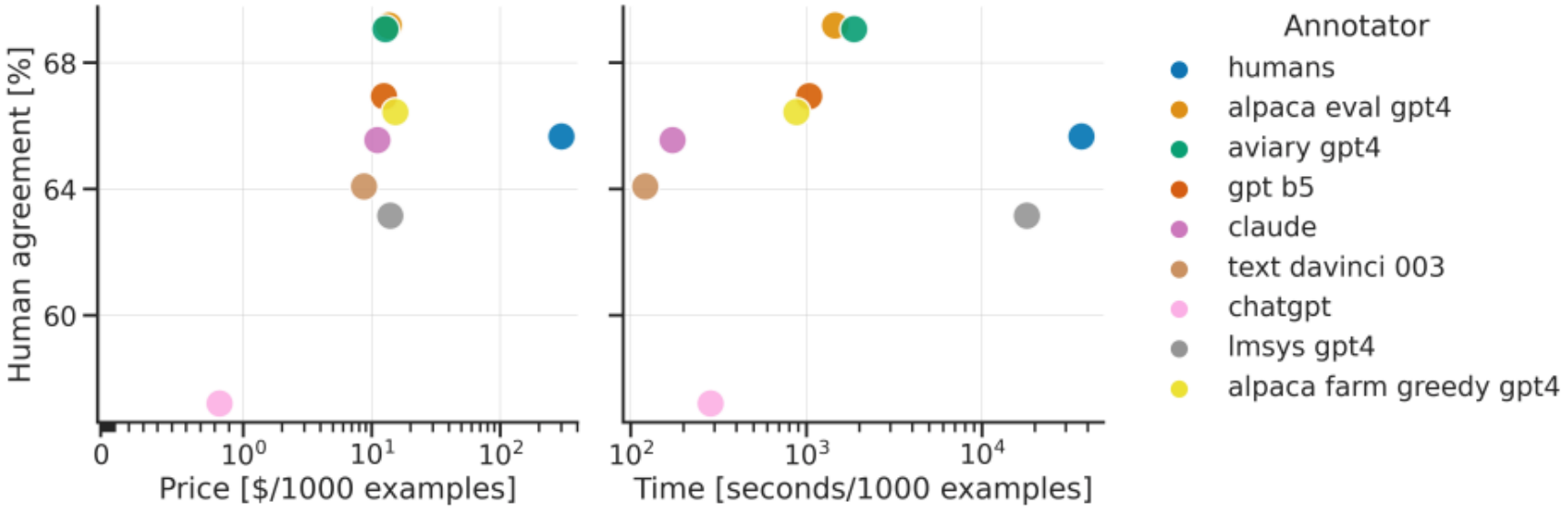
使用大模型的自动评价

- 针对摘要、对话、机器翻译等任务都有不同的评价维度。

Aspect	Task	Definition
Semantic Coverage (COV)	Summ	How many semantic content units from the reference text are covered by the generated text?
Factuality (FAC)	Summ	Does the generated text preserve the factual statements of the source text?
Consistency (CON)	Summ, Diag	Is the generated text consistent in the information it provides?
Informativeness (INF)	Summ, D2T, Diag	How well does the generated text capture the key ideas of its source text?
Coherence (COH)	Summ, Diag	How much does the generated text make sense?
Relevance (REL)	Diag, Summ, D2T	How well is the generated text relevant to its source text?
Fluency (FLU)	Diag, Summ, D2T, MT	Is the generated text well-written and grammatical?
Accuracy (ACC)	MT	Are there inaccuracies, missing, or unfactual content in the generated text?
Multidimensional Quality Metrics (MQM)	MT	How is the overall quality of the generated text?
Interest (INT)	Diag	Is the generated text interesting?
Engagement (ENG)	Diag	Is the generated text engaging?
Specific (SPE)	Diag	Is the generated text generic or specific to the source text?

使用大模型的自动评价

- 在AlpacaFarm中，作者让大模型模拟人类进行标注打分。
- 相较于人工评价，大模型的自动评价更便宜，耗时更少，并且有更高的一致性。

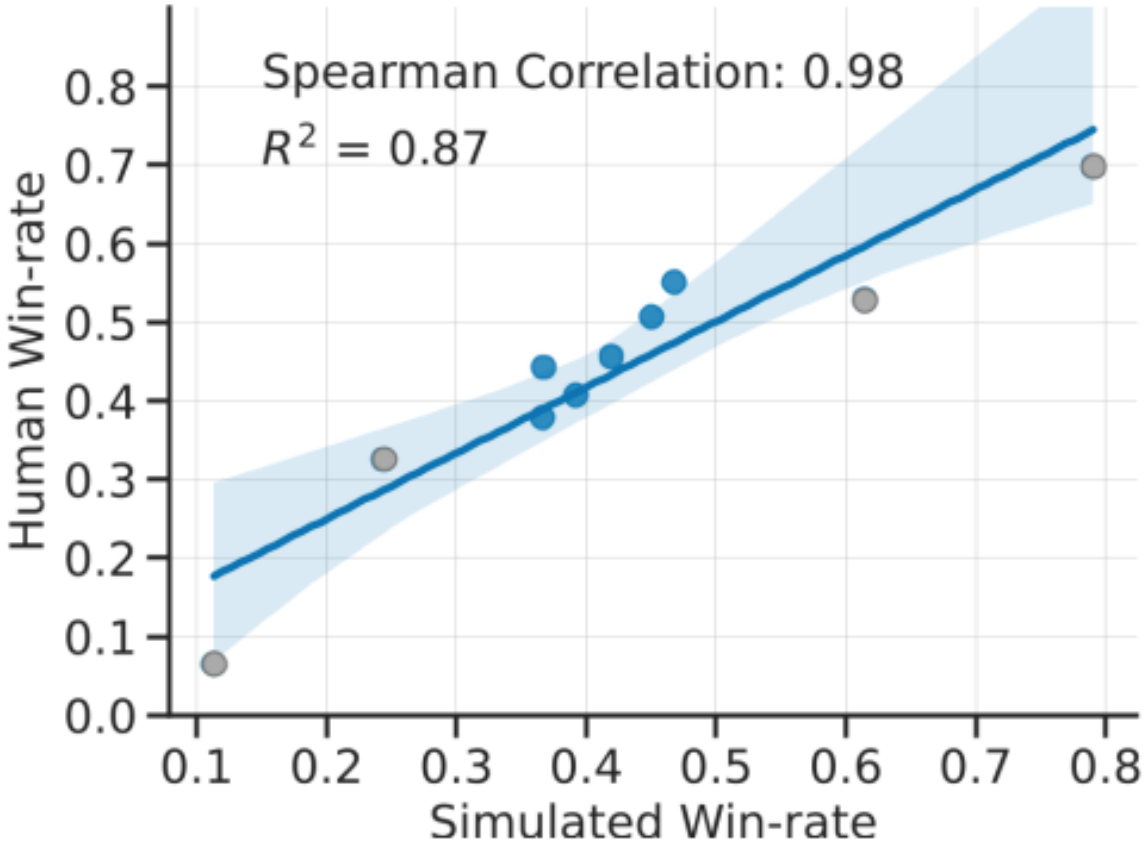


使用大模型的自动评价

- 将带有大模型评分的数据用于RLHF
  - RLHF（Reinforcement Learning from Human Feedback），即**基于人类反馈的强化学习**，是一种结合强化学习和人类反馈的技术，常用于训练大模型使其生成结果更符合人类的偏好和需求。
  - RLHF需要带有对一段文本的人类打分数据，或多段文本的排序数据。
  - 大模型打分能否替代人类打分？RLHF→RLAIF

使用大模型的自动评价

- 使用大模型模拟评分标注和人类评分标注的数据，训练结果非常相近。



如何使用大模型自动评价

- 使用大模型（如ChatGPT，GPT-4o等）自动评估，需要设计合理的prompt，一般需要包括以下方面：
  - 描述需要评价的任务，评价的角度
  - 针对不同的评价角度设计不同的分数及其对应评判标准
  - 大模型评价的输出格式

Prompt for ChatGPT Scoring (Model: GPT-4o)

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output. Your task is to rate the model's responses based on the provided user input transcription [Instruction] and the model's output transcription [Response]. Please evaluate the response from two perspectives: content and style, and provide a score for each on a scale of 1 to 5.

Content (1-5 points):
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that don't contribute to the main point.
4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.

Style (1-5 points):
1 point: The response is poorly suited for speech interaction, possibly including structured elements like lists or being overly complex, disjointed, or difficult to understand.
2 points: The response is somewhat suitable but may be too long, too short, or awkwardly phrased, making it less effective in a speech interaction context.
3 points: The response is generally suitable for speech interaction, but it may have minor issues with length, clarity, or fluency that detract slightly from the overall effectiveness.
4 points: The response is well-suited for speech interaction, with appropriate length, clear language, and a natural flow. It is easy to understand when spoken aloud.
5 points: The response is perfectly suited for speech interaction. It is the ideal length, highly clear, and flows naturally, making it easy to follow and understand when spoken.

Below are the transcription of user's instruction and models' response:
[Instruction]: **{instruction}**
[Response]: **{response}**

After evaluating, please output the scores in JSON format: {"content": content score, "style": style score}. You don't need to provide any explanations.

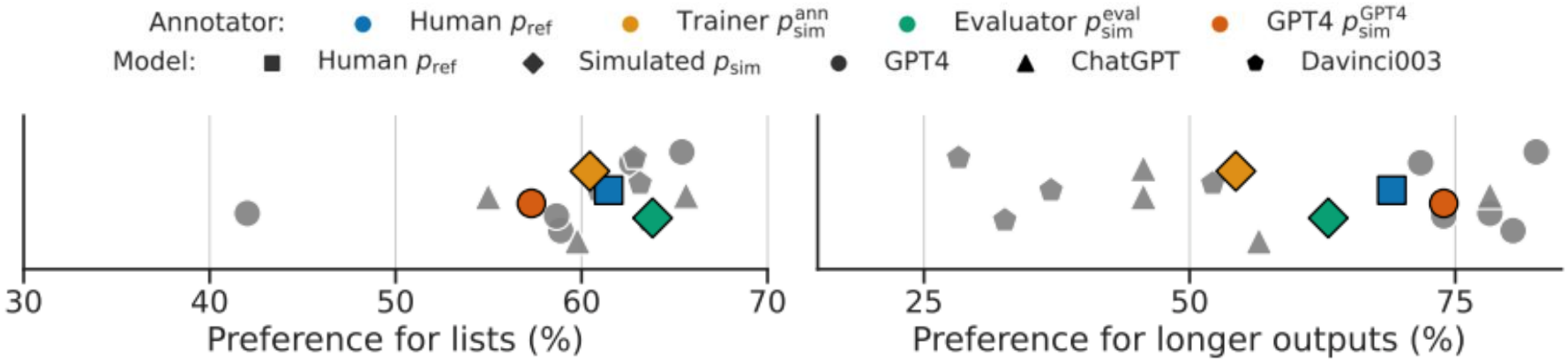
[1] LLAMA-OMNI: SEAMLESS SPEECH INTERACTION WITH LARGE LANGUAGE MODELS

大模型自动评价的局限性

- 缺乏客观性
  - 大模型的评估结果有一定主观性，因为并不是根据固定的算法或规则进行评分。
- 难以制定统一的标准
  - 大模型支持定制化的评价标准，但是这也表明每次评价的基准可能不同。

大模型自动评价的局限性

- 评价存在偏好
  - 大模型存在自我偏好，对于自己生成的文本内容会给予更好的评价[1]。
  - 存在排序偏差，交换两个待评价文本的前后顺序会导致评价出现偏差[1]。
  - 在AlpacaFarm的实验中，大模型偏好包含列表，文本更长的输出。



[1] LLM Evaluators Recognize and Favor Their Own Generations

思考题

- 大模型自动评价还有哪些局限性？
- 如何缓解大模型自动评价中的局限性？
- 能否把典型的自动评价指标和大模型自动评价相结合？

谢谢各位！

Q&A