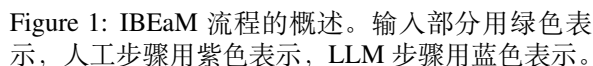


Christian M. Angel, Francis Ferraro
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD 21250 USA
{ change1, ferraro } @umbc.edu

对提示工程的积极研究主题使得大型语言模型 (LLM) 对提示措辞的微小变化敏感这一点显而易见。部分原因可以归结为存在于 LLM 中的归纳偏差。通过将 LLM 的输出作为其提示的一部分，我们可以更容易地创建令人满意的提示措辞。这产生的效果是创建一个与模型中的归纳偏差相匹配的提示。实证上，我们表现出使用这种归纳偏差提取和匹配策略可以将用于分类的 LLM Likert 评分提高多达 19%，并将用于排序的 LLM Likert 评分提高多达 27%。

如果你想建一个露台，你可能会向 LLM 询问如何开始。它可能建议的两个步骤是“购买露台砖块”和“将露台砖块放置在所需区域”。虽然购买露台砖块后放置它们是一个合乎逻辑的行动步骤，但这并不一定意味着你必须立即放置它们，或者你可能想要建造其他同样需要露台砖块的东西。这种不寻常的应用可能会让只见过这两个步骤按这个顺序安排的 LLM 感到困惑，并可能倾向于将砖块的放置标记为购买的结果。

我们用来提高大型语言模型（LLM）在非比较分类任务中表现的方法，是一种我们称之为归纳偏置提取和匹配（IBEaM）的过程，如图 1 所示。通过结合我们对任务的先验知识，即任务可以分解为哪些子任务以及一旦解决后如何组合它们，我们可以更充分地利用 LLM 的不



对于本文中我们研究的每一个任务，我们提示一个大型语言模型 (LLM) 使用一个或多个 10 点 Likert 量表，这些量表可用于生成该任务每个实例的一个或多个得分。这些得分被组合来生成一个总体得分（如果任务是多选，则为多个总体得分），用于进行分类。对于每个任务，我们还评估了一个简单的基线，在该基线中，LLM 会被提示为每个实例提供从 1 到

10 的评分（如果任务是多选，再次则为多个评分）。与使用我们的 IBEaM 方法生成的得分一样，这些评分用于执行下游分类任务。

我们的贡献是：

- 我们定义了归纳偏置提取的方法，并在提示中包括提取的归纳偏置以提高 LLM 的性能。
- 我们证明了你可以描述评估标准，并让 LLM 识别其偏好的措辞。
- 我们表明，一个大型语言模型在生成数字评分方面的能力是有限的，并展示了可以使用 IBEaM 来从大型语言模型生成更好的数字评分。
- 我们进行消融研究，显示 IBEaM 的所有组成部分都是有用的。

2 相关工作

随着生成大语言模型（LLMs）的出现，已经进行了大量研究，以确定它们可以应用于哪些任务以及它们在这些任务中的有效性。虽然 LLMs 展示了使其能够执行以前机器学习方法无效的任务的突出表现，它们也可以应用于在 LLMs 引入之前已经存在的任务。将 LLMs 应用于这些较老任务的动机可能是希望提高执行速度的效率，提高非技术用户的可访问性，或改善精确度方面的最新技术。这种 LLM 应用的广泛范围促使我们评估 IBEaM，通过生成数值得分并使用它们作为执行分类和排序任务的代理。

提高 LLM 在任务上的表现的一种方法是进行思维链推理，即在要求 LLM 评估实例 (Wei et al., 2022) 之前，通过在提示中解决一个或多个例子的过程来引导 LLM。提高 LLM 表现的另一种方法是进行提示工程 (Chen et al., 2025)，这通常是一个通过对提示进行小调整来优化 LLM 下游表现的重复过程。所有提示工程技术从根本上都是试图使提示更好地与 LLM 学习的参数协同工作，换句话说，都是试图找到与 LLM 的归纳偏置最兼容的提示。

本文中研究的所有数据集都包含文本中暗示的某种世界状态的概念。LLMs 能够推理由提示暗示的假设世界状态，以及由其训练数据 (Zhu et al., 2023) 暗示的真实世界状态。这种真实世界状态的编码是 LLMs 的一种归纳偏差形式。

我们在本文中研究的任务类别是分类。Gretz et al. (2023) 选择了几个分类任务和大型语言模型，并比较了这些模型在零样本时的性能与经过微调后的性能。Sun et al. (2023) 还创造了

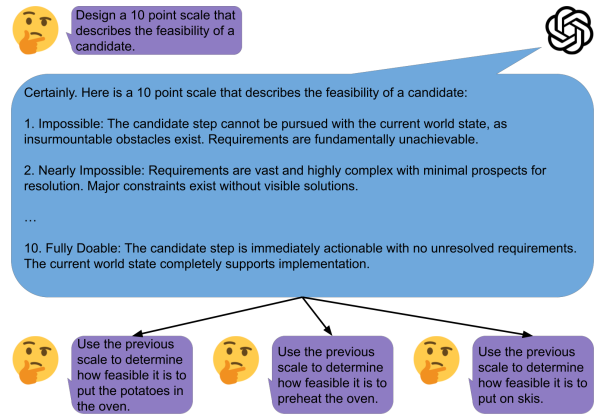


Figure 2: 对于我们的每次实验，我们为每个指标生成一个 Likert 量表，并将其应用于该实验中的所有实例。我们在未来调用 LLM 时将量表作为对话历史的一部分。对所有实例保持量表一致性提高了评级的一致性，并减少了与为每个实例重新生成量表相比的成本和计算时间。

一种新颖的提示技术，以提高大型语言模型执行文本分类的能力。

使用 LLM 生成 Likert 评分已被 Chiang and Lee (2023) 评估。在他们的工作中，他们发现要求 LLM 解释其分配的评分比将 LLM 输出限制为单一评分整数的方法提高了性能。我们将这种解释生成技术应用于 IBEaM，以优化我们的性能。

将问题分解成更易管理的子问题的概念以前也被研究过。LLM-Eval 是一个通过将问题分解成子任务并提示 LLM 完成响应 (Lin and Chen, 2023) 的模式来改进 LLM 输出的系统。LLM-Eval 和我们的工作之间存在一些有趣的技术差异。首先，LLM-Eval 使用模式填充，但我们发现，在 IBEaM 的特定提示设计中，模式填充没有自由形式的文本生成有效。其次，LLM-Eval 使用单一提示来评估多个指标，但对于 IBEaM 的特定提示设计，我们发现对不同的指标使用多个提示最为有效。

还有之前的工作使用了与 LLMs 进行多步骤对话的方法。Trivedi et al. (2023) 提出了一种称为“思维链交替检索” (IRCoT) 的技术。通过这种方法，他们使用一系列提示来执行思维链推理，但在每个提示之间，他们会查询一个知识数据库。从数据库中检索到的信息用于影响序列中下一个提示。虽然我们在 IBEaM 中与我们的提示序列使用了类似的方法，但我们并不查询外部的知识源。相反，我们使用的中间信息是从 LLM 中获得的。

3 方法

将 IBEaM 应用于新任务是一个四步过程。首先，用户必须创建一些组件指标以便能够提取

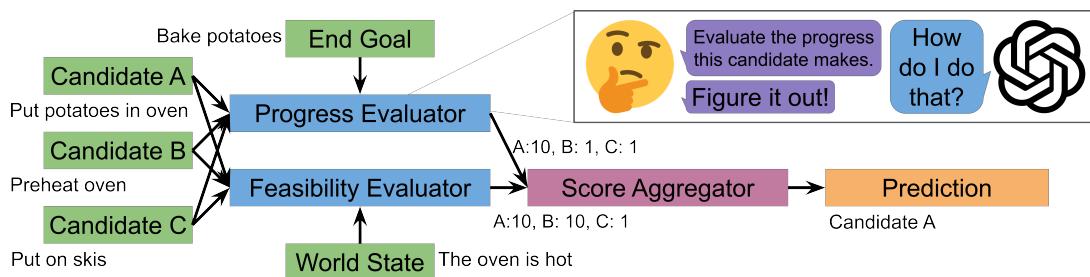


Figure 3: 对我们 WikiHow 任务中使用的 IBEaM 的概述。如图 1 所示，绿色方块表示数据集中的输入。管道的核心是为给定任务设计的一组评估器。对于 WikiHow，我们有一个进度评估器和一个可行性评估器。这些评估器的输出交给适当的聚合器，该聚合器然后为实例做出最终预测。为 SAGA 任务 2 和动作条件任务设计的类似管道分别显示在附录 B 中的图 5 和 6 中。

额外的归纳偏差。其次，用户必须创建一个提示模板，以便将提取的归纳偏差整合到一个评估提示中。第三，用户必须提示 LLM 提取每个组件指标的归纳偏差，最后，用户必须设计一些方法将组件评分组合成最终评估。一旦完成了所有这些步骤，结果就是一个可以用于提高 LLM 在目标任务上性能的管道。

3.1 组件指标的创建

在将 IBEaM 应用于某个问题的过程中，第一步是将问题分解为合理数量的细粒度子问题。这里的目标不是将问题分解得过于细小，而是确定定义任务的高层次维度。例如，如果我们的任务是确定“将土豆放入烤箱”在已经完成“预热烤箱”这一步骤的前提下，是否继续“烤土豆”的过程，那么该任务的两个逻辑维度是进展和可行性。理想情况下，当我们要判断一个步骤是否是朝着目标推进过程的逻辑延续时，我们希望候选步骤既能取得实质性进展，也能具备实质性可行性。其他任务会分解为不同的组成指标，这需要由 IBEaM 用户进行判断。

将任务分解为组件子任务并不是 IBEaM 独有的特性，但 IBEaM 从这一过程中获得了巨大的益处：从 LLM 中跨独特维度获取的 Likert 量表越多，就可以从 LLM 中提取出更多的归纳偏差。IBEaM 从这种额外的归纳偏差提取中得到的好处是对任务进行逻辑子问题分解所产生的直观好处的补充。这些直观好处包括获得来自 LLM 的更详细的输出、获得跨多个维度的评估，以及如果适用，获得对不同于问题答案进行不均匀加权的能力。

3.2 归纳偏差提取

在将 IBEaM 应用于问题的过程中，下一步就是提取 LLM 对每个组件度量的归纳偏差。对于我们在本文中研究的所有任务，这涉及到为每个组件度量提示 LLM 生成一个 Likert 量表。例如，对于之前提到的可行性度量，LLM 可以

生成如图 2 所示的 10 点量表。正如图中所示，我们可以提取模型的归纳偏差，具体表现为其在 Likert 量表中对每个评分的偏好措辞。如果我们尝试手动创建这个量表而不是提示 LLM 生成它，我们将不得不进行一个劳动密集的提示工程过程。

3.3 归纳偏置匹配

在将 IBEaM 应用于问题的过程中，第三步是在提示时匹配 LLM 的归纳偏差。这涉及创建一个提示来评估每个指标，其中包括生成的李克特量表。正如我们在实验中所展示的，用于评估每个指标的提示不需要复杂，并且可以将生成的李克特量表作为 LLM 对话历史的一部分，这使我们能够在实验中跳过任何李克特量表的后处理。在提示中通过使用如下措辞利用该对话历史：“在之前的可行性量表上对以下候选进行 1 到 10 的评分。”这个过程的简化版本如图 2 所示。

这种归纳偏好匹配过程的另一个好处是，我们获得了一个一致的评分尺度，可以应用于任务中的所有实例，这使得与直观的基准方法相比，大型语言模型能够应用一致且更具体的标准。

3.4 组合组件度量

由于将任务分解成多个子问题，我们面临着多个组件评分需要以某种方式组合以形成总体评分的问题。我们组合评分的具体方法会根据任务类型的不同而有所变化。对于像 WikiHow 这样的排序任务，简单地对分数进行求和通常是一个有效且直观的起点。我们测试了求和、乘积、分数排名求和，以及分数排名的乘积。在这些方法中，我们发现分数排名的乘积是最有效的组合方法，而且不需要任何学习的参数。假设 n 是多项选择题的选项数量，这种方法通过为每个指标中的每个候选项分配一个从 0 到 $n - 1$ 的排名来起作用。最终的组合分数是这些排名数的乘积。这种方法的好处在于，与

其他候选项相比，它促进了得分更平衡的候选项，同时仍然提升整体得分较高的候选项。

我们的另外两个任务，SAGA 任务 2 和动作条件任务，是分类任务而不是排序任务。我们决定使用逻辑回归作为这些任务的评分汇总器，关于这一点的理由详见附录 A。

4 数据集

我们使用三个不同的数据集来评估 IBEaM 相较于直观基线的有效性。这些数据集包括 SAGA、动作条件数据集，以及我们自己的基于 WikiHow 的数据集。SAGA 和动作条件数据集使我们能够评估 IBEaM 在分类方面的能力，而我们的基于 WikiHow 的数据集则使我们能够评估 IBEaM 在排序方面的能力。

4.1 WikiHow 程序数据集

我们想要评估 IBEaM 对大型语言模型推理能力的影响，所以选择了 WikiHow 作为我们数据集之一，因为处理这个任务需要理解隐含的世界状态。我们从 WikiHow 中创建了一个五选一的多项选择任务，每个实例包含一个正在进行的目标、若干个已完成的步骤以及五个候选的流程继续步骤。这五个候选步骤由目标步骤和四个干扰项组成。四个干扰项包括一个已经执行过的重复步骤、两个文档外步骤以及一个乱序的未来步骤。这个乱序的未来步骤是和目标及目标目标属于同一篇文章的一部分，但在执行这个步骤之前还需完成另一个步骤。在这些干扰项中，重复步骤难以正确排序，因为重复一个步骤通常是可行的（例如，重新预热烤箱），尽管它已经完成，但不朝目标前进。然而，乱序的未来步骤更具挑战性，因为它确实推进了目标（例如，在烤箱中留置土豆一小时），但大型语言模型需要识别其要求（例如，土豆在烤箱中）与世界状态（例如，土豆在桌上）之间的差距，以确定它不是当前流程的可行继续步骤。

我们为 66 个实例注释了目标、干扰项、先前完成的步骤和总体目标。由于我们的排序方法不需要训练集，因此该数据集是一个纯粹的测试集，不包含训练或验证集的划分。

4.2 总结适用性数据集

ROCStories 数据集由称为故事的实例组成，这些实例具有五个顺序事件，可以通过一个句子 (Mostafazadeh et al., 2016) 进行总结。PASTA 数据集是 ROCStories 的扩展，其将替代事件替换到故事中，这些事件可能与原本的总结句 (Ghosh et al., 2023) 配合，也可能不匹配。故事替代和目标适用性 (SAGA) 数据集是对 PASTA

的注释扩展，其中包括对 PASTA 中这些替代故事与原本总结 (Vallurupalli et al., 2024) 的适用性的人类评估。SAGA 的论文定义了多个可以在 SAGA 上测试的任务，特别是我们将进行任务 2，即一个二分类任务，用于测试模型判断总结是否适用于一个故事的能力。我们选择该数据集是因为它执行起来并不简单，但它也具有简单的宏观和微观 F1 分数评估标准。

4.3 前/后条件分类数据集

“动作条件”数据集也来源于 WikiHow (Wu et al., 2023)，但不同于专注于流程中下一步骤的排序，它定义了 WikiHow 文章中的文本段为前置条件和后置条件。在这个上下文中，前置条件关系表明一个文本段是另一个文本段的要求，而后置条件关系则表明一个文本段是另一个的结果。在一篇 WikiHow 文章中，任何没有标明前置条件或后置条件关系的文本段对，被称为具有 NULL 关系。这些关系不一定是双向的，因此，如果一个文本段是另一个的前置条件，第二个文本段不一定是第一个的前置或后置条件，除非标明如此。对于我们的评估，我们定义一个实例为一对文本段，一个标签，以及包含这两个文本段的最小文本主体。我们选择这个数据集是因为它的完成并非易事，具有简单的宏观和微观 F1 评分评价标准，并且与我们其他两个任务有足够的不同。

5 实验设置

对于所有我们的 IBEaM 实验，每次试验都使用新生成的一组 Likert 量表。即使在提示与另一次试验相同的情况下，IBEaM 和基线都会重新提示 LLM。这是因为我们观察到，即使在给定相同的提示条件下，LLM 的响应往往有细微的变化。在每个试验中，LLM 的响应被缓存以减少推理时间和成本。用于 IBEaM 和基线的 LLM 是 2024-08-06 快照的 GPT-4o (Hurst et al., 2024) 和 LLaMa 3.2 (Grattafiori et al., 2024)。

5.1 WikiHow 步骤候选排名

对于 IBEaM，我们将任务分为进度和可行性两个子任务。每一个子任务都有一个 Likert 量表为其生成，该量表随后被包含在 LLM 的对话历史中。对于每一次试验，Likert 量表都会重新生成，但在每个试验内每个实例中是保持一致的。从我们发现的最有效的分数聚合器是分别对进度和可行性进行评分排名，然后取两个数字排名的乘积。拥有最高乘积的候选者被选为预测。

对于我们的基线，我们要求大型语言模型 (LLM) 根据进行中的目标和已完成的步骤对

每个候选者进行 1 到 10 的评分。预测出的候选者是评分最高的那个。

5.2 摘要分类任务

对于 IBEaM，我们将任务拆分为总结、目标和完成子任务。类似于 WikiHow 任务，这些子任务中的每一个都为每次试验重新生成了一个李克特量表。然而，不同于 WikiHow 任务，SAGA 任务 2 是一项二元分类任务，而不是排名任务，因此我们不能通过排名评分来确定最佳预测。相反，我们使用一个不大于测试集的独立实例集来拟合逻辑回归模型。这个逻辑回归模型将数值评分作为独热编码向量输入，并预测应用性的真假结果，我们为 IBEaM 和基线拟合了不同的逻辑回归模型。完整的流程如附录 B 中的图 5 所示。

SAGA 的测试集由 512 个实例组成，每次试验都使用相同的测试集，这 512 个实例中，419 个有正标签而只有 93 个有负标签。这种标签的不平衡反映了训练集的情况，因此我们在拟合逻辑回归模型时对类别权重进行了调整。每次试验都从训练集中取出一个不同的 512 实例样本来拟合逻辑回归模型。

对于我们的基线，我们再次提示 LLM 对目标与步骤序列的适用性进行 1 到 10 的评分。与我们为 IBEaM 使用的方法类似，我们再次拟合逻辑回归模型以确定哪些评分预测适用性为真或假。对于每次试验，用于拟合基线逻辑回归模型的训练集与用于 IBEaM 的训练集相同。

5.3 前/后条件分类任务

对于 IBEaM，我们将任务分为顺序、相关性和独立性子任务。类似于 WikiHow 任务和 SAGA 任务 2，每个子任务在每次试验中都有一个 Likert 量表重新生成。动作条件任务是一个三元分类任务，而不是二元分类任务，但我们仍然可以应用与 SAGA 任务 2 相同的逻辑回归方法。同样，我们为我们的逻辑回归模型获得了一些训练实例，但它预测的不是正确或错误，而是预测前置条件、后置条件或 NULL。完整的流程如附录 B 的图 6 所示。

Action Conditions 数据集的测试集过于庞大，无法进行完整评估，因此我们在每次试验中从测试集中选取一个不同的 512 个实例样本。类似于 SAGA 任务 2，每次试验也从训练集中选取一个不同的 512 个实例样本，用于拟合逻辑回归模型。

对于我们的基线，我们提示 LLM 给出两个不同的分数，范围从 1 到 10。第一个分数是前提条件分数，第二个分数是后提条件分数。与 SAGA 任务 2 类似，这两个分数被输入到一

		Accuracy		MRR	
		IBEaM	Baseline	IBEaM	Baseline
GPT 4o	Trial 1	57.3 %	50.0 %	0.718	0.598
	Trial 2	59.1 %	49.5 %	0.751	0.610
	Trial 3	63.6 %	49.2 %	0.771	0.605
	Trial 4	58.3 %	47.7 %	0.728	0.561
	Trial 5	61.4 %	44.2 %	0.764	0.560
	Mean	59.9 %	48.1 %	0.746	0.587
LLaMa 3.2	Trial 1	26.8 %	26.0 %	0.472	0.476
	Trial 2	32.3 %	19.8 %	0.513	0.423
	Trial 3	23.8 %	17.3 %	0.464	0.380
	Trial 4	27.3 %	17.6 %	0.498	0.405
	Trial 5	27.9 %	21.5 %	0.505	0.417
	Mean	27.6 %	20.4 %	0.490	0.420

Table 1: 对每个 WikiHow 试验，IBEaM 与基准的性能概述。准确性值为平局提供部分评分。如果目标与一个干扰项得分相同，则给半分，两个干扰项得分时给三分之一，以此类推。

个经过训练的逻辑回归模型中，以确定预测结果应该是前提条件、后提条件还是 NULL。对于每次试验，基线的训练集和测试集与 IBEaM 使用的相同。

6 结果

在我们的所有实验中，我们观察到相关评估指标平均至少有小幅度提高。提升的程度因任务而异，可能归因于 LLM 在基础提示下表现良好的能力。特别是，LLM 在动作条件基线上表现不佳，这导致在应用 IBEaM 时性能提升最大。

6.1 WikiHow 步骤候选排名

如表 1 所示，当对所有 GPT-4o 实验取平均时，我们显示出准确性提升了近 12 个百分点，MRR 提升超过 0.15。LLaMa 在准确性上提高了 7 个百分点，MRR 提高了 0.7。对于该数据集，除了一次 LLaMa 3.2 实验的 MRR 得分之外，所有实验相对于基准都有一致的改进。我们还对每个干扰类型的表现进行了分析，以查看每种干扰类型是否如我们预期般困难。图 4 所示，IBEaM 和基准在每个类别上的表现都如我们预期。特别值得注意的是，每种方法产生的打分并列数量。GPT-4o 的基准，可能由于其预测基于单一的 10 分制标尺，在评估重复干扰和顺序错误干扰时有很多并列。IBEaM 在应用于 GPT-4o 时并列明显减少，这部分可能是因为它结合了多种指标，这允许比基准更精细的评分预测。即便确定性增加，它仍在所有类别中优于基准，包括当将所有并列计为错误预测和将 50 % 的并列计为错误预测时。奇怪的是，当 IBEaM 应用于 LLaMa 3.2 时，这些模式并未表现出来，反而在 OOD 实例中比基准有更多并列，而在重复和 OOD 实例中并列数量相当。

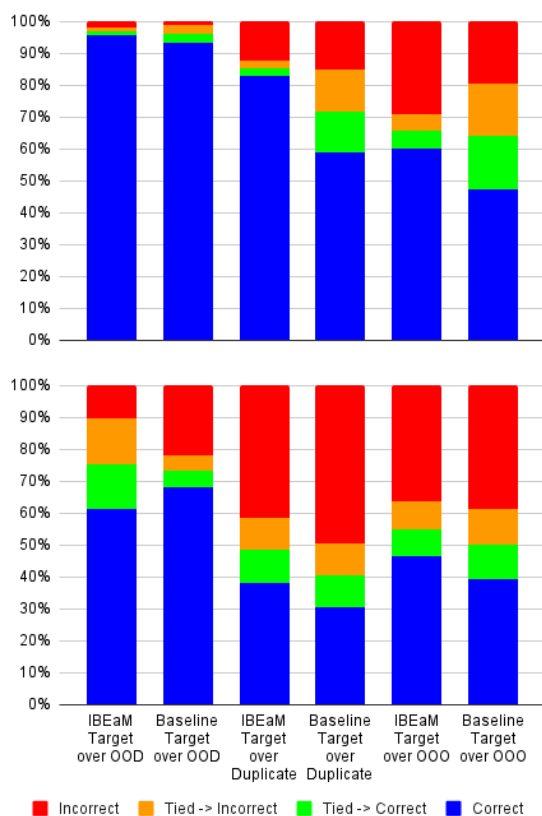


Figure 4: IBEaM 和基准在 WikiHow 数据集上的性能分解。使用 GPT-4o 和 LLaMa 3.2 的性能分别显示在上图和下图。分配的分值是整数，因此平局在正确和错误的预测之间均分。

6.2 摘要和前/后置条件分类任务

如表 2 所示，在我们进行的五个 SAGA 任务 2 试验中，所有的 F1 分数在 GPT-4o 和 LLaMa 3.2 上都显示了改善，唯一的例外是当 IBEaM 应用于 GPT-4o 时的一个微观 F1 分数。对于试验 4，微观 F1 分数低于基线，我们需要注意到测试集中存在显著的标签不平衡，这意味着如果 IBEaM 或基线倾向于将实例标记为适用，微观 F1 分数将增加而宏观 F1 分数会减少。反之亦然，所以这表明在试验 4 中 IBEaM 和基线整体表现差不多。然而，当与基线相比时，其他四个试验显示宏观和微观 F1 分数均一致提高，因此 overall IBEaM 的表现更好。

对于行动条件任务，IBEaM 在所有 GPT-4o 的试验中表现优于基线，并且在 LLaMa 3.2 的五个试验中有四个中表现优于基线，但考虑到行动条件任务是一个三分类问题，IBEaM 和基线在这个任务上的表现都不尽如人意。

7 消融研究

为了确定 IBEaM 的每个组件对整体性能提升的贡献，我们进行了两次消融研究。在第一次

		IBEaM Improvement over Baseline	
		Baseline	
SAGA Task 2	GPT 4o Trial 1	67.4 / 61.6	4.30 / 4.44
	Trial 2	64.3 / 59.5	5.47 / 4.42
	Trial 3	64.3 / 58.5	1.95 / 2.27
	Trial 4	68.9 / 62.0	-1.56 / 1.40
	Trial 5	65.6 / 60.3	8.01 / 4.11
	Mean	66.1 / 60.4	3.63 / 3.33
	LLaMa 3.2 Trial 1	48.0 / 45.9	18.36 / 10.64
	Trial 2	43.2 / 42.6	12.11 / 7.91
	Trial 3	51.4 / 48.8	9.57 / 4.95
	Trial 4	43.8 / 42.9	16.41 / 9.78
Action Conditions	Trial 5	47.7 / 45.7	10.16 / 5.65
	Mean	46.8 / 45.2	13.32 / 7.78
	GPT 4o Trial 1	46.5 / 40.5	0.39 / 1.77
	Trial 2	44.7 / 42.0	2.93 / 1.13
	Trial 3	40.2 / 36.2	13.09 / 10.45
	Trial 4	45.5 / 38.7	4.10 / 5.42
	Trial 5	32.8 / 29.9	18.55 / 18.01
	Mean	42.0 / 37.5	7.81 / 7.36
	LLaMa 3.2 Trial 1	34.0 / 32.5	2.15 / 2.63
	Trial 2	30.3 / 28.8	12.89 / 10.66
	Trial 3	28.5 / 27.8	9.18 / 7.48
	Trial 4	42.4 / 35.5	-5.86 / -2.53
	Trial 5	34.6 / 32.0	8.59 / 6.05
	Mean	33.9 / 31.3	5.39 / 4.86

Table 2: IBEaM 和基线在分类任务中的性能概述。F1 分数乘以 100 并以微观 F1/宏观 F1 表示。改进计算为 (IBEaM F1 - 基线 F1)。使用 GPT-4o 时，除一次 SAGA 任务 2 试验外，IBEaM 始终比基线表现更好。使用 LLaMa 3.2 时，除一次行动条件试验外，IBEaM 始终比基线表现更好。

消融研究中，我们移除了 IBEaM 中 Likert 量表的自校准，但仍保留了分割度量。这有助于我们确定度量分割在多大程度上提高了性能，这应该非零的，因为它允许我们从模型中提取更多的归纳偏置。在第二次消融研究中，我们移除了 IBEaM 中的分割度量，但仍执行自校准。这帮助我们确定自校准在多大程度上提高了性能，这也应该非零的，因为这是归纳偏置匹配发生的关键步骤。

7.1 移除自校准

如表 3 所示，我们看到从 IBEaM 中去除李克特量表校准步骤对平均准确率有很强的负面影响，因为它使平均准确率降低至低于基线。然而，去除校准步骤的一个显著效果是，它还降低了各次试验中准确率的标准差，使其低于未删减的 IBEaM 和基线的标准差。虽然这是一个积极的效果，但与基线相比，准确率的降低并不合理。

7.2 移除度量劈分

再次查看表 3，我们发现在评估 SAGA 任务 2 时，移除度量拆分确实降低了平均准确性。然

	IBeAM	Uncal.	Unsplit	Baseline
Trial 1	71.7 / 66.0	67.0 / 60.8	70.1 / 64.9	67.4 / 61.6
Trial 2	69.7 / 63.9	66.6 / 59.6	60.9 / 58.2	64.3 / 59.5
Trial 3	66.2 / 60.7	67.2 / 61.1	71.5 / 66.8	64.3 / 58.5
Trial 4	67.4 / 63.4	63.7 / 58.4	72.9 / 66.5	68.9 / 62.0
Trial 5	73.6 / 64.4	63.3 / 59.5	59.2 / 56.7	65.6 / 60.3
Mean	69.7 / 63.7	65.5 / 59.9	66.9 / 62.6	66.1 / 60.4
SD	3.04 / 1.92	1.91 / 1.10	6.36 / 4.81	2.05 / 1.45

Table 3: 我们对 SAGA 任务 2 使用 GPT-4o 进行的消融研究概述。F1 分数乘以 100 并表示为微观 F1/宏观 F1。大多数试验中，无分割指标的 IBeAM 表现良好，但其标准差最高。未校准的 IBeAM（具有分割指标）的标准差最低，但在大多数试验中未能超过基线。未消融的 IBeAM 的标准差比未校准的 IBeAM 高，但其准确性的平均提高使其成为一个值得的权衡。

而，更值得注意的是，与未经过消融的 IBeAM 相比，准确性在不同试验中的标准差增加了两倍多。虽然没有拆分度量的 IBeAM 平均而言比基线更准确，但其标准差是基线的三倍是不理想的。

7.3 分析

我们的两个消融研究表明，对于 IBeAM 来说，度量分割和自校准都是必不可少的，这两者使其能够在我们的任务中实现持续的高准确性。值得注意的是，自校准步骤平均提高了 IBeAM 的准确性，使其超过基线，但同时也导致标准差大幅增加。另一方面，度量分割在有无自校准的情况下，对各试验准确性的差异具有强烈的调节作用。当这两个组件结合在一起时，未消融的 IBeAM 比两个消融版本都更准确，同时具有合理标准差，尽管未经校准的 IBeAM 的平均准确性低于基线。这告诉我们 IBeAM 内部的各个组件相辅相成，这种协同作用使得 IBeAM 超过其各部分的总和。

8 结论

在本文中，我们提出了一种名为 IBeAM 的方法，该方法用于书写提示，能够自动提取和应用 LLM 的归纳偏置，以提高其分配得分用于后续排序和分类的能力。LLM 提取其自身归纳偏置的能力可能具有广泛的应用，但我们在多个背景中以具体评估标准量化了可能的性能提升。与我们的基线相比，基线要求 LLM 通过简短描述任务后给出 1 到 10 的评分，而 IBeAM 增加的复杂性在需要数值评分，以及整体排序或分类预测的背景中是值得投入的。

9

局限性

我们所有的实验都是使用 GPT-4o 进行的，这是一种在专有数据集上训练的封闭源码 LLM，以及使用 LLaMa 3.2，这是一种开源 LLM。IBeAM 并不专门依赖于 GPT-4o 或 LLaMa 3.2，但它是一种专门为与 LLM 交互而设计的技术，因此，它受到所使用的 LLM 的相同限制。

References

- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. [Unleashing the potential of prompt engineering for large language models](#). *Patterns*, page 101260.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [A closer look into using large language models for automatic evaluation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.
- Sayontan Ghosh, Mahnaz Koupaee, Isabella Chen, Francis Ferraro, Nathanael Chambers, and Niranjan Balasubramanian. 2023. [Pasta: A dataset for modeling participant states in narratives](#). *Preprint*, arXiv:2208.00329.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Shai Gretz, Alon Halfon, Ilya Shnayderman, Orith Toledo-Ronen, Artem Spector, Lena Dankin, Yannik Katsis, Ofir Arviv, Yoav Katz, Noam Slonim, and Liat Ein-Dor. 2023. [Zero-shot topical text classification with LLMs - an experimental study](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9647–9676, Singapore. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. [Gpt-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- Yen-Ting Lin and Yun-Nung Chen. 2023. [LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models](#). In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende,

Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. [Text classification via large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Sai Vallurupalli, Katrin Erk, and Francis Ferraro. 2024. [Saga: A participant-specific examination of story alternatives and goal applicability for a deeper understanding of complex events](#). Preprint, arXiv:2408.05793.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Te-Lin Wu, Caiqi Zhang, Qingyuan Hu, Alexander Spangher, and Nanyun Peng. 2023. [Learning action conditions from instructional manuals for instruction understanding](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3023–3043, Toronto, Canada. Association for Computational Linguistics.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A 逻辑回归作为评分聚合器

在分类任务中使用 IBEaM 需要我们根据分数值创建某种决策边界。根据经验，我们发现将所有子指标的一热编码向量输入到逻辑回归中是创建决策边界的最有效方法。然而，这需要为逻辑回归模型使用一个训练集，而这在排序任务中是不需要的。因此，对于 SAGA 任务 2 和动作条件任务，我们抽取训练集中的一个小

样本，并通过 IBEaM 管道处理。这些任务的标签和 LLM 的响应然后用于拟合逻辑回归模型。

虽然逻辑回归组件不是 IBEaM 的一部分，但它有助于评估 IBEaM 生成的评分相较于基准产生的评分的改进。由于我们使用 LLM 生成评分而不是直接的分类预测，我们需要一种方法将评分转换为二元分类。逻辑回归为我们完成了这一点，同时由于当评分更接近线性可分时会表现得更好，因此仍然具有信息价值。无论哪个提示策略能产生更接近线性可分的结果，就是更能确定其预测的提示策略。虽然确定性不是下游结果的直接代理，但为逻辑回归使用独立的训练集也会将评分质量的下降转化为准确性的下降。

B IBEaM 管道

图 5 和 6 分别展示了我们用于评估 SAGA 任务 2 和动作条件任务的流程设置。

C 完整分类结果

表 4 展示了表 2 的扩展版本，其中直接比较了基线和 IBEaM 的 F1 分数。

D 大语言模型对话

图 7、8 和 9 显示了 WikiHow 任务的示例对话。

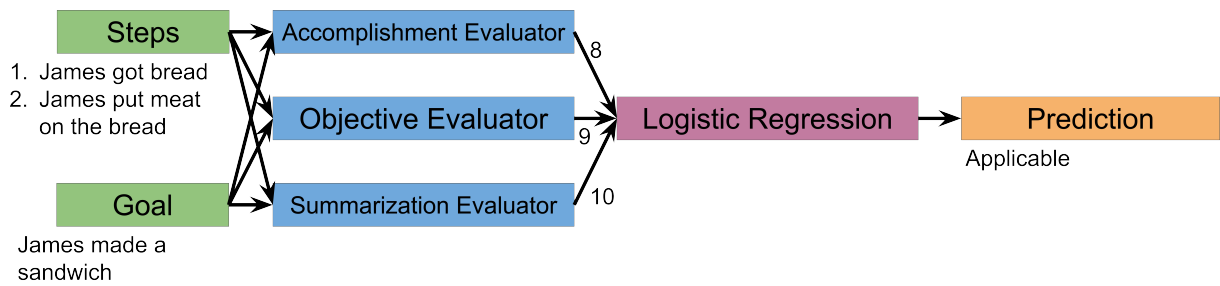


Figure 5: IBEaM 在 SAGA 任务 2 中的使用概览。

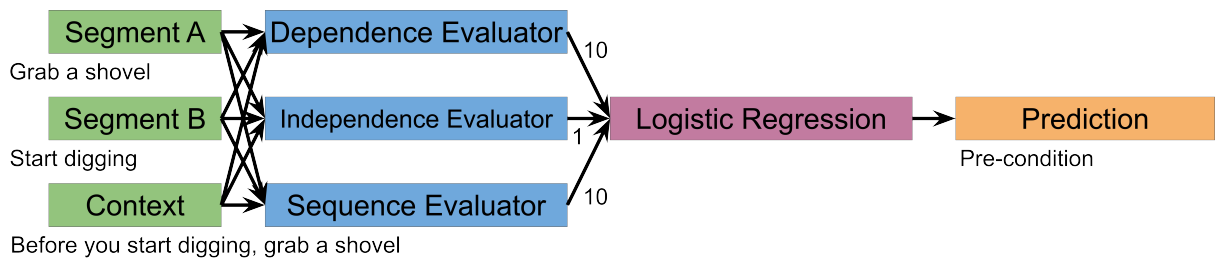


Figure 6: IBEaM 在动作条件任务中的使用概述。

		SAGA Task 2				Action Conditions			
		Micro F1		Macro F1		Micro F1		Macro F1	
		IBEaM	Baseline	IBEaM	Baseline	IBEaM	Baseline	IBEaM	Baseline
GPT 4o	Trial 1	0.717	0.674	0.660	0.616	0.469	0.465	0.423	0.405
	Trial 2	0.697	0.643	0.639	0.595	0.477	0.447	0.431	0.420
	Trial 3	0.662	0.643	0.607	0.585	0.533	0.402	0.467	0.362
	Trial 4	0.674	0.689	0.634	0.620	0.496	0.455	0.441	0.387
	Trial 5	0.736	0.656	0.644	0.603	0.514	0.328	0.479	0.299
	Mean	0.697	0.661	0.637	0.604	0.498	0.420	0.448	0.375
LLaMa 3.2	Trial 1	0.664	0.480	0.565	0.459	0.361	0.340	0.351	0.325
	Trial 2	0.553	0.432	0.505	0.426	0.432	0.303	0.394	0.288
	Trial 3	0.609	0.514	0.537	0.488	0.377	0.285	0.353	0.278
	Trial 4	0.602	0.438	0.526	0.429	0.365	0.424	0.329	0.355
	Trial 5	0.578	0.477	0.514	0.457	0.432	0.346	0.380	0.320
	Mean	0.601	0.468	0.530	0.452	0.393	0.339	0.362	0.313

Table 4: 关于 IBEaM 和基线在分类任务中的表现概述。使用 GPT-4o 时，IBEaM 除了在一次 SAGA 任务 2 的试验中例外外，一直表现优于基线。使用 LLaMa 3.2 时，IBEaM 除了在一次行动条件试验中例外外，一直表现优于基线。

User

Imagine you are in a world where you are working towards some goal and you have already done some number of steps. Design a 10 point scale that describes how doable some candidate step is. On this scale, 10 should represent the highest degree of doability and 1 should represent the lowest degree of doability.

Assistant

Creating a 10-point scale to assess the doability of a candidate step in achieving a goal requires careful consideration of various factors such as feasibility, resources, time, skills, and potential barriers. Here's a proposed scale:

- 1 - 不可行：由于当前资源和知识的限制，这一步几乎不可能实现。存在重大障碍难以克服，需要显著改变环境或资产。
- 2 - 极不可能：这一步骤面临重大的障碍，并且需要当前无法获得的资源或技能。需要重大的转变或突破才能继续进行。
- 3 - 非常困难：这一步骤只有在特定和罕见的条件下才能实现。需要大量的时间和精力投入，并且可能依赖于超出控制范围的外部因素。
- 4 - 富有挑战性：需要克服若干障碍；其中一些可能是模糊的或未知的。这一步需要大量的计划、努力，并可能需要学习新技能或获取资源。
- 5 - 中等难度：这一步需要一定的努力才能完成，但涉及处理已知的障碍。需要清晰的计划和适度的资源。存在一定的风险或不确定性。
- 6 - 颇为可行：通过勤奋工作和解决问题可以实现这一步。挑战和资源需求已确定，伴随一些可管理的风险。
- 7 - 可行：此步骤使用现有资源可以直接完成。预期的挑战是常规的，但可以通过周密的计划和执行来减轻。
- 8 - 非常可行：鉴于当前的能力和资源，轻松实现。可能需要注意小障碍，但在正常工作范围内是可以管理的。
- 9 - 非常可行：可以使用现有工具和技能轻松实现。预期的挑战或阻力很小。
- 10 - 轻松：这一步可以利用现有资源和能力顺利快速地完成。没有可以预见的挑战，结果几乎是有保障的。

This scale takes into account factors such as resource availability, skill level, time constraints, external dependencies, and potential challenges, providing a comprehensive assessment of a step's feasibility in the pursuit of a goal.

User

Imagine you are in a world where your objective is to Assemble your paper bag so you can make a paper bag and you have done the following:

- 将剪好的纸放在您面前的平面上。
- 将纸张的底边向上折叠 2 英寸（5 厘米），并用力压折痕。
- 定位上下边缘的中心点。

For the following query, do not assume that the requirements of the following step are fulfilled unless they are known to be so from the world state, and start your response with a rating from 1 to 10 on the previous doability scale that considers the gap between the current world state and the requirements of the following step: Fold the sides of the bag into place.

Assistant

Rating: 5 - Moderate Difficulty

Assessing the doability of the next step, “Fold the sides of the bag into place,” involves several considerations based on the current world state:

1. 任务理解：假设您知道根据所裁剪纸袋的形状和设计应该如何折叠边缘，您对这项任务的熟悉程度会影响其可操作性。
2. 要求和准备：纸张在一个平面上，你已经折叠了底边并找到了中心点。但是，是否存在用于侧边折叠的额外标记或指南没有说明。
3. 资源和工具：这项任务似乎涉及手工折叠，除了已提到的内容外，不需要额外的资源。如果纸张预先经过折线或有指南，会使任务更为简化。
4. 技能水平：折纸需要基本的技能，但在没有经验或指导的情况下实现精确和干净的折叠可能具有挑战性，这会影响袋子的整洁度和结构完整性。
5. 潜在障碍：在没有特定折叠线或额外指示的情况下，可能需要一些试错，这可能使任务变得复杂。

Based on these considerations, folding the sides involves moderate difficulty due to potential lack of precise guidelines and the manual dexterity required for accurate folds, though it's achievable with effort and focus.

Figure 7: 与 GPT-4o 进行对话，以确定在 WikiHow 任务中某一步的可行性。

User

Imagine you are in a world where you are working towards some goal and you have already done some number of steps. Design a 10 point scale that describes the amount of progress that some candidate step makes. On this scale, 10 should represent the highest degree of progress and 1 should represent the lowest degree of progress.

Assistant

Creating a 10-point scale to assess the progress of a candidate step towards a goal involves defining specific criteria for each level. Here's a potential scale:

1. 最低进展 (1/10): 该步骤贡献的进展可以忽略不计。更像是对以前行动的重复, 增量效益很少或没有。
2. 非常低的进展 (2/10): 这一步带来了一个小的好处或变化, 但对整体目标的影响微小且几乎不明显。
3. 进展不大 (3/10): 该步骤带来了一些进展, 但主要解决了目标的次要或外围方面, 并没有显著推动目标前进。
4. 低于平均进度 (4/10): 该步骤带来了可衡量的进展, 但它并没有显著改变当前状态或朝目标迈进的动力。
5. 适度进展 (5/10): 这一步提供了一种明确而积极的贡献, 代表了一种必要但并不突破性的进步。
6. 高于平均进展 (6/10): 该步骤通过有效解决与目标相关的重要方面或克服显著障碍, 提供了稳固的进展。
7. 显著进展 (7/10): 此步骤带来了有意义的飞跃, 解决了关键问题或为进一步发展打下了基础。
8. 高进展 (8/10): 这一步骤带来了显著的收益, 推动了可观的动力, 并使目标更加接近。
9. 非常高的进展 (9/10): 这一步实现了重大的进展, 克服了显著的障碍, 在很大程度上改变了实现目标的前景。
10. 最大进展 (10/10): 这一步代表了一次最佳移动, 产生了变革性的影响, 并推动了目标的大量完成或实现。由于其有效性, 它甚至可能重新定义目标的参数。

This scale provides a structured way to assess and compare the efficacy of different steps towards a specific goal, helping prioritize actions and allocate resources efficiently.

User

Imagine you are in a world where your objective is to Assemble your paper bag so you can make a paper bag and you have done the following:

- 将剪下的纸放在您面前的平坦表面上。
- 将纸张的下边缘向上折叠 2 英寸（5 厘米），并用力压实折痕。
- 找到顶边和底边的中心点。

For the following query, if the results of the following step have already been done in the current world state, do not count that as progress, and start your response with a rating from 1 to 10 on the previous progress scale that considers the amount of progress that the following step makes: Fold the sides of the bag into place.

Assistant

6/10

Above Average Progress : By folding the sides of the bag into place, you are addressing a crucial step in the assembly process. This action contributes effectively to the overall structure and stability of the paper bag. It represents notable advancement from the preparatory steps taken earlier, moving you closer to completing the functional shape of the bag. This progress is solid and necessary, though other steps will still be needed to achieve a fully assembled paper bag.

Figure 8: 与 GPT-4o 进行对话，以确定一项步骤在 WikiHow 任务中取得的进展。

User

Imagine you are in a world where your objective is to Assemble your paper bag so you can make a paper bag and you have done the following:

- 将剪好的纸放在你面前的平坦表面上。
- 将纸张的底边向上折叠 2 英寸（5 厘米），并用力折出折痕。
- 找出顶部和底部边缘的中心点。

For the following query, start your response with a rating from 1 to 10. If you wanted to Assemble your paper bag so you can make a paper bag, how good of a continuation is the following step? Fold the sides of the bag into place.

Assistant

8. 在将袋子的底边折叠和压平并找到中心点后，将袋子的侧边折叠到位是合乎逻辑的下一步。这一步对于形成袋子的结构是至关重要的，但可以更具体一些。为了获得更高的评价，说明中可以包括使用中心点来确保对称，并提及如何折叠或固定侧边的位置。

Figure 9: WikiHow 任务的一个示例基线提示和回应。