7. 大语言模型的科学评估与对齐

邀请了CMU李磊教授，探讨大型语言模型（LLMs）是否能够公平地评估和改进模型生成的表现，分析这些模型是否真正掌握了特定知识，还是仅仅凭运气回答问题。课程内容包括展示评估LLMs在知识密集和语言生成任务上的科学方法，观察使用LLMs作为评估者时出现的自偏现象，并讨论后期训练方法，以更好地优化和调整LLMs，使其与人类判断和评价更加一致。本课程旨在通过科学手段，实现对LLMs的客观有效评估和优化。

2024版本：李磊、唐杰、杜晋华、罗开荣、潘至璇、江楚萌、陈俊杰

## 引言

ChatGPT、LLaMA和Claude等大语言模型，能够基于概率序列预测文本，展示出了强大的语言理解和生成能力，已经彻底改变了我们的生产方式。这些模型能够高效地完成诸如翻译、摘要、内容编辑和创意生成等任务，在代码生成、回答复杂问题甚至推荐餐馆等场景中也表现出色。

然而，如何科学地评估LLM生成的内容质量仍然是一个巨大的挑战。评估方法通常需要结合基于参考的指标（如BLEU和ROUGE）和基于源的评估（如质量估计模型），前者依赖于已知的标准答案进行比较，而后者则无需明确的参考答案即可评估输出质量。从研究趋势来看，评估方法正在从规则驱动的指标发展为学习驱动的方法，包括有监督和无监督的模型，例如BERTScore、BLEURT和COMET等。然而，这些方法也面临诸多限制，例如过于关注表层形式差异或者需要大量人工标注数据的支持。

此外，使LLM输出与人类期望对齐的过程也至关重要。通过强化学习与奖励模型的结合，模型能够不断优化其生成能力。一种典型的奖励机制框架是，通过比较和排序不同的输出，结合人类反馈，来逐步提高模型的质量生成能力。然而，在评估LLM时，不仅需要应对传统的指标，还需要应对开放式任务、多维目标以及缺乏明确答案的场景，这些挑战仍然有待人们探索和研究。

本章将探讨LLM评估与对齐背后的科学方法，对评估LLM生成质量面临的挑战进行深入探讨，分析LLM作为评价者的可信度、可解释性文本生成评价、LLM中的知识评估等问题，并对LLM后训练对齐的一些前沿方法进行概述，如在线行为偏好优化、迭代细化与精细化反馈等。

## 主讲介绍



李磊教授是卡内基梅隆大学（Carnegie Mellon University）语言技术研究所（Language Technologies Institute）的助理教授。他的研究兴趣包括机器翻译、可信生成式人工智能、智能体型大语言模型（agentic LLM）以及人工智能驱动的药物研发。他获得了上海交通大学学士学位和卡内基梅隆大学博士学位。李磊曾获2021年ACL最佳论文奖、2019年CCF青年精英奖、2017年CCF杰出演讲者、2017年吴文俊人工智能奖，以及2012年ACM SIGKDD博士论文奖（亚军）。他在ICLR 2023中被评为杰出领域主席（Notable Area Chair）。他的团队在WMT 2020机器翻译竞赛中赢得了五个翻译方向的第一名以及最佳语料过滤挑战奖。

此前，李磊曾担任加州大学圣塔芭芭拉分校的终身副教授（Associate Professor, Tenured）、字节跳动AI实验室创始主任、百度首席科学家以及加州大学伯克利分校的博士后研究员。他是IEEE TPAMI和TMLR的副主编，并在ACL、EMNLP、NAACL、NeurIPS、ICML、ICLR、KDD、AAAI、IJCAI、WSDM和CIKM等多个顶级会议中担任组织者和领域主席/资深程序委员会成员。他曾领导开发了字节跳动的机器翻译系统（VolcTrans）和小明机器人（Xiaomingbot）自动写作系统，其多项算法已转化为产品（今日头条、抖音、TikTok、飞书），服务全球超过十亿用户。

李磊曾在ACL 2024、NAACL 2024、CCF ADL 2022、ACL 2021、EMNLP 2019、NLPCC 2019、NLPCC 2016和KDD 2010等国际会议上主讲八场教程，并于2014年担任在美国俄勒冈州波特兰举办的“概率编程促进机器学习暑期学校”（Probabilistic Programming for Advancing Machine Learning Summer School）讲师。

## 目录

### LLM作为评估者

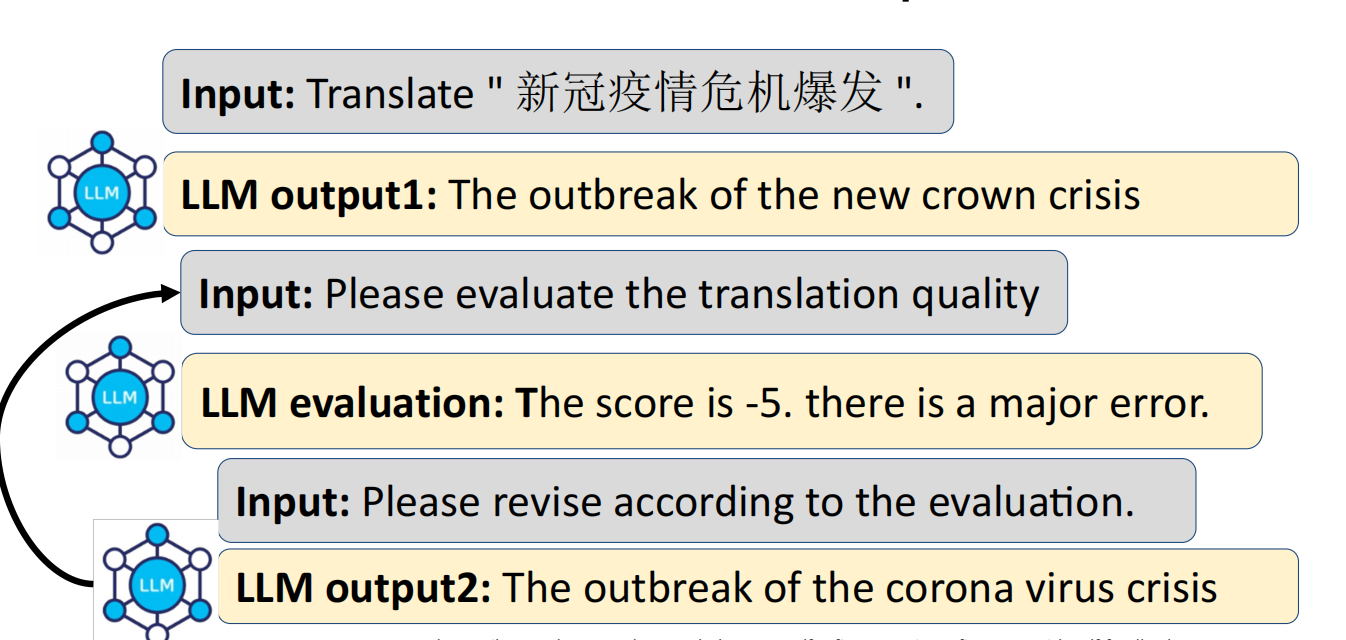
### 评估 LLM 生成质量

### 后训练对齐

## 7.1 LLM作为评估者

本小节探讨了大语言模型（LLM）在自我评估与优化中的表现和局限性。我们先简单地介绍LLM自我评估与优化的流程，再介绍其中的几点主要洞察。

### 7.1.1 LLM的自我评估与优化

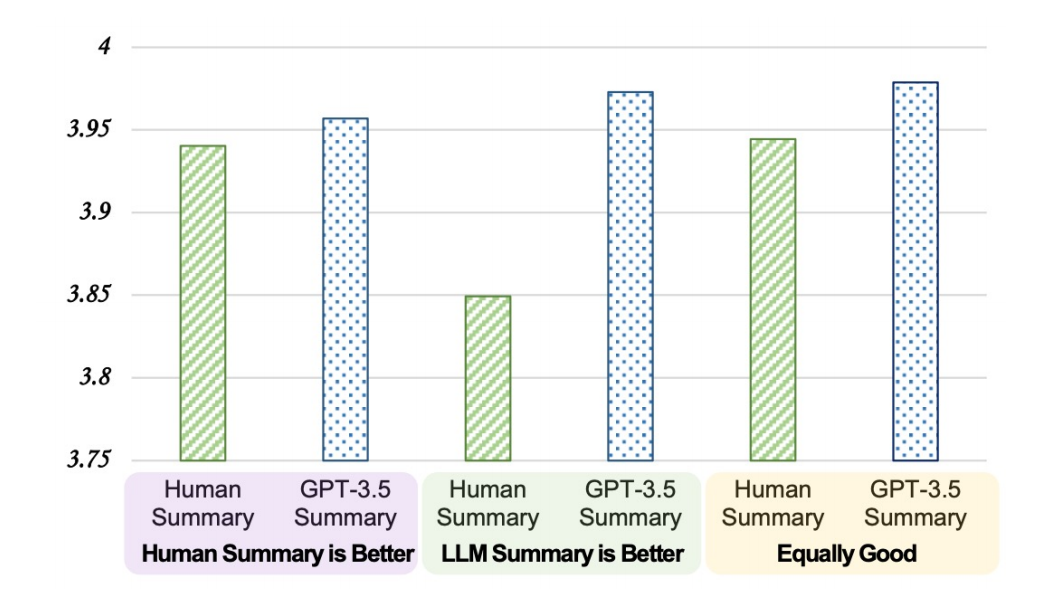


如图展示的是一个LLM自我评估与优化的实际例子，包含以下三个步骤：

1. **执行翻译任务**：输入句子“新冠疫情危机爆发”，LLM 生成初始翻译结果 LLM output1，即“The outbreak of the new crown crisis”。
2. **翻译质量评估**：输入请求评估翻译质量，LLM 给出评价结果 LLM evaluation：评分为“-5”，指出翻译中存在重大错误。
3. **修正翻译，迭代优化**：根据评估反馈，输入请求进行修改，LLM 生成修正后的翻译结果 LLM output2，即“The outbreak of the coronavirus crisis”。

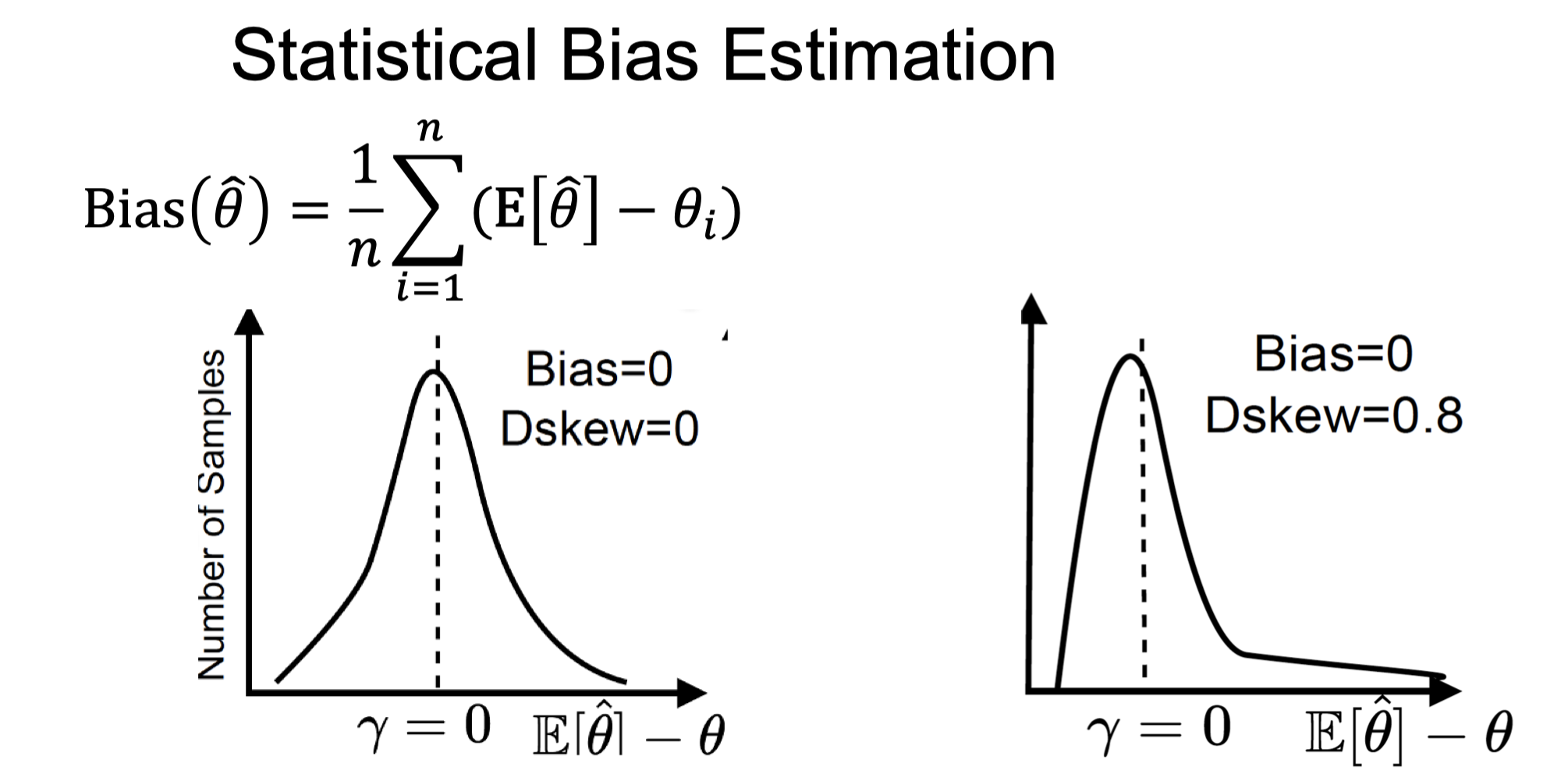
### 7.1.2 LLM评估者的自我偏差

研究表明，大语言模型（LLM）评估者表现出显著的自我偏差。如下图所示，GPT-3.5作为评估者对其生成的内容会给予更高的评分。



### 7.1.3 偏差的量化

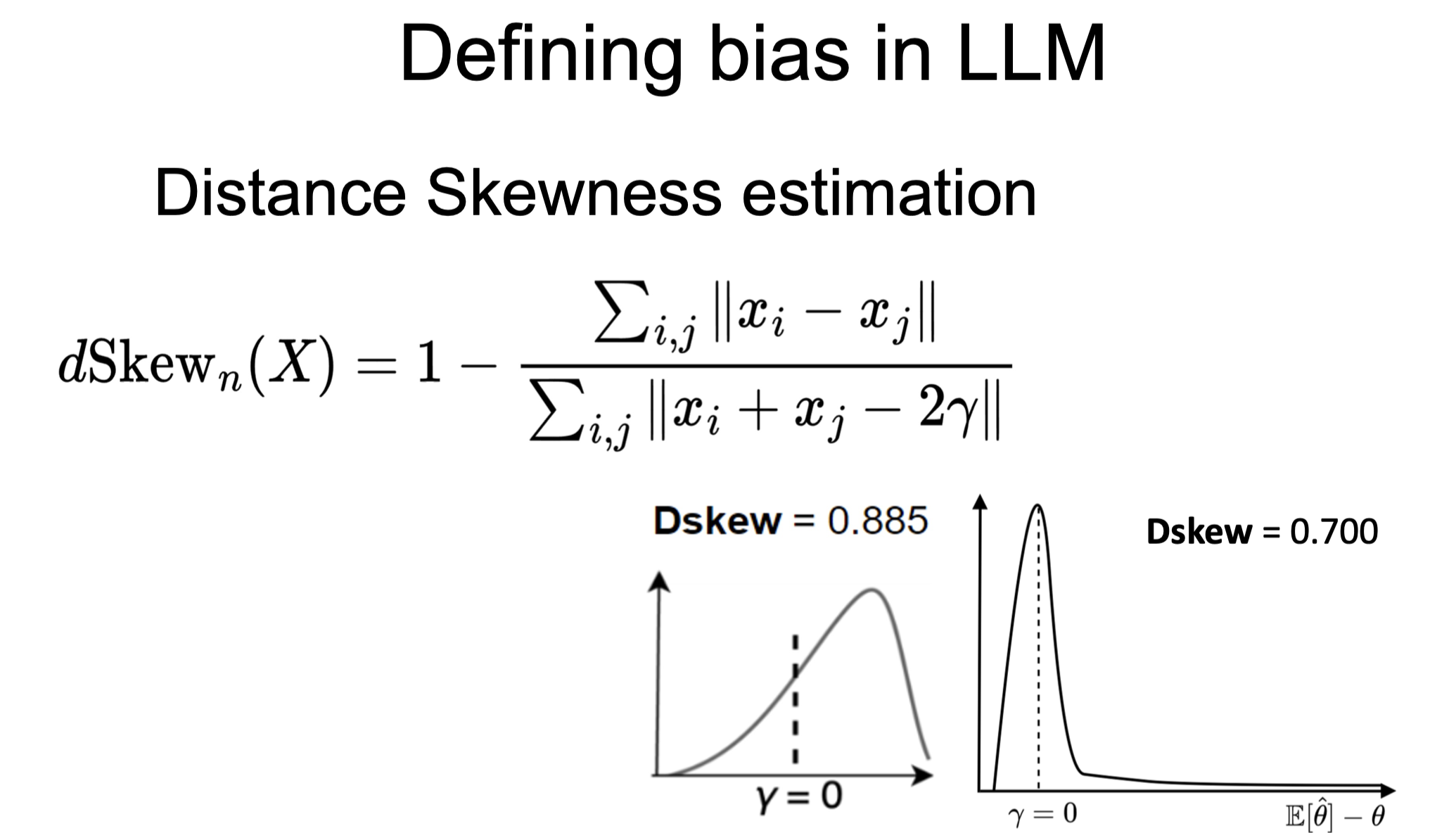
偏差的公式定义为估计值的期望与真实值之间的差异的平均值，用来衡量估计值的准确性。此外，图片提到一个指标“Dskew”（偏态），用于描述分布的对称性或偏斜程度。当偏态为0时，分布是完全对称的；当偏态不为0时，分布呈现一定的偏斜。通过不同的偏态值，图片展示了偏差为零但分布形状可能发生改变的情况。这表明，即使偏差为零，分布的形状也会对估计结果的可靠性产生影响。



为了在大语言模型（LLM）中定义偏差，研究者提出了一种称为“距离偏态估计”（Distance Skewness Estimation）的指标，用以量化数据分布的不对称性。公式定义为：

descript

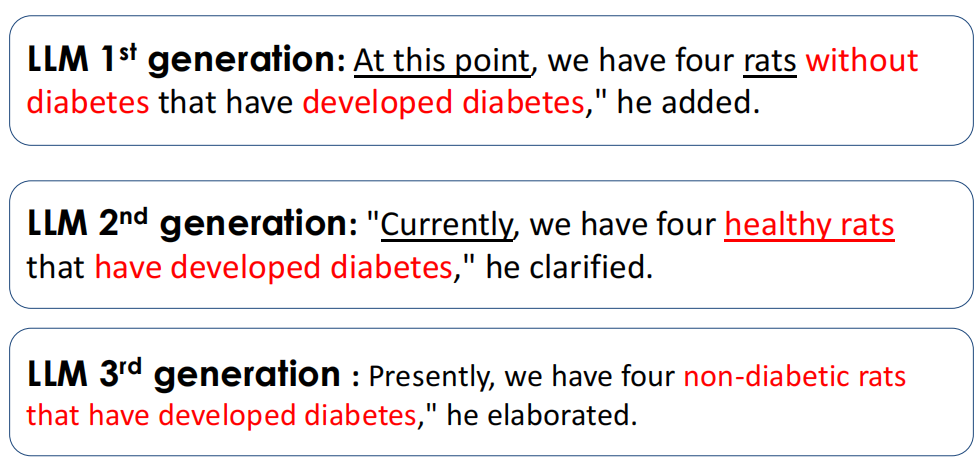
该公式通过比较样本点间距离与偏移后距离之间的比例，反映分布的不对称程度。



图中、展示了两个分布示例，分别对应不同的偏态值（Dskew）。当 Dskew = 0.885 时，分布具有较明显的偏斜，曲线向右倾斜；当 Dskew = 0.700 时，偏斜程度较小，但仍存在不对称性。这表明，偏态（Dskew）能够有效描述分布形状的变化，而不仅仅依赖均值或方差。

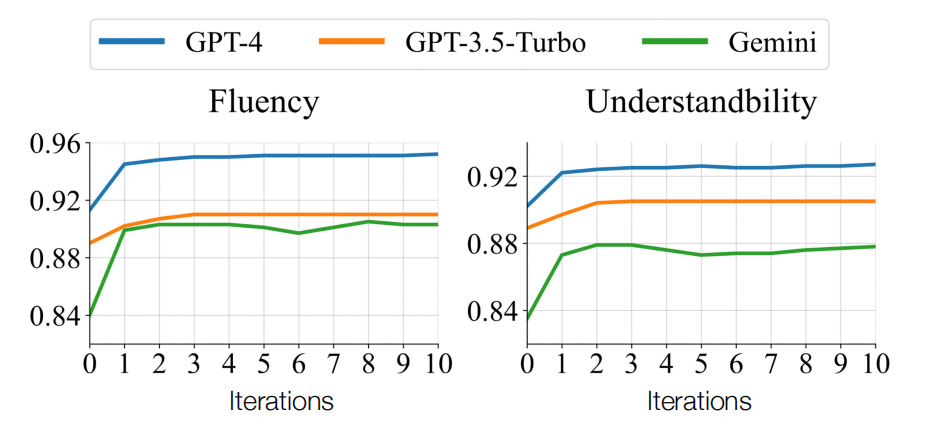
### 7.1.4 自我偏差的放大

这种自我偏差在自我优化和自我奖励过程中会被进一步放大。比如下面的例子，任务要求模型翻译Yoruba语"Ní bayii a ni àwon eku oloshu merin F ko ni dayabetesi telele to F ni ayabetesi,” o she afikun." （中文意思是“目前，我们有一些 4 个月大的小鼠没有糖尿病，但它们以前患有糖尿病。他说” ）。模型最初翻译为"At this point, we have four rats without diabetes that have developed diabetes," he added. 人类和模型对初始输出的评分均为-11，表明模型在初始阶段能够正确识别问题的严重性。然而，在后续的自我改进过程中，模型对句子进行了多次修改，试图改进内容。第一次修正后，句子变为"Currently, we have four healthy rats that have developed diabetes," he clarified，虽然语法更流畅，但引入了新的逻辑矛盾，即“健康的老鼠”无法同时“患上糖尿病”。此时，模型的自评分从-11提高到-10，显示出对自身改进的高估。第二次修正后，模型生成的句子为“目前，我们有四只非糖尿病老鼠患上了糖尿病"Presently, we have four non-diabetic rats that have developed diabetes," he elaborated，仍然未解决逻辑矛盾，但模型却将自评分进一步提升至0，完全忽略了逻辑错误的存在，而人类评分始终保持在-11。

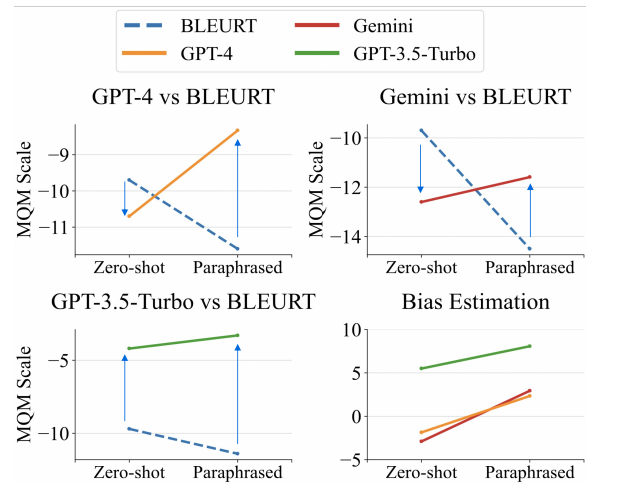


通过这些过程可以看出，LLM 的自我评估机制存在显著局限性，模型倾向于高估自己的改进效果，逐渐拉大与人类评价的差距。尽管句子表面流畅，但逻辑矛盾未被识别或解决，这反映了模型在逻辑理解和一致性检查能力上的不足。这些问题揭示了 LLM 自我改进机制的弱点，表明其在复杂逻辑问题上的表现仍需进一步优化，同时需要改进自我评估标准以更贴近人类认知，增强模型对逻辑和语义问题的识别能力。

**流畅性与质量**：那么如果自我优化不能提升模型生成的质量，那么在自我优化的过程中，什么东西是被改善了的呢？结论是：自我优化可以提升文本的流畅性，但不一定能改善整体质量。



**风格偏好**：另外，研究者发现了一种特定的自我偏差，称作风格偏好，即LLM倾向于偏好符合其自身“风格”的文本。



## 7.2 评估LLM生成质量

### 7.2.1 文本生成质量评估方法

在文本生成质量评估中，常用指标包括BLEU、BERTScore、COMET、BLEURT和SEScore2等。例如，在某翻译任务中，“COVID-19”被误译为“new crown”，导致得分仅为20/100。这类评估通常结合Transformer编码器（如BERT/XLMR）提取特征，利用人工评分或自动化指标细化解释。然而，这些指标的可解释性不强，即使知道文本生成质量不高也不能直接得到改进的方法。例如，GPT-4生成文本时，可能包含原文中不存在的信息，这类错误需通过细化解释加以标注。因此，我们希望得到可解释的文本生成评估指标。同时，我们希望可解释的评估指标满足：

* 与专家标注高度一致；
* 提供细粒度错误解释；
* 具备泛化能力。

### 7.2.2 可解释的文本生成评估：InstructScore

#### 7.2.2.1 InstructScore简介

**使用 GPT-4 作为检查模型 (Use GPT-4 as a checking model)**

* 人类定义可能的失败模式，并将其转化为检查项（checklist）。
* GPT-4 被用作辅助检查的工具，通过执行简单的任务（如生成答案、信息提取等），评估模型生成的输出是否符合预期。

**输出检查过程 (Detailed Checking Process)**

* 参考数据（Reference）和模型输出（Candidate）会被比对。
* 通过 GPT-4 判断候选答案的正确性，例如验证输出是否包含正确的科学术语或信息，以及是否符合上下文逻辑。
* 如果答案不正确，会标注失败模式，例如"未包含预期信息"或"语言表达不明确"。

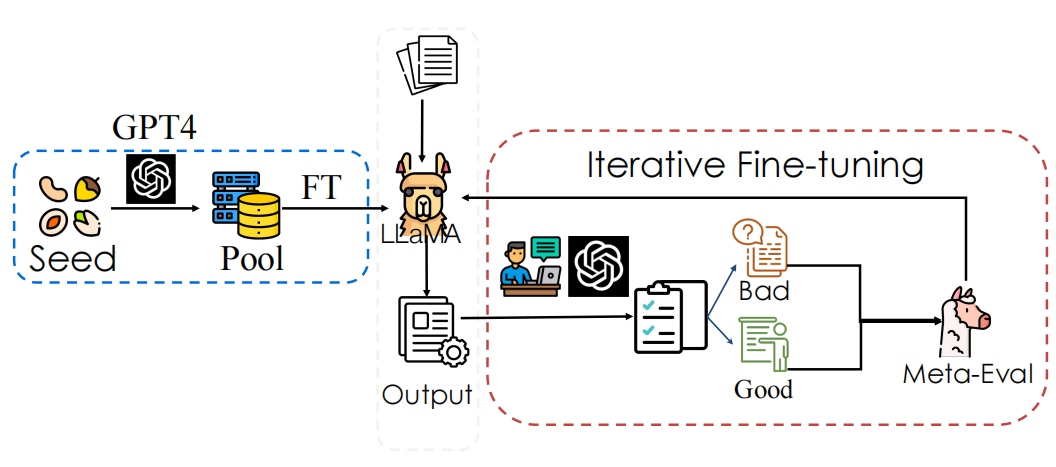
**自动化反馈 (InstructScore: Automatic Feedback)**

* 系统会根据错误和正确的标注，对每个样本生成评分表。
* 自动计算最终得分（InstructScore），用于衡量模型对指令的理解和执行质量。



**优化与迭代 (InstructScore: Refinement)**

* 基于评分结果，模型会通过迭代微调的方式不断优化其能力。
* 反馈闭环确保模型能够逐步提高，强化其在特定指令任务中的表现。



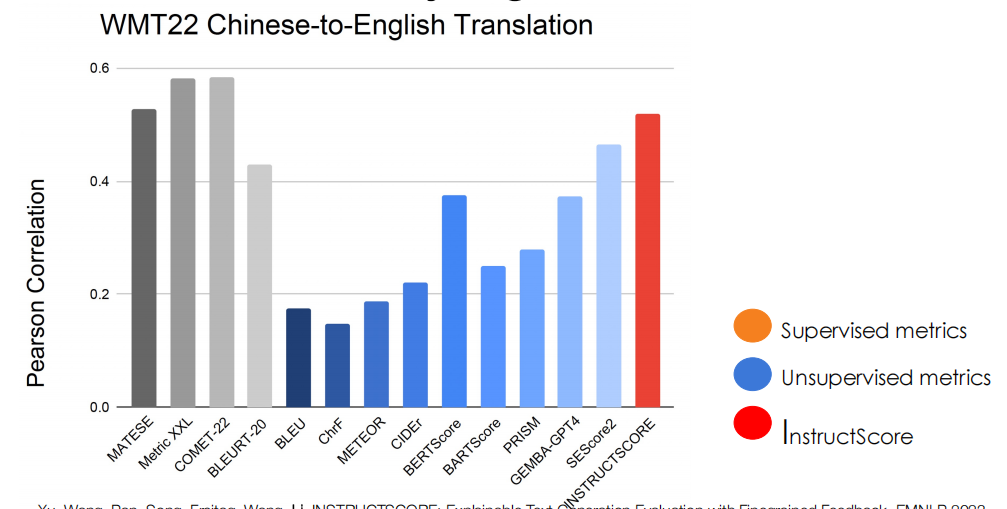
#### 7.2.2.2 InstructScore的亮点

1. **细粒度的可解释性**：提供对错误及其类型的详细洞察。
2. **高度贴合人类评估**：生成的评价与人类判断高度一致。
3. **通用性强**：无需人工评分，可适用于多种任务。

#### 7.2.2.3 InstructScore在机器翻译上的表现

下图展示了不同评估指标在WMT22中英翻译任务上的效果对比，其中横轴为不同的评估方法，纵轴为与人工评价的皮尔逊相关系数（Pearson Correlation），即评估方法与人类判断的一致性程度。图中将评估指标分为三类：监督指标（橙色）、无监督指标（蓝色）和InstructScore（红色）。

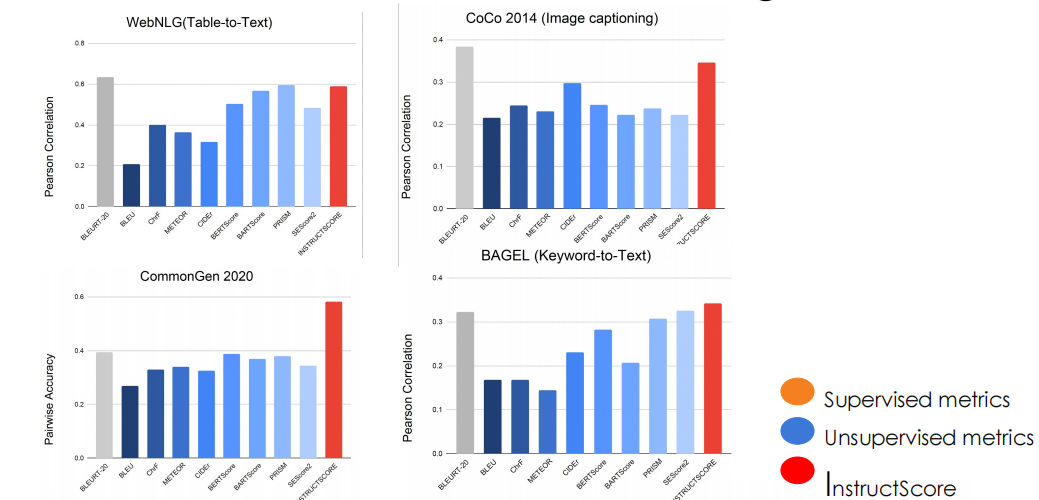
从图中可以看出，InstructScore的表现显著优于其他评估指标，说明它在评估机器翻译质量时与人类评价的相关性最高。而传统的监督指标（如COMET和BLEURT）和无监督指标（如BERTScore和PRISM）相比，虽然表现出一定水平的相关性，但整体低于InstructScore。这突出了InstructScore在机器翻译评估中的潜力与优势，表明通过指令调优的模型可以更加精准地判断翻译质量。



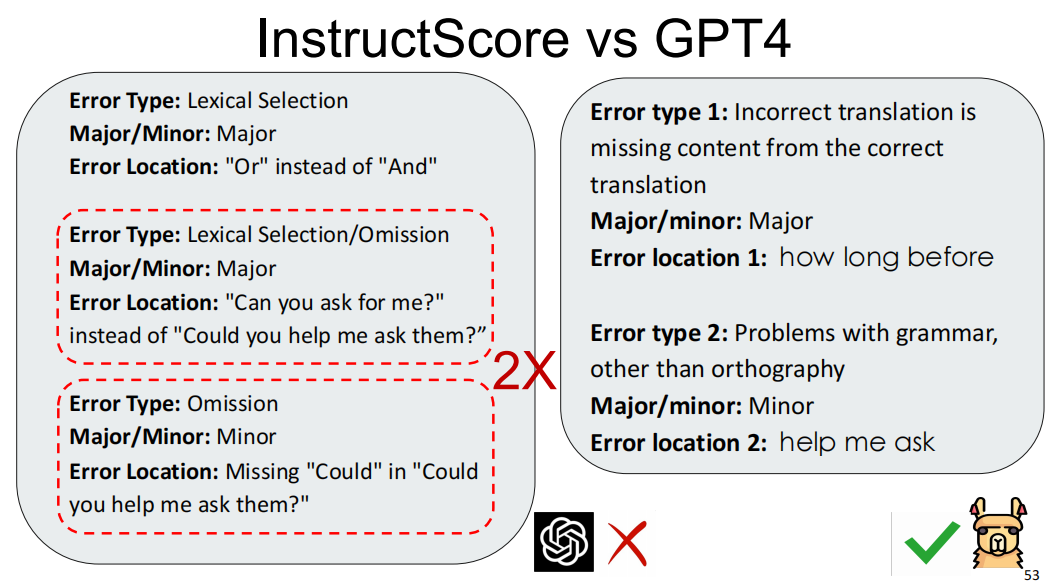
#### 7.2.2.4 InstructScore在文本生成上的表现

下图展示了InstructScore在多个文本生成任务中的评估性能表现，包括WebNLG（表格到文本生成）、CoCo 2014（图像描述生成）、BAGEL（关键词到文本生成）和CommonGen 2020（概念组合生成）。每个子图的纵轴表示与人工评价的相关性或准确性（如皮尔逊相关系数或成对准确率），横轴则列出了不同的评估指标。

从图中可以看出，无论是监督指标（橙色柱状）、无监督指标（蓝色柱状）还是InstructScore（红色柱状），InstructScore在所有任务中均表现出更高的相关性或准确性，特别是在WebNLG和BAGEL任务中优势明显。这表明InstructScore能够更好地捕捉文本生成质量，与人工判断的契合度更高。相比传统的评估方法（如BLEU、BERTScore），InstructScore通过精细化的反馈机制，显现出其在文本生成任务评估中的显著潜力和优越性。



#### 7.2.2.5 InstructScore与GPT-4作为评估指标的对比

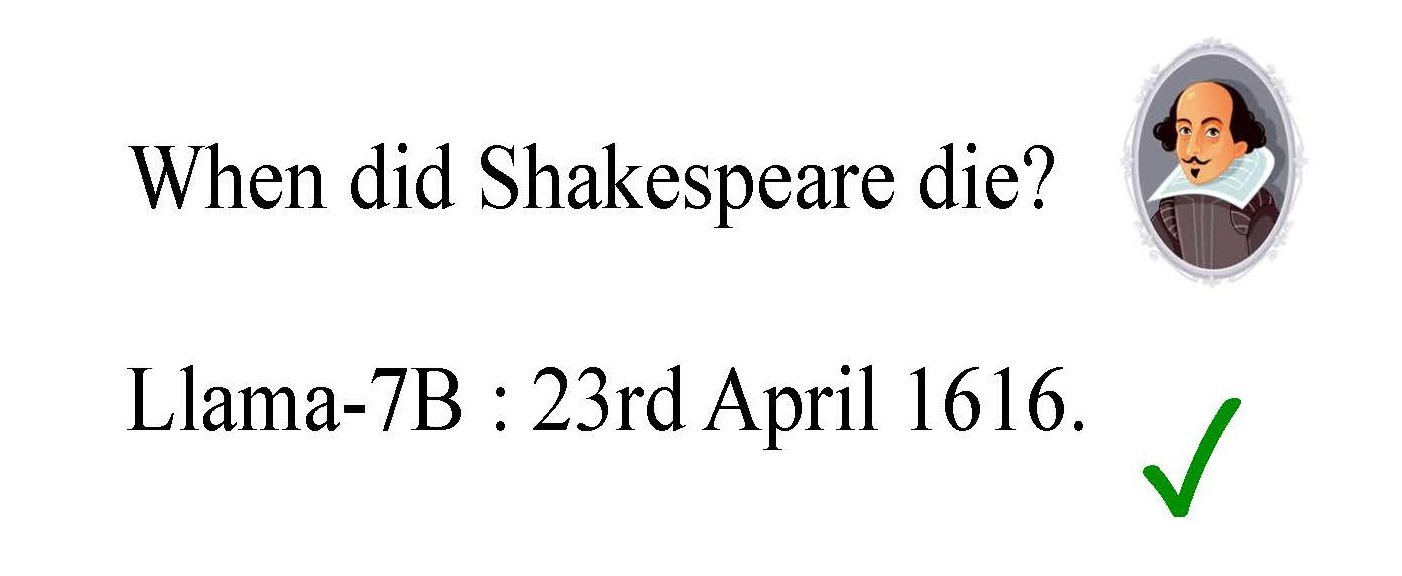


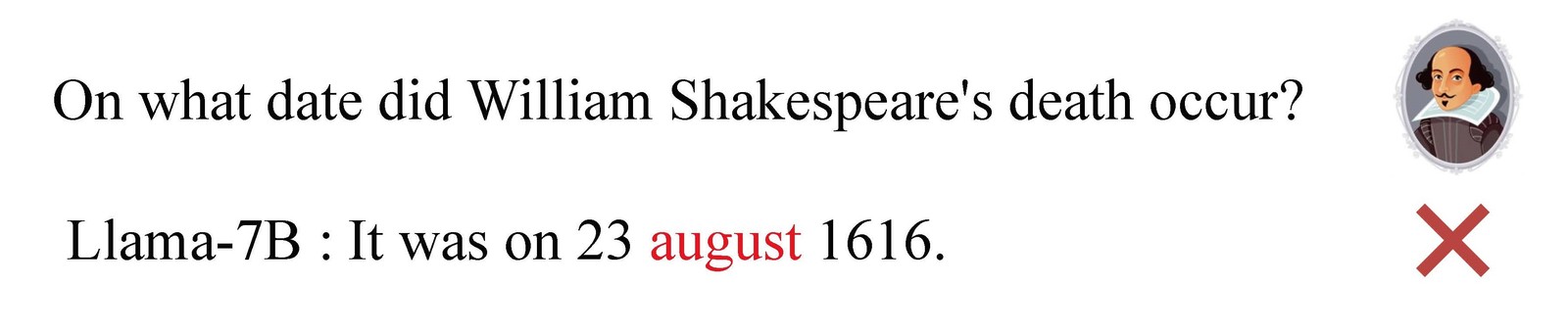
左侧是Instruct Score 错误分类，右侧是GPT-4 错误分类

InstructScore相比GPT-4在评估语言模型输出质量方面具有明显的优势。首先，它对错误类型进行了更细致的分类，例如词汇选择错误（Lexical Selection）、遗漏错误（Omission）等，并明确标注了错误的严重性（Major/Minor）和具体位置（Error Location），使分析更精准。其次，它能够更全面地捕捉错误，不仅指出词汇选择和遗漏问题，还识别语法层面的细微问题，例如“Could”缺失的情况。同时，InstructScore区分了不同错误的严重性，有助于开发者理解错误对整体输出质量的影响。此外，InstructScore在类似场景下表现出更高的一致性和全面性，便于定位系统不足之处。最后，它的分析内容更具解释性，通过提供具体的错误位置和详细描述，让使用者能够快速理解问题的来源及其解决方向。因此，InstructScore在细致性、全面性、解释性和精准性上明显优于GPT-4的评估方法。

### 7.2.3 评估LLM中知识的方法：KaRR

LLM可能生成不可靠的答案。以LLaMA-7B为例，当两次用不同的指令询问莎士比亚的死亡日期时，LLaMA-7B给出了不同的日期，这表明了LLM在生成答案时的不确定性。





#### 7.2.3.1 知识评估的必要性

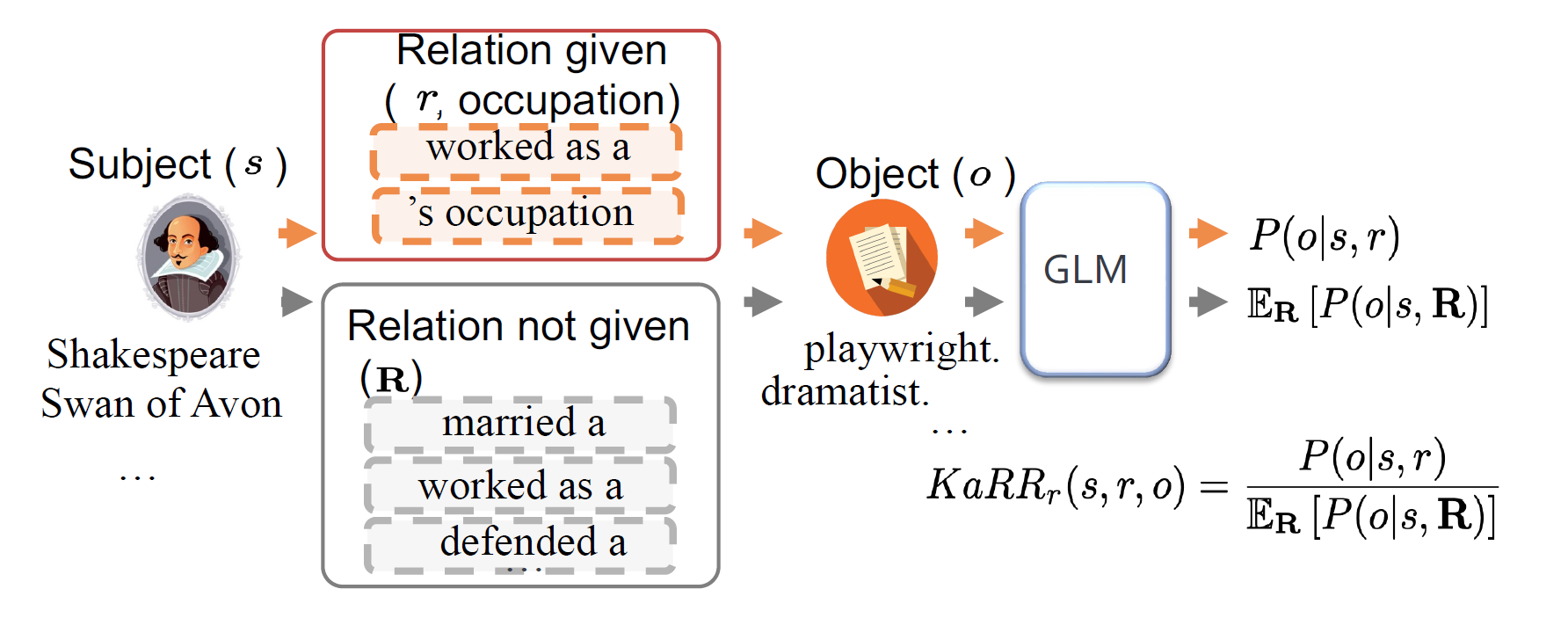
知识评估的必要性在于确保LLM的可靠性和准确性。人们需要区分文本生成是否源自真实知识还是仅与给定文本高度共现。例如，“William Shakespeare的职业是”生成的结果是作家可能是由于真实知识，而“John Smith的职业是”生成的结果是作家则更可能是因为给定的文本共现的概率较高。为了评估LLM的知识，研究人员提出了多个事实性问题，如“莎士比亚何时去世？”和“威廉·莎士比亚的死亡日期是什么？”，观察LLM对这些问题的回答情况。知识评估结果将直接影响到人们对LLM生成内容的信任度。若发现LLM生成的知识存在不一致性，则可对其进行纠正。

#### 7.2.3.2 知识评估风险比例（KaRR）

##### KaRR指标

在统计学中，人们通常使用风险比（Risk Ratio）来评估暴露（如治疗或风险因素）与结果（如疾病）之间的关联强度。

类似地，研究者们提出了知识评估风险比（KaRR），用于评估主体（subject）和关系（relation）符号对大型语言模型（LLM）生成目标符号（object）的能力的联合影响。



其结构分为以下几个部分：

1. **主体（Subject, s）**：例如图中的“莎士比亚（Shakespeare）”或“Avon之天鹅（Swan of Avon）”。
2. **关系（Relation, r 或 R）**：
   1. **给定具体关系（Relation given, r）**：如“worked as a ...'s occupation”（从事某职业）。
   2. **未给定具体关系（Relation not given, R）**：表示更广义的、未明确指定的关系，如“married a”、“worked as a”、“defended a”等。
3. **目标（Object, o）**：LLM需要生成的目标符号，例如“剧作家（playwright）”或“剧本作家（dramatist）”。
4. **GLM（生成式语言模型）**：通过输入主体和关系（或缺失关系），模型生成目标符号，并输出概率 P(o∣s,r) 和 ER[P(o∣s,R)]。
5. **KaRR公式**：

KaRR 的定义为

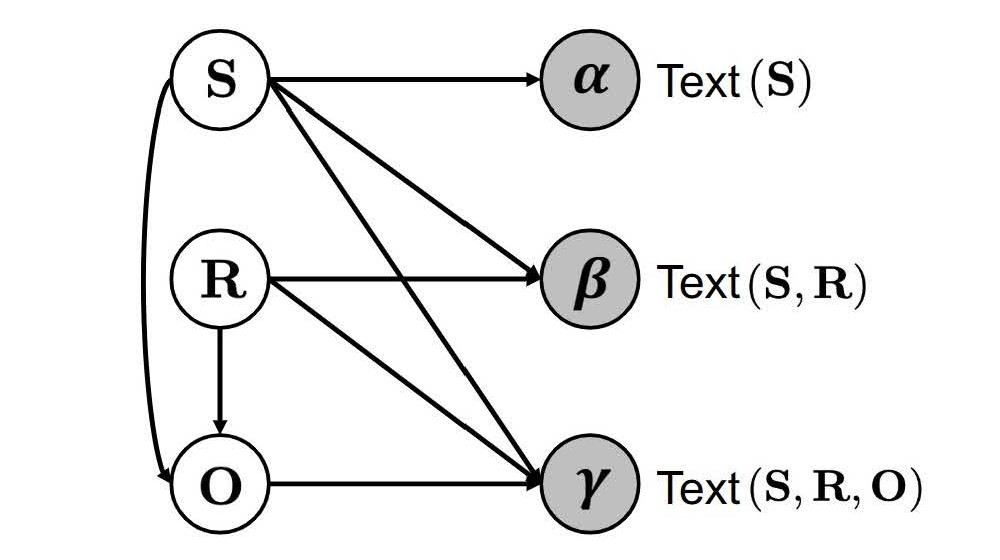
descript

公式的含义是，将在给定具体关系 r 时生成目标 o 的概率，与在未给定具体关系时的期望概率进行对比，以衡量关系 r 的作用。

总结来说，KaRR 量化了主体和具体关系符号的组合对 LLM 生成目标的影响，帮助评估模型在知识生成中的表现。

##### KaRR计算

通过图模型分解潜在变量与可观测变量之间的关系，从而计算 KaRR 。KaRR 衡量了在特定关系 R 下，LLM 生成目标符号 O 的能力，同时对比了未特定关系下的生成期望。这种分解方法使得模型的知识评估更加系统和精确。



**具体概率计算：**

1. **计算 P(o∣s,r)**： 通过分解关系生成过程，公式为：

descript

即为所有文本形式（β）中生成目标 O 的联合概率的总和。

1. **进一步分解**： 联合概率可写为：

descript

其中：

* 1. descript：从主体 S 和关系 R 生成具体文本 β 的概率。
  2. descript：在具体文本 β 下生成目标 O 的概率。

##### KaRR评估表现

**KaRR 数据集**

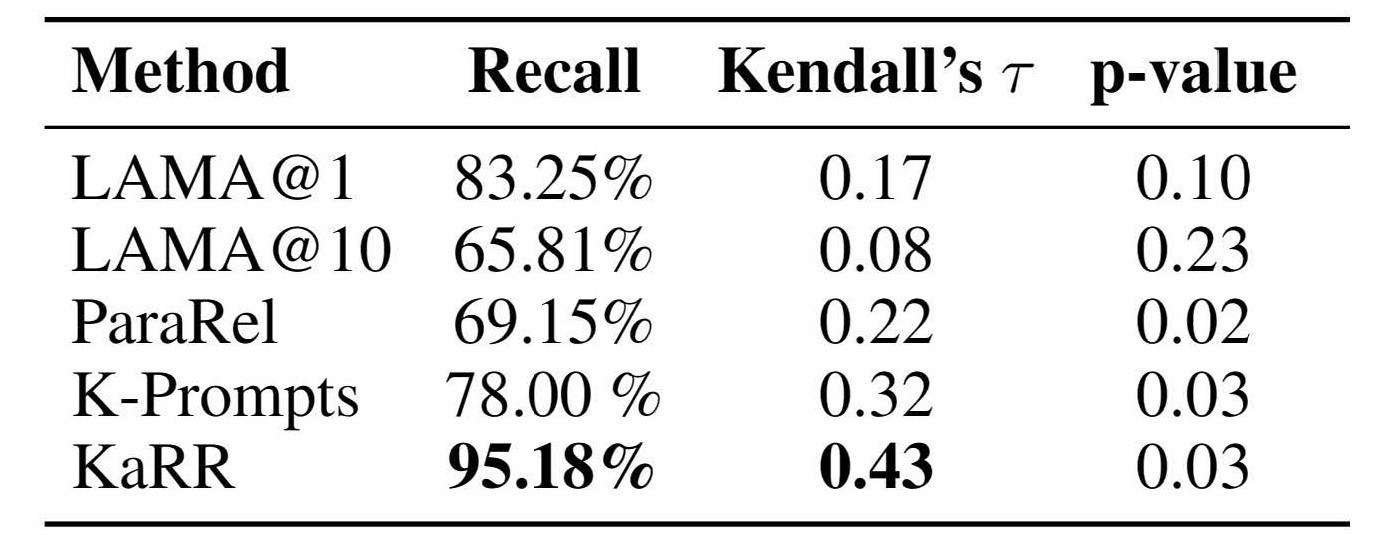
* **覆盖范围广泛**：包含 100 万个实体，涉及 600 种关系。
* **知识模板示例**：使用模板化的方式表示知识。例如：“P36”（首都）："[X] 是 [Y] 的首都。"
  + - “P19”（出生地）："[X] 的出生地是 [Y]。"
  + 每种关系都设计了多种表达形式以覆盖不同的语言风格。

**人工注解过程**

* **注解阶段**：每位注解者编写 3 个提示（prompt）来探测模型的知识。
  + 根据模型生成的回答对提示进行优化，直到生成内容与目标答案匹配。
* **评分阶段**：另一组注解者对模型生成的答案进行评分（0 或 1），判断其是否准确。

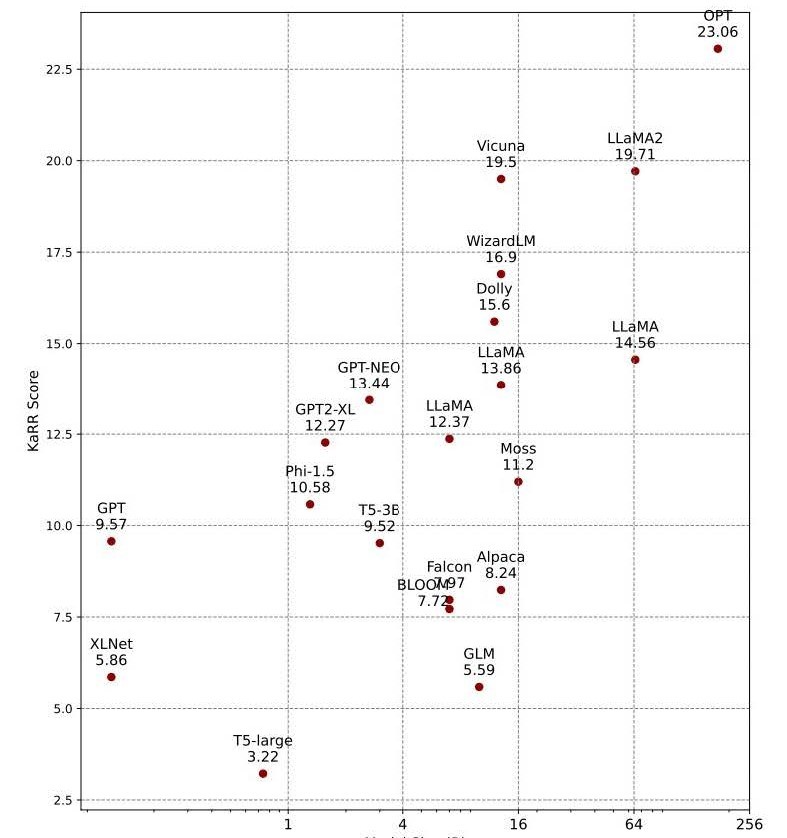
**评估结果：人类评估与模型评分的关联性**

* **关联性分析**：使用 Kendall tau 相关性评估各种方法的得分与人类评估排名之间的关系。
* **发现**：
  + 与人工评估高度相关，评估偏差较小。

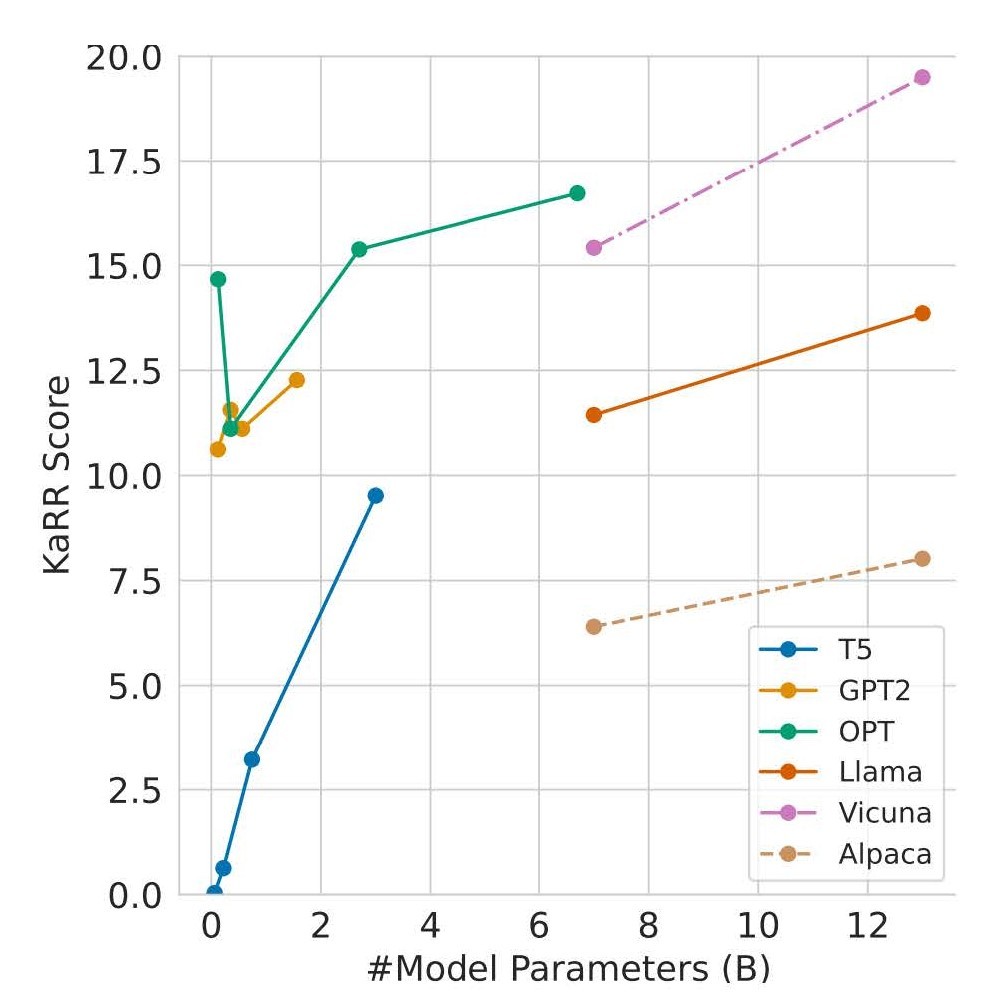


**模型表现：KaRR 分数**

* 小型和中型语言模型在生成正确事实方面表现不稳定。
* 通过使用来自更知识丰富模型的数据进行微调，可提升模型的知识表现。



* 大型模型通常具备更多事实性知识。
  + 不同模型的扩展收益存在差异，例如从 T5-small 到 T5-3B 的变化。

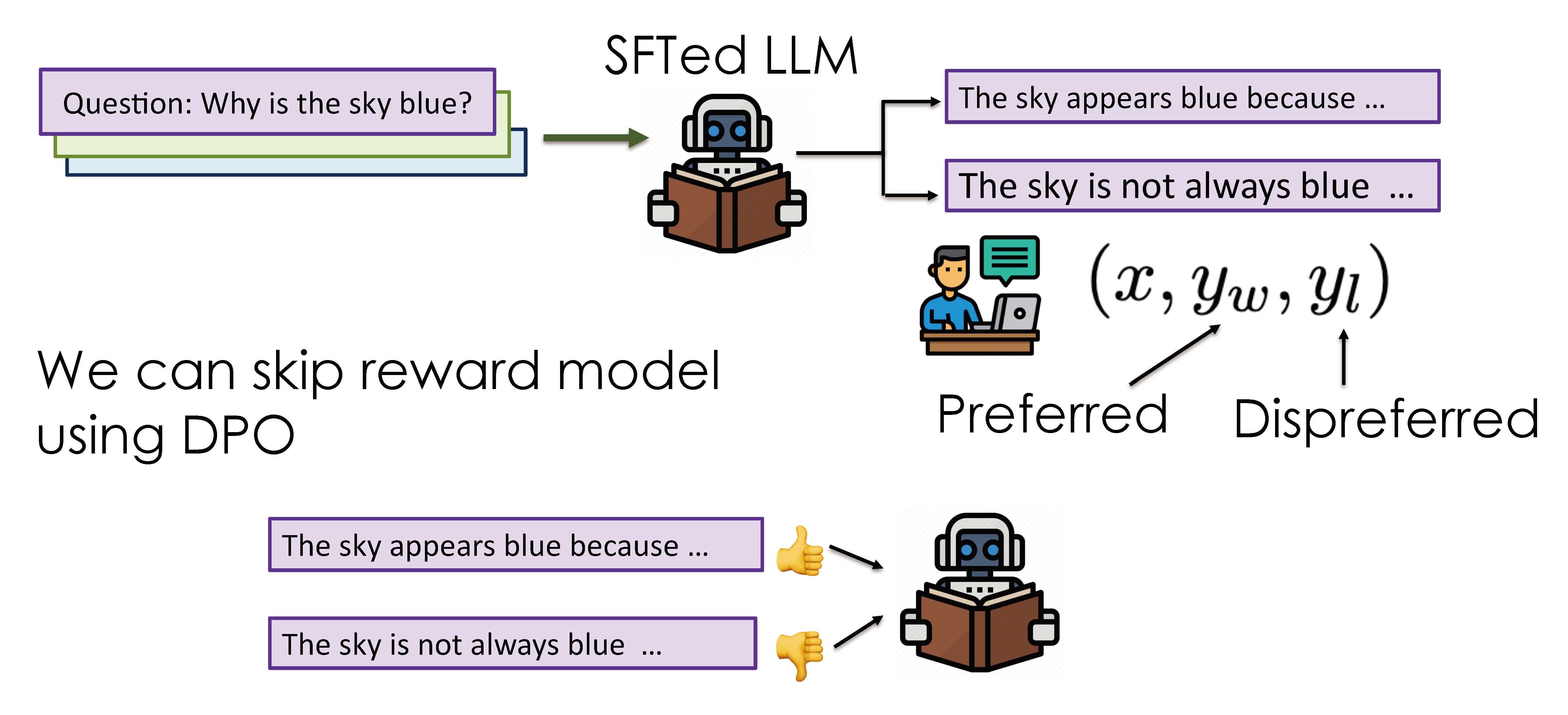


## 7.3 后训练对齐

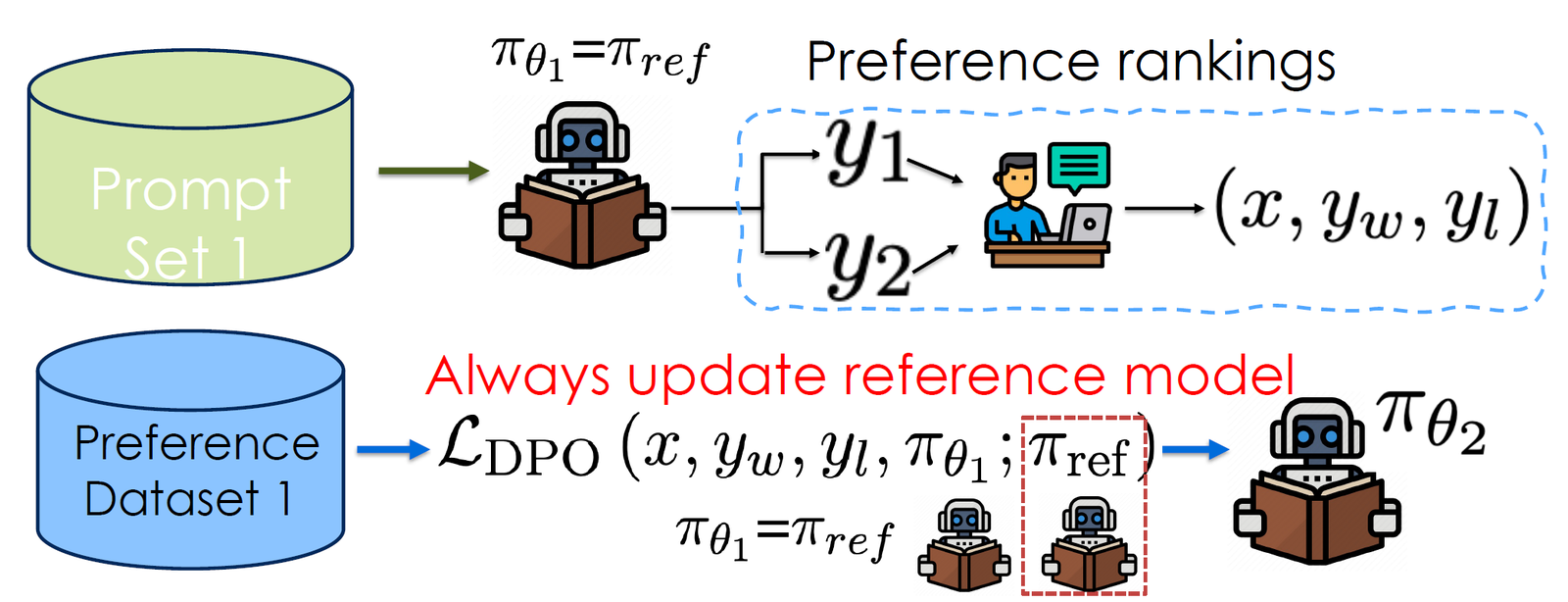
### 7.3.1 行为偏好优化（BPO）

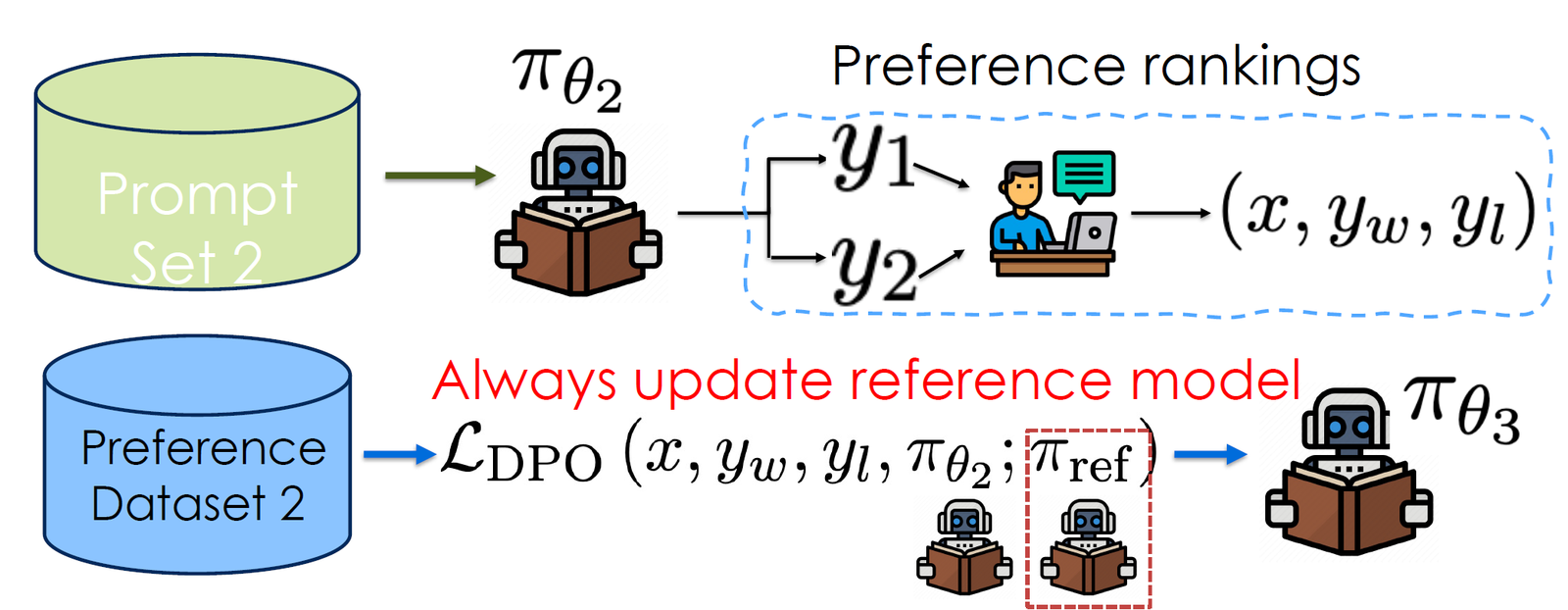
现有方法在模型对齐上仍然存在诸多挑战，例如模型评估的自我偏倚、离线优化数据不足，以及在线优化过程中参考模型更新的滞后性等。为了克服这些问题，研究者提出了一种全新的行为偏好优化（Behavior Preference Optimization, BPO）方法，通过动态调整参考模型和行为模型之间的接近性，显著提升了在线对齐的效果。

传统上，大语言模型的优化主要依赖于人类反馈进行偏好标注。具体来说，人工标注优选和非优选回答，构建偏好数据集，并利用这些数据训练奖励模型，从而提升模型的指令执行能力。这种基于人类反馈的奖励建模（RLHF）通常采用Bradley-Terry模型，通过强化学习优化语言模型。然而，奖励建模面临过拟合的风险，尤其是在训练数据规模有限的情况下。直接偏好优化（DPO）应运而生，通过固定的偏好数据集和参考模型，减少了奖励建模中的过拟合问题。DPO利用方差损失或合页损失等机制优化偏好对齐，提供了一种高效的离线对齐方法。然而，DPO的局限性也十分明显：它依赖于离线收集的数据集，而无法适应动态变化的偏好需求，同时固定的参考模型往往难以跟随行为模型的变化，从而导致对齐效果的下降。

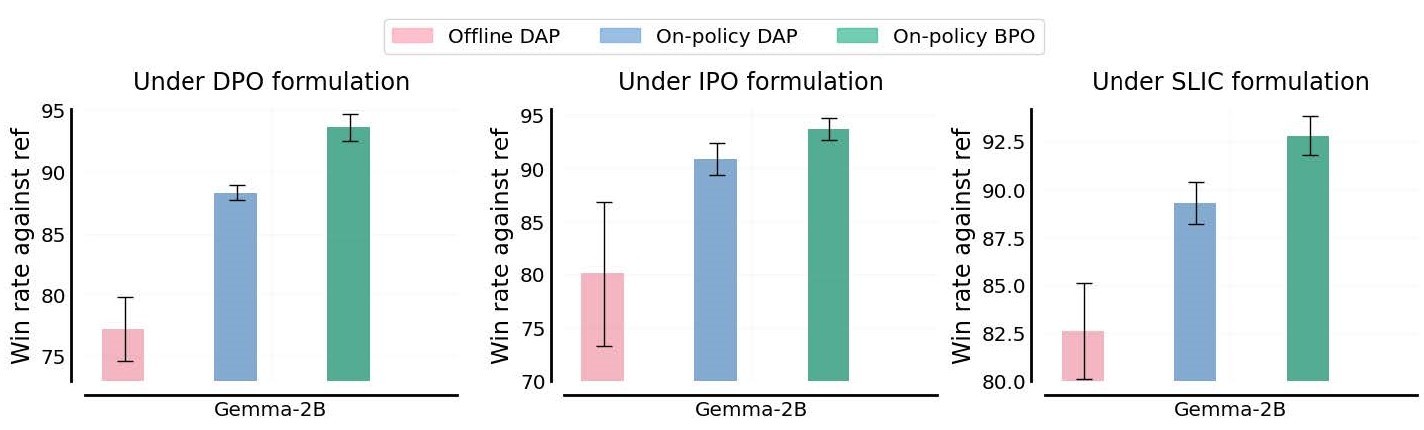


为了克服这些不足，研究者提出了行为偏好优化（BPO）这一新算法。BPO的核心思想在于，始终保持参考模型与行为模型的接近性，以提升在线对齐的效果。在具体实现上，BPO通过逐步迭代数据收集与偏好排序（即Prompt Set和Preference Dataset），在每个阶段都更新参考模型，使其始终贴近行为模型。此外，BPO引入了一种基于模型权重平均更新的机制，即通过LoRA（Low-Rank Adaptation）对行为模型的权重进行独立更新并融合，从而提高优化的效率和稳定性。这一动态调整机制有效克服了传统方法中由于参考模型固定导致的局限性。





在实验部分，研究者设计了多项任务，将BPO与现有的在线和离线对齐方法进行对比。结果表明，BPO在生成质量、偏好对齐和用户满意度等指标上均显著优于基线模型，特别是在经过多轮数据收集后，其对齐效果提升尤为显著。实验进一步揭示了参考模型与行为模型的接近性是实现优异对齐效果的关键因素，而非单纯依赖参考模型本身的质量提升。这一发现凸显了BPO的独特优势，即通过动态调整参考模型，使得对齐过程更加贴合行为模型的实际表现。



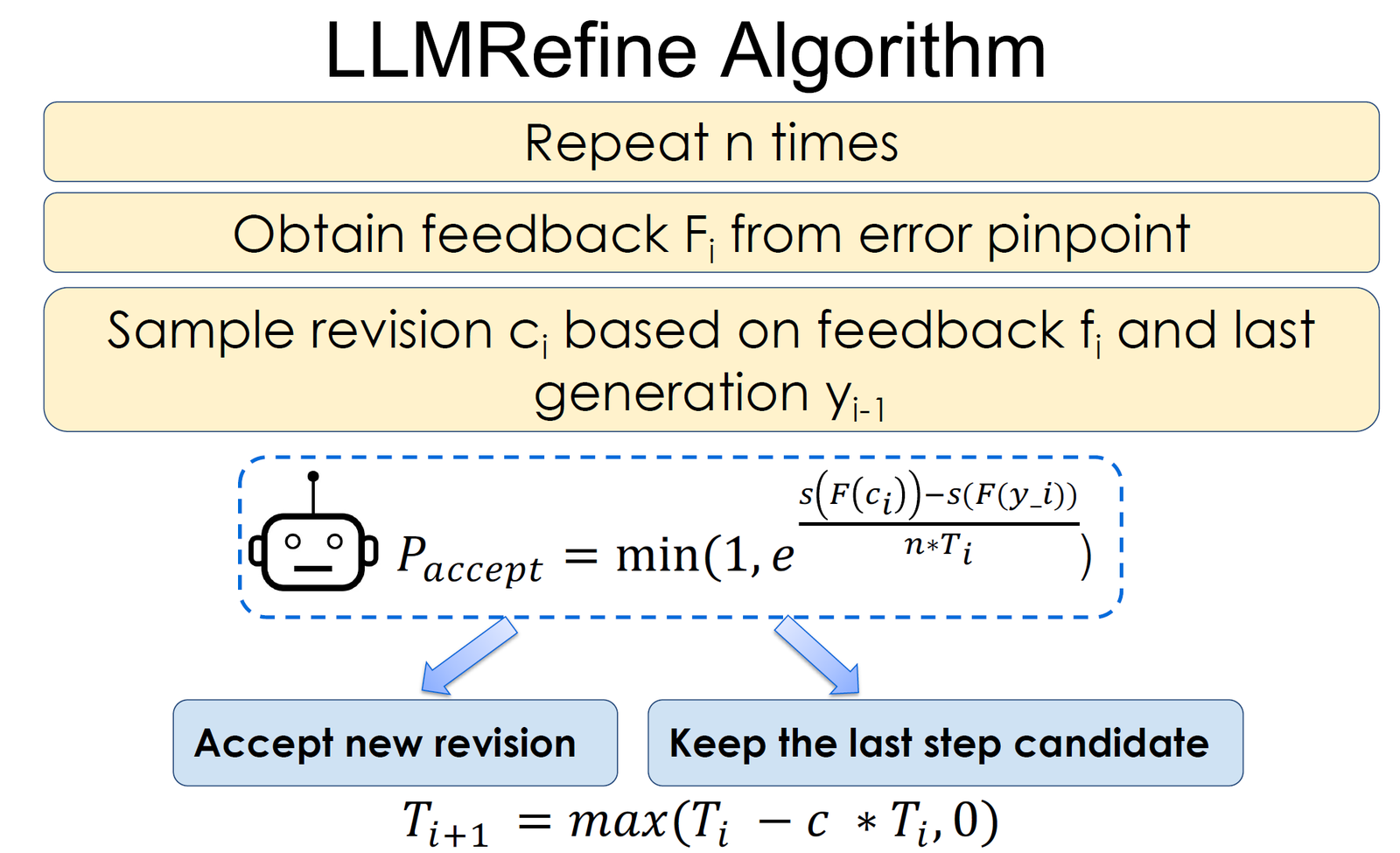
综上所述，BPO为大语言模型的优化提供了一种全新的范式，通过动态调整和权重平均更新，克服了传统方法在离线数据和固定参考模型上的局限性。其在多个任务中的卓越表现表明，这一方法不仅具有重要的理论价值，也具备广泛的实际应用潜力，为未来LLM优化和用户偏好对齐指明了新的方向。

### 7.3.2 精细化反馈（LLMRefine）

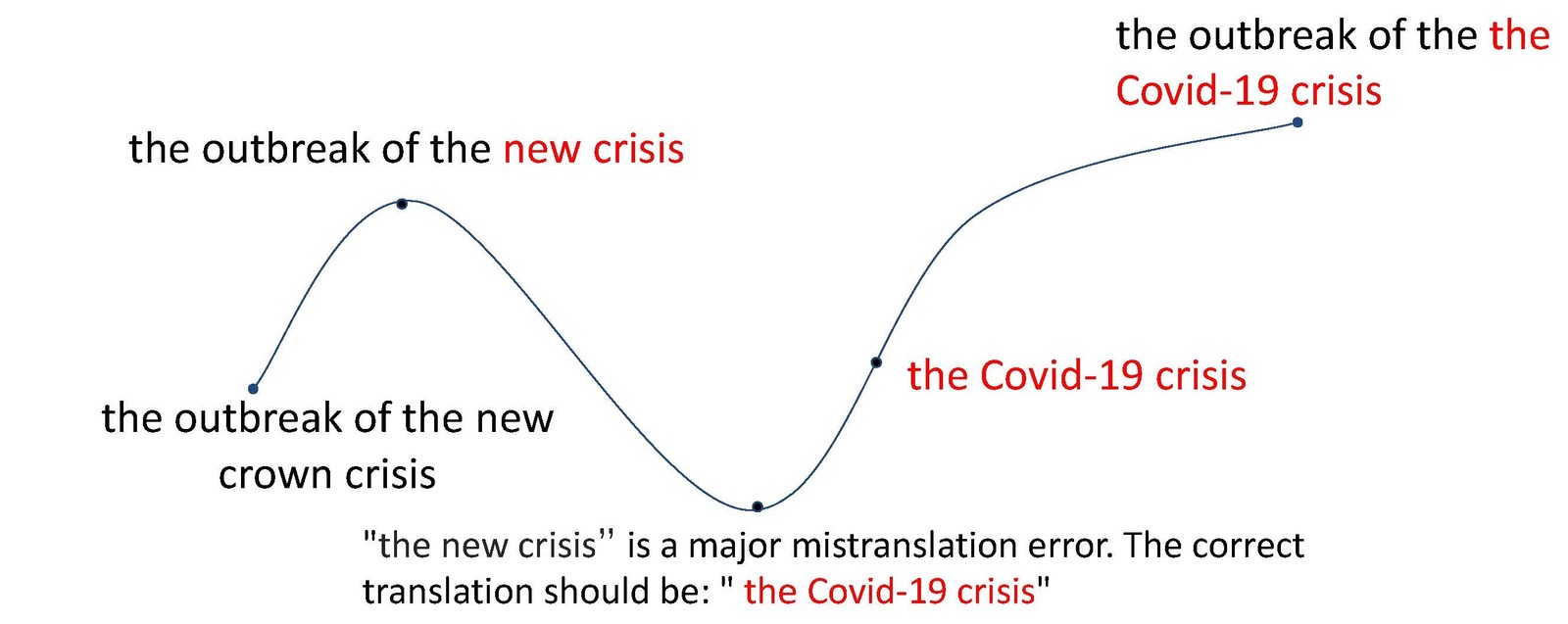
本部分探讨了在LLM中利用细致反馈来引导其改进输出质量的方法。具体来说，本部分以翻译任务为核心展开，展示了如何通过不同类型的反馈来指导LLM进行翻译修正。翻译质量的评估往往涉及多层次的衡量标准，包括语义的准确性和用词的恰当性。然而，简单的改进请求或二元反馈（如仅指出正确或错误）通常缺乏足够的信息量，使得模型难以理解需要改进的具体细节。例如，当输入是“新冠疫情危机爆发”时，模型可能生成“the outbreak of the new crown crisis”。虽然语法正确，但术语“new crown”显然不符合规范。在此情况下，仅仅告诉模型“翻译有误”并不能有效地指导其生成更优解。

为了克服上述问题，研究提出了包括四种类型反馈在内的分析框架。第一种是简单的改进请求，例如直接要求模型“改进翻译”，但未提供具体建议；第二种是二元反馈，指出翻译错误，但未指明问题所在；第三种是标量反馈，通过给定评分（如70/100）为模型提供粗略的评估；第四种则是精细化反馈，通过明确指出错误的位置和类型（如术语使用不当），为模型优化提供更直接的指导。精细化反馈在这一框架中表现出显著优势，因为它能够帮助模型理解问题的具体来源，从而更精准地调整输出。

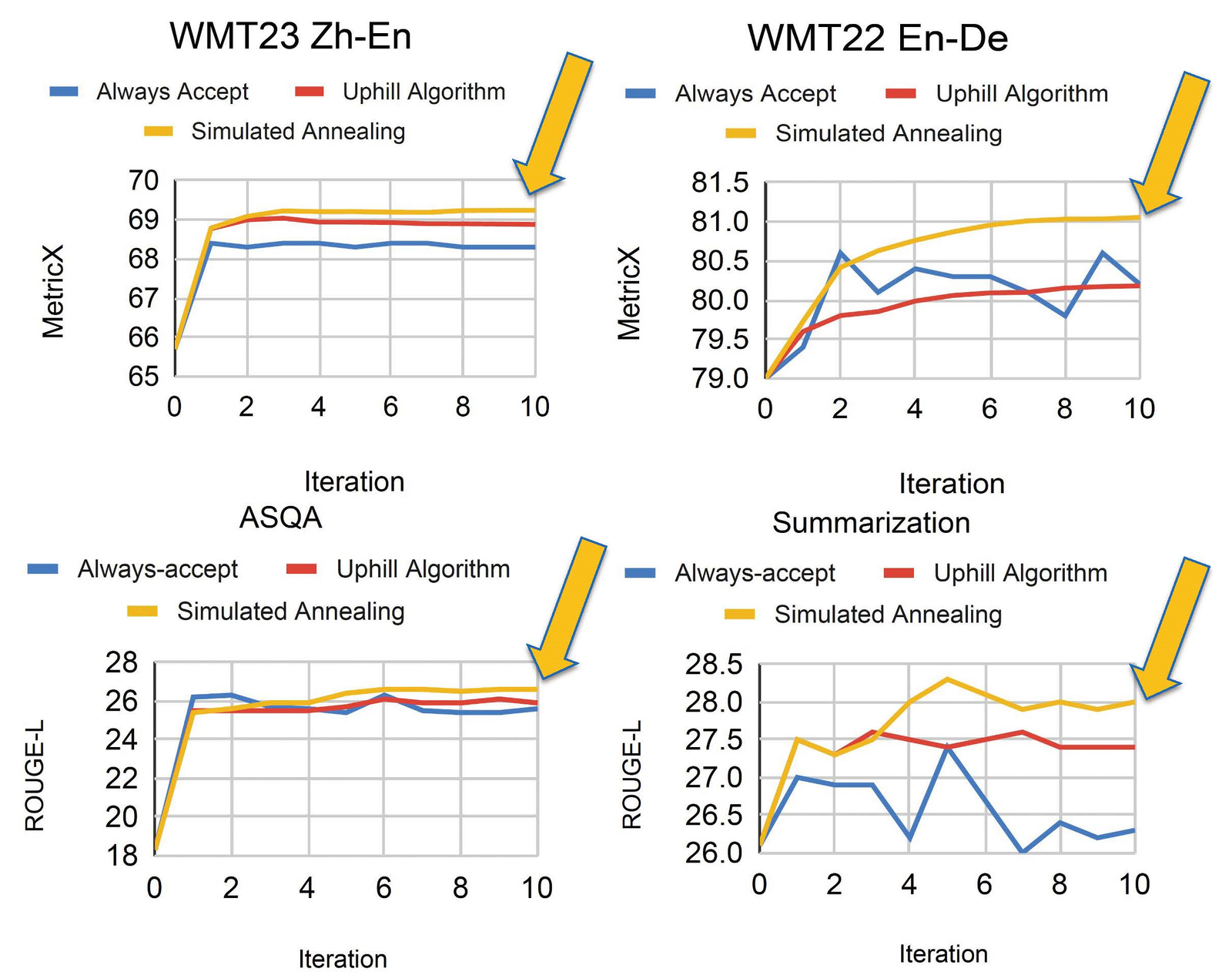
基于上述反馈类型的讨论，研究进一步提出了一种算法迭代优化方法，称为LLMRefine。该方法的核心思想是利用精细化反馈在多轮迭代中逐步优化模型的生成结果。具体而言，首先，通过识别翻译输出中的具体错误点生成精细化反馈，例如指出“new crown”为术语错误；其次，根据反馈采样新的翻译候选版本，并评估修订是否优于上一版本；最后，结合模拟退火（Simulated Annealing）等优化算法动态调整接受新修订的概率，以探索可能的最优解。



例如，经过多轮优化后，原始翻译“the outbreak of the new crown crisis”最终被改进为“the outbreak of the Covid-19 crisis”，这一过程展现了精细化反馈在模型优化中的强大潜力。



这一方法不仅在翻译任务中表现突出，还显示出其在摘要生成和问答等场景中的广泛适用性。通过动态调整优化策略，模型能够适应多样化任务需求，同时提升生成结果的整体质量。这一过程还表明，简单的二元反馈往往不足以支持复杂任务的优化，而精细化反馈由于其明确的操作性，为解决复杂问题提供了新的解决方案。



研究表明，通过精细化反馈结合迭代优化算法，大型语言模型能够实现显著改进。这一发现对未来LLM在人机交互中的应用具有重要意义。无论是翻译、问答，还是更复杂的语言生成任务，精细化反馈的引入都为优化过程提供了清晰的方向。随着相关算法的进一步发展，LLM有望在生成任务中实现更高水平的精确性和语义一致性。

综上所述，LLMRefine的研究不仅深化了对LLM优化机制的理解，还为实践中如何更有效地改进模型输出质量提供了重要指导。这一方法为提升LLM性能奠定了坚实基础，并为未来探索更加智能的人机交互奠定了理论框架和实践路径。

## 结语

本章节探讨了大语言模型（LLM）评估及对齐的相关问题和未来研究方向。通过对自偏差、可解释评估方法（InstructScore）和知识评估方法（KaRR）的讨论，本文揭示了当前LLM评估在复杂知识、RAG任务及智能体性能测评方面的挑战与进展。同时，后训练对齐方法如在线偏好优化（BPO）和细粒度反馈的迭代改进（LLMRefine）为提升模型生成质量和对齐学习提供了新的视角。未来研究应聚焦于更加稳健的对齐学习方法，并继续探索对复杂任务和开放式生成的高效评估，为大语言模型的进一步优化奠定坚实基础。

## 参考资料

李磊老师主页：<https://lileicc.github.io/index.html>

BPO: Supercharging Online Preference Learning by Adhering to the Proximity of Behavior LLM：<https://arxiv.org/abs/2406.12168>

Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement：<https://arxiv.org/abs/2402.11436>

Statistical Knowledge Assessment for Large Language Models：<https://arxiv.org/abs/2305.10519>

Learning Personalized Alignment for Evaluating Open-ended Text Generation：<https://arxiv.org/abs/2310.03304>