An Actionable Framework For Assessing Bias and Fairness in Large Language Model Use Cases

作者：Dylan Bouchard

单位：CVS Health

文章地址：https://arxiv.org/abs/2407.10853

项目地址：https://github.com/cvs-health/langfair **200+** Star

联系方式：dylan.bouchard@cvshealth.com

作者背景：目前，我是CVS Health的首席应用科学家/总监，在那里我领导人工智能研究和开源软件开发，重点是人工智能安全。我创建并维护了两个用于人工智能安全的开源Python库：UQLM，一个使用最先进的不确定性量化技术进行大模型幻觉检测的工具包；LangFair，一个用于特定用例大模型偏见/公平性评估的工具包。

摘要

大型语言模型（LLMs）可能以多种方式表现出偏见。这些偏见可能为受保护属性内的某些群体创造或加剧不公平的结果，包括但不限于**性别、种族、性取向或年龄**。在本文中，我们提出了一个决策框架，允许从业者确定在特定LLM用例中使用哪些偏见和公平性指标。为了建立这个框架，我们定义了LLMs的偏见和公平性风险，将这些风险映射到LLM用例的分类法中，然后定义各种指标来评估每种类型的风险。我们不仅仅关注模型本身，而是通过在LLM用例层面定义评估来考虑提示特定风险（不同的提示输入可能引发不同的偏见）和模型特定风险（模型本身固有的偏见特征），其特征是一个模型和一个提示群体。此外，由于所有评估指标都仅使用LLM输出来计算，我们提出的框架非常实用且易于为从业者付诸行动。为了简化实施，框架中包含的所有评估指标都在本文的配套Python工具包LangFair中提供。最后，我们的实验表明偏见和公平性在不同用例中存在显著差异，强调了用例层面评估的重要性。

1 介绍

当前大型语言模型在处理各种任务方面的多功能性使得在模型层面评估偏见和公平性变得困难。现有方法主要依赖于包含预定义提示的基准数据集、掩码标记或无掩码句子，假设这些方法能够充分捕获特定的偏见或公平性风险[Gallegos等，2023]。然而，这些评估没有考虑到提示特定风险，而这些风险已被证明会显著影响LLM产生偏见和不公平回应的可能性。此外，据我们所知，目前的文献没有提供一个框架来有效地将LLM用例与评估偏见和公平性的合适指标相匹配。

为了解决这些局限性，我们提出了一个在用例层面定义的LLM偏见和公平性评估框架。我们的框架从Saleiro等人提出的分类公平性框架中汲取灵感，通过考虑任务、提示的相关特征和利益相关者的价值观，使从业者能够将LLM用例映射到一组合适的偏见和公平性评估指标。这种评估方法的独特之处在于它采用了"自带提示"的方法，即从LLM对来自从业者用例的实际提示的回应中计算指标。我们的框架专为定义明确的用例而设计，其中提示从已知群体中抽样，从而允许针对特定应用定制的偏见和公平性评估。

为了介绍这个框架，我们首先从文献中定义了LLMs的偏见和公平性风险，并将这些风险映射到用例分类法中。对于每个风险类别，我们然后提出了各种评估指标，并讨论了它们的输入要求、计算方法、所评估的风险以及应该应用它们的情况。作为这项工作的一部分，我们还引入了各种新颖的偏见和公平性指标。具体来说，这些新指标包括面向回忆的要点评估替身（ROUGE）、双语评估替身（BLEU）和余弦相似度的反事实适应版本，以及一套基于刻板印象分类器的指标，这些指标是从类似的基于毒性分类器的指标中改编而来的。出于实用性考虑，我们将LLM偏见和公平性指标的选择限制在那些仅需要LLM生成输出进行计算的指标上。

为了简化框架的实施，本文所包含的所有偏见和公平性指标都由本文的配套Python工具包LangFair提供。在实践中，用户提供来自其用例的提示样本和他们选择的LLM，LangFair简化了LLM响应的生成过程并计算适用于其用例的相关指标。该工具包提供了一种与模型无关、用户友好的方式来为现实世界的用例实施我们的评估框架。

最后，我们进行了一系列实验来评估几个文本生成和摘要用例中的偏见和公平性。具体而言，我们构建了6个独特的用例，其特征是三组提示和两个LLM的组合。我们发现偏见和公平性在不同用例中存在显著差异，这强调了用例层面评估的重要性。

2 LLM用例的偏见和公平风险

下面我们提供几个初步定义，这些定义将在后续章节中使用。

2.1 初步定义

大型语言模型（LLM）。LLM $\mathcal{M}: \mathcal{X} \to \mathcal{Y}$ 是一个预训练的、基于Transformer的模型，它将文本序列 $X \in \mathcal{X}$ 映射到输出 $\hat{Y} \in \mathcal{Y}$，其中 $\mathcal{X}$ 表示所有可能的文本输入集合（即用户提示），而 $\hat{Y}$ 的形式取决于具体的LLM和用例 。设 θ 为 M 的参数，使得 $\hat{Y} = \mathcal{M}(X; \theta)$。

提示群体。提示群体，记为 $\mathcal{P}_X$，是LLM输入（用户提示）的集合。为了刻画定义明确的用例，我们随后提及"已知的提示群体"，表明从业者拥有关于提示领域的信息，并能够从 $\mathcal{P}_X$ 中抽取具有代表性的样本。例如，一个提示群体可能由临床笔记组成，其中每个单独的提示包括一组笔记，附有让LLM生成摘要的具体指令。

大型语言模型用例。 LLM用例由一个LLM $\mathcal{M}(X; \theta)$ 和一个提示群体 $\mathcal{P}_X$ 所刻画。为了记号简洁，LLM用例在此后将记为 $(\mathcal{M}, \mathcal{P}_X)$。LLM用例基于 $\mathcal{M}(X; \theta)$ 从提示群体 $\mathcal{P}_X$ 抽取的 N 个提示样本 $X_1, \ldots, X_N$ 生成的有限回应集合进行评估。

受保护属性群体。受保护属性群体 $G \in \mathcal{G}$ 表示由共同身份特征所刻画的人群子集，其中 $\mathcal{G}$ 是一个分割 。

受保护属性群体词典。 受保护属性群体词典 $A \in \mathcal{A}$ 是对应于受保护属性群体 $G \in \mathcal{G}$ 的词汇集合。在此后的表述中，为了记号简洁，$\mathcal{G}$（类似地，$\mathcal{A}$）既可以表示每个 $G \in \mathcal{G}$（类似地，每个 $A \in \mathcal{A}$）的分割，也可以表示它们的并集。受保护属性群体"男性"的受保护属性词汇为 { 'he', 'son', 'his', 'him', 'father', 'man', 'boy', 'himself', 'male', 'brother', 'sons', 'fathers', 'men', 'boys', 'males', 'brothers', 'uncle', 'uncles', 'nephew', 'nephews' }。考虑男性对女性的例子是按照 COBS}和 holistic等先前研究建立的惯例使用的。需要强调的是，这种用法纯粹是为了保持一致性，并不意图暗示只有两种性别。我们承认并尊重更广泛的性别身份谱系。

反事实输入对。反事实输入对是一对提示 X' 和 X''，它们在各个方面都相同，只是前者提及受保护属性群体 G'，而后者提及受保护属性群体 G'' 。对于LLM用例 $(\mathcal{M}, \mathcal{P}_X)$，反事实输入对的评估集记为 $(X'_1, X''_1), \ldots, (X'_N, X''_N)$。为了创建每一对，从包含受保护属性词典 $\mathcal{A}$ 中词汇的提示子集中抽取提示，即 $\mathcal{P}_{X|\mathcal{A}} = \{X : X \in \mathcal{P}_X, X \cap \mathcal{A} \neq \emptyset\}$，并通过反事实替换获得反事实变体。这里，反事实替换指使用词汇替换将对一个群体的提及替换为对应另一个群体的类似词汇。例如，$(\text{'then he went to the store'}, \text{'then she went to the store'})$ 将是性别的反事实输入对的一个例子。

通过无感知实现公平 (FTU)。 给定受保护属性词典 $\mathcal{A}$，如果对于每个 $X \in \mathcal{P}_X$，都有 $X \cap \mathcal{A} = \emptyset$，则LLM用例 $(\mathcal{M}, \mathcal{P}_X)$ 满足FTU。简而言之，FTU 意味着LLM用例的所有提示都不包含任何受保护属性词汇的提及 。

2.2 大模型偏见与公平性风险

在本节中，我们定义了适用于LLM的各种偏见和公平性风险，并定义了相应的理想标准。具体而言，这些风险包括**毒性、刻板印象、反事实公平性和分配性伤害。**

2.2.1 毒性

我们将有毒文本定义为任何攻击性语言，该语言 1) 对社会群体发起攻击、发出威胁或煽动仇恨或暴力，或 2) 包括使用贬义词汇、侮辱或任何其他专门针对和贬低社会群体的表达形式。为了将此形式化，我们引入了一个相应的理想标准，称为非毒性。

非毒性。设 $\mathcal{T}$ 表示所有有毒短语的集合。如果对于每个 $X \in \mathcal{P}_X$，都有 $\mathcal{M}(X; \theta) \cap \mathcal{T} = \emptyset$，则LLM用例 $(\mathcal{M}, \mathcal{P}_X)$ 表现出非毒性。

2.2.2 刻板印象

刻板印象是在LLM背景下应该考虑的一种重要社会偏见类型 。我们遵循定义，将刻板印象定义为对受保护属性群体的负面概括，通常通过各个群体与刻板印象术语关联频率的差异来反映。由提出的相应理想标准被称为平等群体关联。

平等群体关联 。对于两个受保护属性群体 G', G'' 和一个中性词集合 W，如果对于每个 $w \in W$，都有 $P(w \in \hat{Y} | \hat{Y} \cap A' \neq \emptyset) = P(w \in \hat{Y} | \hat{Y} \cap A'' \neq \emptyset)$，则LLM用例 $(\mathcal{M}, \mathcal{P}_X)$ 满足平等群体关联。简而言之，平等群体关联要求 W 中的每个中性词在 $\mathcal{M}$ 的输出中出现的概率相等，无论提及哪个受保护属性群体。

2.2.3 反事实公平性

在许多情况下，LLM因输入提示中包含不同的受保护属性词汇而产生实质性不同的输出是不可取的，在其他条件相等的情况下。遵循先前的工作，我们将这个概念称为反事实公平性的（缺乏）。根据上下文和利益相关者的价值观，从业者可能希望评估LLM用例在提示中包含不同受保护属性词汇时导致的整体内容或情感差异。下面，我们提出相应的公平性理想标准，称为**反事实不变性。**

反事实不变性。 对于两个受保护属性群体 G', G''，如果对于指定的不变性度量 $\upsilon(\cdot, \cdot)$，不变性度量的期望值小于某个容忍水平 ϵ：

$$\mathbb{E}[\upsilon(\mathcal{M}(X'; \theta), \mathcal{M}(X''; \theta))] \leq \epsilon,$$

其中 (X', X'') 是对应于 G', G'' 的反事实输入对 \citep{survey}，则LLM用例 $(\mathcal{M}, \mathcal{P}_X)$ 满足反事实不变性。这个修改后的反事实不变性定义放宽了严格相等性。注意，一些评估反事实不变性的度量评估的是相似性而非差异性，这意味着不等式将被颠倒。

2.2.4 分配性伤害

分配性伤害，将其定义为资源或机会在不同受保护属性群体间的不平等分配，在机器学习公平性文献中得到了广泛研究。在这项工作中，我们基于群体公平性来测量分配性伤害，定义如下。

群体公平性。给定两个受保护属性群体 G', G'' 和一个容忍水平 ϵ，如果

$$|B(\mathcal{M}(X; \theta) | G') - B(\mathcal{M}(X; \theta) | G'')| \leq \epsilon,$$

其中 B 是应用于 $\mathcal{M}$ 的统计性能度量（例如假阴性率），以受保护属性群体成员身份为条件，则LLM用例 $(\mathcal{M}, \mathcal{P}_X)$ 满足群体公平性。这里，以 G 为条件意味着在输入提示的子集上计算 B，该子集要么直接提及群体 G，要么在个人级别提示粒度的情况下，对应于属于群体 G 的个体。注意，B 的选择将取决于上下文和利益相关者的价值观。

2.3 将偏见和公平性风险映射到LLM用例

在第2.1节中，我们基于模型和已知的提示群体来刻画LLM用例。在这里，我们将用例分为三个基于任务的组别：**1) 文本生成和摘要，2) 分类，和 3) 推荐。**注意，我们的用例分割是按任务进行的，这可以通过各种方式控制，如LLM的微调、通过少样本提示提供示例，或在系统或用户提示中包含指令。我们提出的框架旨在用于大规模应用，其中生成响应的数量使得详尽的人工审查变得不现实。对于从业者手动评估每个生成输出的场景，如果与偏见和公平性相关的担忧可以通过审查输出的个人有效解决，那么我们提出的评估可能是不必要的。描述和示例在表1中提供。

使用样例分类	描述	例子	偏见/公平性风险
文本生成和摘要	LLM生成的文本输出不受预定义的类集或列表元素的约束	创建针对个人的个性化联系信息；总结临床笔记	毒性，刻板印象，反事实公平性
分类	LLM在一组预定义的类中对文本输入进行分类	对客户支持咨询的意向进行分类，以便分配协助；将客户反馈分为正面或负面，以分配后续跟进	分配性伤害
推荐	LLM产生推荐集合	生成推荐产品列表；生成推荐新闻文章列表	反事实公平

2.3.1 文本生成和摘要

我们首先考虑大语言模型(LLM)生成文本输出但不局限于预定义类别集合（如正面vs负面）或列表元素（如推荐产品）的使用场景。为了简洁起见，我们将这类使用场景统称为"文本生成和摘要"，但要说明的是，这个类别还可以包含机器翻译、问答等其他任务。此类使用场景的一个例子是使用大语言模型为客户外联撰写个性化消息。这一类别的使用场景存在在输出中生成有害文本的风险。此外，如果这些使用场景未能满足FTU（公平待遇原则），即提示中包含对受保护属性的提及，它们还会面临延续刻板印象或表现出反事实不公平的风险。

2.3.2 分类

大语言模型已被广泛用于文本分类任务。在偏见和公平性的背景下，区分文本输入是否能够映射到受保护属性是很重要的，这种映射可能通过直接提及受保护属性群体来实现，或者在个人层面提示粒度的情况下，对应于属于特定受保护属性群体的个人。例如，使用大语言模型将客户反馈分类为正面或负面，以便分配适当的后续处理，这就是个人层面分类使用场景的一个例子。与机器学习中传统的个人层面分类问题类似，这些使用场景存在分配性伤害的风险。另一方面，不涉及个人层面数据且满足FTU（公平待遇原则）的分类使用场景则不会面临这些偏见和公平性风险。

2.3.3 推荐

推荐是大语言模型的另一个潜在应用，例如使用大语言模型向客户推荐产品。研究表明，当接触到受保护属性信息时，用作推荐引擎的大语言模型可能会产生歧视性行为。因此，如果大语言模型推荐使用场景不满足FTU（公平待遇原则），它们就会面临反事实不公平的风险。

3 偏见和公平性评估指标

我们提出的框架涵盖三个不同的用例类别：1）文本生成和摘要，2）分类，以及3）推荐。对于每个类别，我们提出了多种评估指标来解决适用的偏见和公平性风险。出于实用性考虑，我们将指标的选择限制在仅需要大语言模型生成输出进行计算的指标上。由于输入要求的实际限制，我们在框架中省略了基于嵌入的指标（使用大语言模型的单词或句子隐向量表示进行计算和基于概率的指标（利用大语言模型的预测词元概率））。虽然这些指标需要访问大语言模型的上游架构，但我们框架中包含的指标不需要。重要的是，我们注意到专注于下游任务的指标（与本框架中纳入的指标一致）已被证明比从嵌入或词元概率衍生的指标更可靠。为确保我们的指标定义准确反映框架的用例特定性质，我们将每个指标置于从已知提示总体$\mathcal{P}_X$中抽取的大小为N的评估样本的上下文中。

3.1 文本生成和摘要用例的指标

我们根据第2.2节中概述的适用偏见和公平性风险，对文本生成和摘要用例的指标进行分类。具体包括毒性指标、刻板印象指标和反事实公平性指标。毒性指标利用预训练的毒性分类器（如Perspective API https://perspectiveapi.com）为大语言模型的输出分配毒性分数。刻板印象指标评估刻板印象词汇与受保护属性词汇的相对共现情况，或利用预训练的刻板印象分类器为大语言模型的输出分配刻板印象分数。最后，反事实公平性指标通过因果视角评估公平性，利用在其他方面相同的提示中受保护属性群体提及的变化。各类别中的选定指标详述如下。

3.1.1 毒性指标

毒性指标利用预训练的毒性分类器$T : \mathcal{Y} \to [0, 1]$，将文本输入映射到毒性分数。在此，我们在大语言模型用例$(\mathcal{M}, \mathcal{P}_X)$的背景下定义毒性指标。具体而言，这些指标基于从$\mathcal{P}_X$中抽取的提示样本$X_1, \ldots, X_N$，在由$\mathcal{M}$生成的每个提示$m$个响应的集合$\{\hat{Y}_{ij}\}_{j=1}^{m}$上进行计算。在实践中，同一提示的响应变化可以通过为大语言模型选择非零温度或利用不同的随机种子来实现。下面，我们定义三个毒性指标：**期望最大毒性、毒性概率和毒性比例。**这三个指标的取值范围均为0到1，较小的值表示更高程度的公平性。

期望最大毒性 (EMT) . EMT估计前m个生成结果中预测毒性概率的最大值：

$$EMT = \frac{1}{N} \sum_{i=1}^{N} \max_{1 \le j \le m} T(\hat{Y}_{ij}).$$

注意该指标的标准选择是$m = 25$。

毒性概率 (TP) . TP计算为在前m个生成结果中至少有一个毒性预测（$T(\hat{Y}) \ge 0.5$）的经验概率：

$$TP = \frac{1}{N} \sum_{i=1}^{N} I(\max_{1 \le j \le m} T(\hat{Y}_{ij}) \ge 0.5).$$

与EMT一样，该指标的标准选择是$m = 25$。

毒性比例 (TF) . TF定义为被分类为毒性的生成结果的比例：

$$TF = \frac{1}{N} \frac{1}{m} \sum_{i=1}^{N} \sum_{j=1}^{m} I(T(\hat{Y}_{ij}) \ge 0.5),$$

该指标有效估计了$\mathcal{M}$在来自$\mathcal{P}_X$的提示上生成的响应包含毒性文本的可能性。注意虽然该指标的标准选择是$m = 1$，但如果采样大的N不可行，在实践中可能更倾向于选择较大的m值。

3.1.2 刻板印象指标

刻板印象指标旨在识别大语言模型输出中可能存在的与受保护属性相关的有害刻板印象。由于这些指标依赖于受保护属性群体的提及，如果大语言模型用例满足FTU（公平待遇原则），这些指标可能就不必要了。在刻板印象指标中，我们区分基于受保护属性词汇与刻板印象词汇共现的指标，以及利用刻板印象分类器的指标。

暂时跳过

3.1.2.1 基于共现的指标

在本节中，我们概述了一组基于受保护属性词汇与相关刻板印象词汇相对共现来评估刻板印象风险的指标。这些指标有效评估了第2节中定义的平等群体关联的满足程度。我们定义了两个基于共现的刻板印象指标：**共现偏见分数和刻板印象关联。**

共现偏见分数 (COBS) . 给定两个受保护属性群体G', G''及其相关的受保护属性词汇集合A', A''，一个刻板印象词汇集合W，一个停用词集合$\mathcal{S}$，以及一个大语言模型用例$(\mathcal{M}, \mathcal{P}_X)$，COBS的完整计算如下：

$$cooccur(w, A|\hat{Y}) = \sum_{w_j, w_k \in \hat{Y}, w_j \neq w_k} I(w_j = w) \cdot I(w_k \in A) \cdot \beta^{dist(w_j, w_k)}$$

$$P(w|A) = \frac{\sum_{i=1}^{N} cooccur(w, A|\hat{Y}_i) / \sum_{i=1}^{N} \sum_{\tilde{w} \in \hat{Y}_i} cooccur(\tilde{w}, A|\hat{Y}_i) \cdot I(\tilde{w} \notin \mathcal{S} \cup \mathcal{A})}{\sum_{i=1}^{N} \sum_{a \in A} C(a, \hat{Y}_i) / \sum_{i=1}^{N} \sum_{\tilde{w} \in \hat{Y}_i} C(\tilde{w}, \hat{Y}_i) \cdot I(\tilde{w} \notin \mathcal{S} \cup \mathcal{A})}$$

$$COBS = \frac{1}{|W|} \sum_{w \in W} \log \frac{P(w|A')}{P(w|A'')},$$

其中$C(x, \hat{Y}_i)$表示x在$\hat{Y}_i$中的计数，$dist(w_j, w_k)$表示w_j和w_k之间的词元数量。上述共现函数$cooccur(w, A|\hat{Y})$计算每次w在$\hat{Y}$中出现时，在以w为中心的上下文窗口内找到的来自A的词汇的加权计数。虽然引入了该指标的两个版本——一个使用固定上下文窗口，另一个使用无限上下文窗口——本框架只纳入了无限上下文窗口版本。在他们的工作中，COBS使用$\beta = 0.95$。注意对于$\tilde{w} \in \mathcal{S} \cup \mathcal{A}$，函数$cooccur(\tilde{w}, A|\hat{Y}_i)$和$C(\tilde{w}, \hat{Y}_i)$被乘以零，以便从这些计数中排除停用词和受保护属性词汇。简而言之，COBS计算大语言模型$\mathcal{M}$生成包含$w \in W$与A'相对于A''共现的输出的相对可能性。该指标的可能取值范围为$(-\infty, \infty)$，值越接近0表示公平性程度越高。

刻板印象关联 (SA) . 考虑一组受保护属性群体$\mathcal{G}$，相关的受保护属性群体词典集合$\mathcal{A}$，以及相关的刻板印象词汇集合W。此外，设$C(x, \hat{Y})$表示词汇x在输出$\hat{Y}$中出现的次数，P^{ref}表示参考分布，TVD表示总变差距离。\footnote{\cite{holistic}推荐的参考分布是均匀分布。总变差距离衡量概率分布之间的距离。}对于给定的大语言模型$\mathcal{M}(X; \theta)$和从$\mathcal{P}_X$中抽取的提示样本$X_1, \ldots, X_N$，SA的完整计算如下：

$$\gamma(w|A') = \sum_{a \in A'} \sum_{i=1}^{N} C(a, \hat{Y}_i) I(C(w, \hat{Y}_i) > 0)$$

$$\pi(w|A') = \frac{\gamma(w|A')}{\sum_{A \in \mathcal{A}} \gamma(w|A)}$$

$$P^{(w)} = \{\pi(w|A') : A' \in \mathcal{A}\}$$

$$SA = \frac{1}{|W|} \sum_{w \in W} TVD(P^{(w)}, P^{\text{ref}}).$$

简言之，SA衡量一组刻板印象关联词汇在受保护属性群体中的相对共现情况。注意虽然COBS和SA都评估平等群体关联，但COBS采用成对计算，而SA采用属性级计算。SA的取值范围为0到1，较小的值表示更高的公平性。

3.1.2.2 利用刻板印象分类器的指标

研究表明，刻板印象分类器可以成为评估大语言模型用例中刻板印象风险的有效工具。stereoclass使用https://huggingface.co/Narrativa/distilroberta-finetuned-stereotype-detection作为其分类器。我们通过扩展第3.1.1节中概述的毒性指标，引入了三个刻板印象指标，利用预训练的刻板印象分类器$St : \mathcal{Y} \to [0, 1]$，而非毒性分类器。具体而言，这些指标包括：**期望最大刻板印象、刻板印象概率和刻板印象比例**。这三个指标的取值范围均为0到1，较小的值表示更高程度的公平性。我们在下面定义这些指标。

期望最大刻板印象 (EMS). EMS与EMT类似，估计前m个生成结果中预测刻板印象概率的最大值：

$$EMS = \frac{1}{N} \sum_{i=1}^{N} \max_{1 \leq j \leq m} St(\hat{Y}_{ij}).$$

遵循EMT的惯例，实践者可能希望该指标使用$m = 25$。

刻板印象概率 (SP). 与TP类似，SP衡量在前m个生成结果中至少有一个刻板印象预测（$St(\hat{Y}) \geq 0.5$）的经验概率：

$$SP = \frac{1}{N}\sum_{i=1}^{N} I(\max_{1\leq j\leq m} St(\hat{Y}_{ij}) \geq 0.5),$$

为了与TP的惯例保持一致，实践者可能希望该指标使用$m = 25$。

刻板印象比例 (SF). SF作为TF的扩展，衡量被预测包含刻板印象的生成结果的比例：

$$SF = \frac{1}{N}\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m} I(St(\hat{Y}_{ij}) \geq 0.5),$$

有效估计了$\mathcal{M}$在来自$\mathcal{P}_X$的提示上生成的响应包含刻板印象的可能性。注意虽然类似毒性指标EMT的标准选择是$m = 1$，但如果采样大的N不可行，在实践中可能更倾向于选择较大的m值。

3.1.3 反事实公平性指标

反事实指标旨在评估当输入提示中提及不同受保护属性时，在其他条件相等的情况下大语言模型输出的差异。给定两个受保护属性群体G', G''，我们在大语言模型用例$(\mathcal{M}, \mathcal{P}_X)$的背景下定义这些指标。特别地，这些指标在由$\mathcal{M}$生成的反事实响应对样本$(\hat{Y}_1', \hat{Y}_1''), \ldots, (\hat{Y}_N', \hat{Y}_N'')$上进行评估，这些响应对来自从$\mathcal{P}_{X|\mathcal{A}}$中抽取的反事实输入对样本$(X_1', X_1''), \ldots, (X_N', X_N'')$。在实践中，反事实替换可以通过利用一个受保护属性群体词典到另一个的映射来实现。例如，女性到男性的词典映射可以包括如下替换：{'she': 'he', 'hers': 'his', 'her': 'him', 'herself': 'himself', 'female': 'male', 'females': 'males', 'woman': 'man', 'women': 'men', 'girl': 'boy', 'girls': 'boys', 'daughter': 'son', 'daughters': 'sons', 'mother': 'father', 'mothers': 'fathers', 'sister': 'brother', 'sisters': 'brothers', 'aunt': 'uncle', 'aunts': 'uncles', 'niece': 'nephew', 'nieces': 'nephews', 'lady': 'gentleman', 'ladies': 'gentlemen', 'grandmother': 'grandfather', 'grandmothers': 'grandfathers'}。需要注意的是，映射不需要详尽无遗，只要有足够的覆盖率来生成大量反事实输入对样本即可。注意，在大的N不可行的情况下，实践者可以选择为每个反事实输入对生成多个响应对，就像第3.1.1节中毒性指标所做的那样。

这些指标，我们将其分类为**反事实相似度指标和反事实情感指标**，分别通过利用反事实输入对之间观察到的大语言模型输出变化来量化文本相似性和情感的差异。由于它们依赖于输入提示中受保护属性的提及，如果大语言模型用例满足FTU，则无需使用这些指标。

3.1.3.1 反事实相似度

反事实相似性指标根据指定的不变性度量υ衡量从反事实输入对生成的输出的相似性，即$\upsilon(\mathcal{M}(X';\theta), \mathcal{M}(X'';\theta))$。这些指标有效评估大语言模型用例是否满足第2.2节中定义的反事实不变性属性。υ的一个例子是精确匹配，但一些研究认为该指标过于严格。我们引入了三个不那么严格的反事实相似性指标：**\textit{反事实ROUGE-L}、\textit{反事实BLEU}和\textit{反事实余弦相似性}**，它们是最先进文本相似性指标的扩展。前两个使用词元序列重叠评估相似性，取值范围为0到1。第三个使用句子嵌入评估相似性，取值范围为-1到1。对于每个指标，较大的值表示更高程度的公平性。

反事实ROUGE-L (CROUGE-L). 我们引入CROUGE-L，定义为反事实生成输出对上ROUGE-L分数的平均值。CROUGE-L的完整计算如下：

$$r_i' = \frac{LCS(\hat{Y}_i', \hat{Y}_i'')}{len(\hat{Y}_i')}$$

$$r_i'' = \frac{LCS(\hat{Y}_i'', \hat{Y}_i')}{len(\hat{Y}_i'')}$$

$$CROUGE\text{-}L = \frac{1}{N}\sum_{i=1}^{N}\frac{2r_i'r_i''}{r_i'+r_i''},$$

其中$LCS(\cdot,\cdot)$表示两个大语言模型输出之间词元的最长公共子序列，$len(\hat{Y})$表示大语言模型输出中的词元数量。CROUGE-L指标有效使用ROUGE-L通过相对于生成文本长度的最长公共子序列(LCS)来评估相似性。

鉴于其依赖于匹配词元序列，实践者在计算CROUGE-L之前应该在反事实输出对中掩码受保护属性词汇。例如，假设对于反事实输入对$(\hat{X}', \hat{X}'') =$（What did he do next', What did she do next），大语言模型生成输出对$(\hat{Y}', \hat{Y}'') =$（then he drove his car to work, `then she drove her car to work'）。在这种情况下，这两个响应实际上是相同的。掩码词元$\{$he', she', his', her'$\}$在计算上实现了这一点。

反事实BLEU (CBLEU). 我们将CBLEU定义为反事实生成输出对上BLEU分数的平均值。CBLEU的完整计算如下：

$$precision_b(\hat{Y}_i', \hat{Y}_i'') = \frac{\sum_{snt\in\hat{Y}_i'}\sum_{b\text{-}gram\in snt}\min(C(b\text{-}gram, \hat{Y}_i'|\hat{Y}_i''), C(b\text{-}gram, \hat{Y}_i''))}{\sum_{\tilde{snt}\in\hat{Y}_i'}\sum_{b\text{-}gram\in\tilde{snt}}C(b\text{-}gram, \hat{Y}_i')}$$

$$BLEU(\hat{Y}_i', \hat{Y}_i'') = \min(1, \exp\{1-\frac{len(\hat{Y}_i'')}{len(\hat{Y}_i')}\})(\prod_{b=1}^{4}precision_b(\hat{Y}_i', \hat{Y}_i''))^{1/4}$$

$$CBLEU = \frac{1}{N}\sum_{i=1}^{N}\min(BLEU(\hat{Y}_i', \hat{Y}_i''), BLEU(\hat{Y}_i'', \hat{Y}_i')),$$

其中snt表示大语言模型输出中的一个句子，$len(\hat{Y})$表示大语言模型输出中的词元数量，$C(b\text{-}gram, \hat{Y}_i')$表示$b\text{-}gram$在$\hat{Y}_i'$中出现的次数，$C(b\text{-}gram, \hat{Y}_i'|\hat{Y}_i'')$表示在$\hat{Y}_i''$中也出现的前提下$b\text{-}gram$在$\hat{Y}_i'$中出现的次数\citep{bleu1, google_bleu}。为了实现对称性，我们在平均之前取每个反事实对的两个BLEU分数中的最小值。出于与CROUGE-L相同的原因，实践者在计算CBLEU之前应该在反事实输出对中掩码受保护属性词汇。

反事实余弦相似性 (CCS). 给定句子转换器$\mathbf{V}: \mathcal{Y}\to\mathbb{R}^d$，我们将CCS定义为：

$$CCS = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{V}(Y_i')\cdot\mathbf{V}(Y_i'')}{\|\mathbf{V}(Y_i')\|\|\mathbf{V}(Y_i'')\|},$$

即大语言模型用例中反事实生成输出对之间余弦相似性的平均值。

3.1.3.2 反事实情感偏差

反事实情感度量用于测量反事实生成的输出对之间的情感一致性。为了实现这一目标，这些度量利用预训练的情感分类器 $Sm: \mathcal{Y}\to[0,1]$。我们概述了两种反事实情感度量：由CSB提出的严格反事实情感公平性，以及该度量的扩展版本弱反事实情感公平性。CSB使用Google Cloud情感API和基于BERT的情感分类器。两种度量的取值范围均为 $[0,1]$，数值越小表示公平性程度越高。

严格反事实情感公平性 (SCSP) 。 SCSP计算应用于反事实生成的大语言模型输出的情感分类器输出分布之间的Wasserstein-1距离：

$$SCSP = \mathbb{E}_{\tau\sim\mathcal{U}(0,1)}|P(Sm(\hat{Y}')>\tau) - P(Sm(\hat{Y}'')>\tau)|,$$

其中 $\mathcal{U}(0,1)$ 表示均匀分布。上式中，$\mathbb{E}_{\tau\sim\mathcal{U}(0,1)}$ 是在由 $\mathcal{M}$ 生成的反事实响应对样本 $(\hat{Y}_1', \hat{Y}_1''), \ldots, (\hat{Y}_N', \hat{Y}_N'')$ 上经验计算得出的，这些样本来自从 $\mathcal{P}_{X|\mathcal{A}}$ 中抽取的反事实输入对样本 $(X_1', X_1''), \ldots, (X_N', X_N'')$。

弱反事实情感公平性 (WCSP)。我们引入WCSP，定义为情感分类器应用于反事实生成的大语言模型输出对时预测情感率的差异。给定用于二值化情感得分的阈值 τ，该度量定义如下：

$$WCSP = |\frac{1}{N}\sum_{i=1}^{N} I(Sm(\hat{Y}_i') > \tau) - \frac{1}{N}\sum_{i=1}^{N} I(Sm(\hat{Y}_i'') > \tau)|.$$

在实践中，从业者可以根据利益相关者的价值观和所使用的情感分类器选择合适的 τ 值。

3.2 分类用例的度量

众所周知，分类模型可能对某些受保护属性群体产生不公平的结果。设分类LLM用例定义为被赋予分类任务的LLM，记为$\mathcal{M}^{(c)}$，以及提示词总体$\mathcal{P}_X$。在此，我们提出针对二元分类用例的度量，其中$\mathcal{M}^{(c)} : \mathcal{X} \to \{0, 1\}$，需要注意的是，评估多类分类的公平性是从二元情况的直接扩展。

在本节的其余部分，我们假设给定分类LLM用例的$\mathcal{P}_X$中的每个提示词对应一个受保护属性群体。在此假设下，传统的机器学习公平性度量可以被应用。相应地，我们在二元预测$\hat{Y}_1, \ldots, \hat{Y}_N$上定义这些度量，这些预测由提示词样本$X_1, \ldots, X_N \in \mathcal{P}_X$生成，一些度量还包含相应的真实值$Y_1, \ldots, Y_N$。这些度量有效地评估了两个受保护属性群体$G'$和$G''$之间的群体公平性（见第\ref{sec:bias_def}节），统计结果度量B的选择取决于利益相关者的价值观（例如，假阴性与假阳性的相对成本）。

我们区分了表示公平性度量（仅使用预测计算）和基于错误的公平性度量（使用预测和真实值计算）。每个公平性度量测量一对群体级度量之间的绝对差异。此计算产生0到1之间的值范围，较小的值表示更高的公平性水平。请注意，虽然我们考虑的群体公平性是受保护属性群体之间的差异，但类似的度量也可以作为比率计算。

3.2.1 二元分类的表示公平性度量

表示公平性度量旨在确定受保护属性群体是否在分类器生成的正向预测中得到充分表示。我们建议从业者将这套度量保留给群体级预测患病率（即属于正类预测的比例）应该大致相等的分类LLM用例。例如，在预测哪些申请人符合工作资格时，可能希望针对男性和女性的正向预测率相等。然而，在预测某些疾病时，例如，这可能不是期望的模型行为。在我们的框架中，我们包含一个表示公平性度量，**人口统计平等**，它测量群体级预测患病率的绝对差异。

人口统计平等 (DP)。DP计算群体级预测患病率的绝对差异：

$$DP = |P(\hat{Y} = 1|G = G') - P(\hat{Y} = 1|G = G'')|,$$

其中$\hat{Y}$表示模型预测，$P(\cdot)$表示基于从$\mathcal{P}_X$抽取的样本提示词生成的预测的经验概率。

3.2.2 二元分类的基于错误的公平性度量

基于错误的公平性度量旨在确定受保护属性群体之间是否存在模型性能差异。为了解决基于错误的公平性问题，我们在框架中包含了两个专注于假阴性的度量：**假阴性率差异**和**假遗漏率差异**，以及两个专注于假阳性的度量：**假阳性率差异**和**假发现率差异**。根据aequitas，我们建议从业者评估对于分配辅助性（惩罚性）干预措施的用例，评估各群体间假阴性（阳性）的差异。这些度量定义如下。

假阴性率差异 (FNRD)。 FNRD测量群体级假阴性率的绝对差异：

$$FNRD = |P(\hat{Y} = 0|Y = 1, G = G') - P(\hat{Y} = 0|Y = 1, G = G'')|,$$

其中Y表示对应于$\hat{Y}$的真实值，$P(\cdot)$表示基于从$\mathcal{P}_X$抽取的提示词样本生成的预测的经验概率。注意假阴性率测量实际阳性$(Y = 1)$被错误分类为阴性$(\hat{Y} = 0)$的比例。FNRD等价于EOP提出的等机会差异度量。

假遗漏率差异 (FORD) 。 FORD测量群体级假遗漏率的绝对差异：

$$FORD = |P(Y = 1|\hat{Y} = 0, G = G') - P(Y = 1|\hat{Y} = 0, G = G'')|,$$

其中Y表示对应于$\hat{Y}$的真实值，$P(\cdot)$表示基于从$\mathcal{P}_X$抽取的提示词样本生成的预测的经验概率。假遗漏率不专注于实际阳性，而是计算预测阴性($\hat{Y} = 0$)被误分类的百分比。因此，与FNRD类似，较高的FORD表示各群体间假阴性可能性的更大差异。

假阳性率差异 (FPRD)。 FPRD测量群体级假阳性率的绝对差异：

$$FPRD = |P(\hat{Y} = 1|Y = 0, G = G') - P(\hat{Y} = 1|Y = 0, G = G'')|,$$

其中Y表示对应于$\hat{Y}$的真实值，$P(\cdot)$表示基于从$\mathcal{P}_X$抽取的提示词样本生成的预测的经验概率。注意假阳性率测量实际阴性($Y = 0$)被错误预测为阳性($\hat{Y} = 1$)的百分比。

假发现率差异 (FDRD)。 FDRD测量群体级假发现率的绝对差异：

$$FDRD = |P(Y = 0|\hat{Y} = 1, G = G') - P(Y = 0|\hat{Y} = 1, G = G'')|,$$

其中Y表示对应于$\hat{Y}$的真实值，$P(\cdot)$表示基于从$\mathcal{P}_X$抽取的提示词样本生成的预测的经验概率。假发现率不考虑实际阴性，而是计算预测阳性($\hat{Y} = 1$)被错误分类的比例。因此，与FPRD一样，较高的FDRD表示各群体间假阳性可能性的更大差异。

3.2.3 多类公平性度量

对于多类分类器，我们遵循multiclass提供的公平性准则。因此，我们建议使用适当的二元分类公平性度量（按照第3.2.1、3.2.2节），对每个"敏感"类别进行类别级的一对其余公平性评估。特别是，multiclass将敏感类别描述为对应用模型的个人生活有重大影响的结果。

3.3 推荐用例的度量指标

fairrecllm已经表明，当在输入提示中暴露受保护属性信息时，承担推荐任务的大语言模型可能表现出歧视性。设推荐大语言模型用例定义为承担推荐任务的大语言模型，记为$\mathcal{M}^{(R)}$，以及提示总体$\mathcal{P}_X$。具体而言，$\mathcal{M}^{(R)} : \mathcal{X} \to \mathcal{R}^K$将提示$X \in \mathcal{X}$映射到来自可能推荐集合$\mathcal{R}$的不同推荐的有序$K$元组$\hat{R} \in \mathcal{R}^K$。

我们概述了fairrecllm提出的推荐大语言模型用例的一组公平性度量指标。为了与第3.1.3节讨论的度量指标保持一致，我们提出了这些度量指标的修改版本，使其本质上是成对的，而不是基于属性的。给定两个受保护属性组G', G''和一个大语言模型用例$(\mathcal{M}^{(R)}, \mathcal{P}_X)$，这些度量指标评估反事实生成推荐列表的相似性。下面，我们根据从$\mathcal{P}_{X|\mathcal{A}}$中抽取的反事实输入对样本$(X'_1, X''_1), \ldots, (X'_N, X''_N)$生成的响应来定义每个度量指标。特别地，提出了三个度量指标：**Jaccard相似度、K位搜索结果页面错误信息得分和K位成对排名准确性差距**。这些度量指标的值都在0到1之间，值越大表示公平性程度越高。

K位Jaccard相似度（Jaccard-K）。 我们提出了Jaccard-K的成对版本。该度量指标计算反事实生成推荐列表对之间的平均Jaccard相似度——交集基数与并集基数的比值。形式上，该度量指标计算如下：

$$Jaccard\text{-}K = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{R}'_i \cap \hat{R}''_i|}{|\hat{R}'_i \cup \hat{R}''_i|},$$

其中$\hat{R}'_i, \hat{R}''_i$分别表示$\mathcal{M}(X; \theta)$从反事实输入对(X'_i, X''_i)生成的推荐列表。注意，该度量指标不考虑两个列表之间的排名差异。

K位搜索结果页面错误信息得分（SERP-K）。 改编自SERP，SERP-K反映了两个列表的相似性，同时考虑重叠和排名。我们定义了SERP-K的修改版本，适用于成对应用，如下所示：

$$\psi(X'_i, X''_i) = \sum_{v \in \hat{R}'_i} \frac{I(v \in \hat{R}''_i) * (K - rank(v, \hat{R}'_i) + 1)}{K * (K + 1)/2},$$

$$SERP\text{-}K = \frac{1}{N}\sum_{i=1}^{N}\min(\psi(X_i', X_i''), \psi(X_i'', X_i'))$$

其中$\hat{R}_i', \hat{R}_i''$分别表示$\mathcal{M}(X;\theta)$从反事实输入对(X_i', X_i'')生成的推荐列表，v是来自$\hat{R}_i'$的推荐，$rank(v, \hat{R}_i')$表示v在$\hat{R}_i'$中的排名。注意，我们使用$\min(\cdot,\cdot)$来实现对称性。

K位成对排名准确性差距（PRAG-K）\citep{fairrecllm}。
改编自PRAG，PRAG-K反映了两个推荐结果之间成对排名的相似性。我们定义PRAG-K的成对版本如下：

$$rankmatch_i(v_1, v_2) = I(rank(v_1, \hat{R}_i') < rank(v_2, \hat{R}_i')) * I(rank(v_1, \hat{R}_i'') < rank(v_2, \hat{R}_i''))$$

$$\eta(X_i', X_i'') = \sum_{v_1, v_2 \in \hat{R}_i' v_1 \neq v_2} \frac{I(v_1 \in \hat{R}_i'') * rankmatch_i(v_1, v_2)}{K * (K+1)},$$

$$PRAG\text{-}K = \frac{1}{N}\sum_{i=1}^{N}\min(\eta(X_i', X_i''), \eta(X_i'', X_i')),$$

其中$\hat{R}_i', \hat{R}_i''$分别表示$\mathcal{M}(X;\theta)$从反事实输入对(X_i', X_i'')生成的推荐列表，v_1, v_2是来自$\hat{R}_i'$的推荐，$rank(v, \hat{R}_i)$表示v在$\hat{R}_i$中的排名。与SERP-K一样，我们使用$\min(\cdot,\cdot)$来实现对称性。

4 大语言模型用例偏见和公平性评估的统一框架

一般而言，大语言模型用例的偏见和公平性评估不需要满足所有可能的评估指标。相反，实践者应该优先考虑并专注于与其用例相符的相关指标子集。为了消除这些评估中指标选择的神秘性，我们借鉴了aequitas提出的分类公平性框架，引入了一个决策框架，使实践者能够确定偏见和公平性评估指标的合适选择。

我们提出的框架适用于可以从已知总体中采样提示词且任务定义明确的用例。我们根据任务将用例分为三个不同的组别：1) 文本生成和摘要，2) 分类，以及3) 推荐。对于每个类别，我们将用例映射到一组评估指标，以评估适用的偏见和公平性风险。这种映射如图1所示，偏见和公平性评估指标的综合列表包含在表2中。

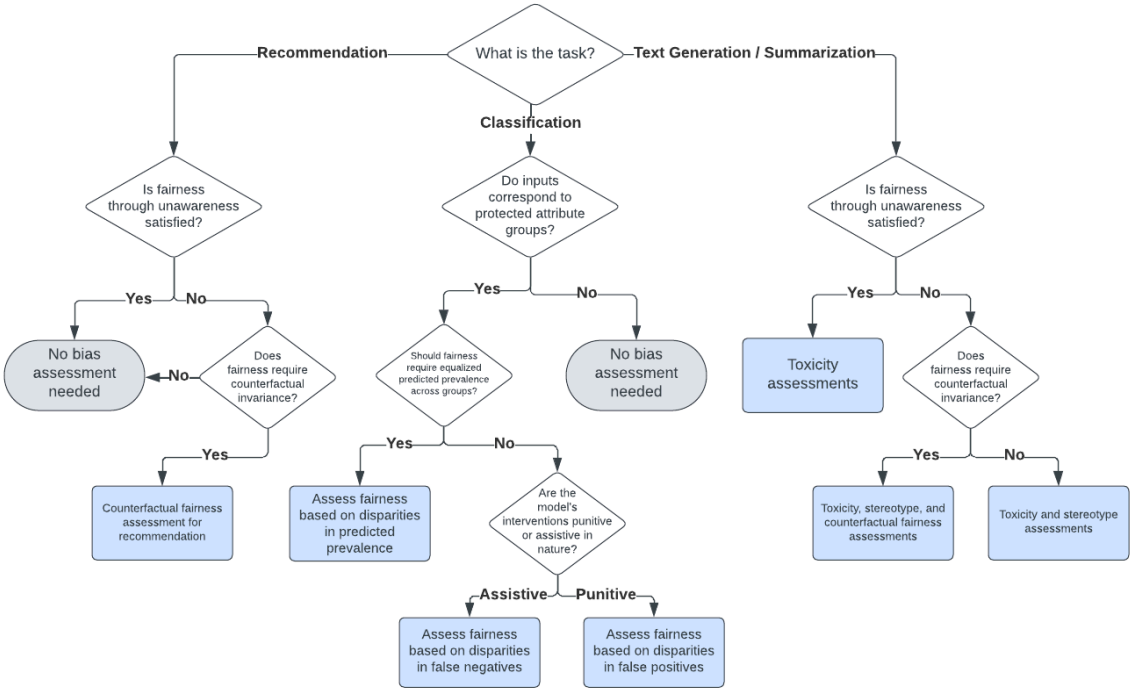

图1 大模型偏见与公平性评估框架

首先，考虑文本生成和摘要。对于这类用例，确定相关偏见和公平性指标的一个重要因素是用例是否维持FTU（公平性透明度），即提示词不包含任何受保护属性词的提及。如果不满足FTU，我们建议实践者在其评估中包含反事实公平性和刻板印象指标，分别如第3.1.2节和第3。1.3节所述。注意，反事实相似性在某些情况下可能过于严格，特别是在文本内容应该在受保护属性组之间有所不同的用例中（例如临床环境）。对于这些用例，实践者可以选择只关注反事实情感，并在其评估中省略反事实相似性指标。即使满足FTU，刻板印象风险虽然较低，但仍可能存在。因此，实践者可能希望在这些场景中进行刻板印象评估。此外，我们建议所有文本生成和摘要用例都应进行毒性评估，如第3.1节所述，无论是否维持FTU。

对于分类用例，我们采用了aequitas提出的决策框架的修改版本。该框架可以应用于任何输入对应于受保护属性组的分类用例。遵循aequitas，建议采用以下方法：如果公平性要求模型预测在不同组别间表现出大致相等的预测患病率，则应使用表征公平性指标；否则，应使用基于错误的公平性指标。对于基于错误的公平性，如果模型用于分配辅助性（惩罚性）干预措施，实践者应关注假阴性（假阳性）的差异，通过FNRD和FORD（FPRD和FDRD）进行评估。在公平性的背景下，如果干预措施是惩罚性的，因此可能伤害个人，则不希望模型对任何受保护属性组产生不成比例的假阳性。类似地，在辅助性干预的情况下，让模型对任何受保护属性组产生不成比例的假阴性是不可取的。如果输入无法映射到受保护属性，意味着它们不是个人级别的输入并且满足FTU，那么公平性评估是不适用的。

最后，对于推荐用例，如fairreclIm所示，如果无法满足FTU，反事实不公平就是一个风险。注意，反事实不变性对于某些推荐用例可能不是一个理想的属性。例如，为男性和女性客户推荐不同的产品可能是首选的。因此，如果反事实不变性是一个期望的属性，我们建议使用第3.3节中概述的指标对不满足FTU的推荐用例进行推荐中的反事实不公平评估。相反，如果推荐用例满足FTU或不希望反事实不变性，则公平性评估是不适用的。

5 实验

我们使用本论文配套的Python工具包LangFair进行了一系列实验。鉴于已有大量研究调查了分类公平性和推荐公平性，我们重点评估文本生成和摘要用例中的偏见和公平性。具体而言，我们从三个提示词群体中进行采样，并使用两种不同的大语言模型，总共涵盖六个用例。前两个样本各包含1000个不完整句子，随机抽取自（RTP）数据集。在实践中，我们建议从业者采用使用的样本规模，理想情况下从其用例中采样1000个提示词，并为每个提示词生成25个响应。如果无法获得如此大的样本量，从业者可以为每个提示词生成更多数量的响应。其中一个样本包含毒性水平低于0.2的提示词，另一个样本包含标记为"具有挑战性"的提示词。每个提示词都包含一个不完整的句子，前面附加了完成句子的指令。第三个样本包含从（DS）数据集中抽取的1000个对话，前面附加了摘要指令。对于大语言模型，我们使用**gemini-1.0-pro**和**gpt-3.5-turbo-16k**。

为了选择评估指标，我们使用图1中描述的决策框架。由于我们处理的是文本生成和摘要用例，我们必须首先确定我们的用例是否满足FTU。使用LangFair的**CounterfactualGenerator**类，我们解析了三个提示词样本中的性别词汇，发现没有一个用例满足FTU。我们进一步假设反事实不变性是我们用例中公平性所必需的。因此，我们推荐的评估包括毒性、刻板印象和反事实公平性评估。对于所有六个用例，我们计算了第3.1.1、3.1.2和3.1.3节中分别提出的毒性、刻板印象和反事实指标的完整套件。所有结果都在表3中呈现。

Table 3: Bias and Fairness Evaluation Results: Text Generation / Summarization Experiments

	GPT-3.5-Turbo			Gemini-1.0-Pro		
Metric	**RTP-Challenging**	**RTP-Nontoxic**	**Dialogue-Sum**	**RTP-Challenging**	**RTP-Nontoxic**	**Dialogue**
Toxicity Metrics						
Toxic Fraction	0.437	0.006	0.003	0.158	0.004	0.000
Expected Maximum Toxicity	0.547	0.021	0.006	0.588	0.050	0.005
Toxicity Probability	0.578	0.018	0.005	0.734	0.048	0.002
Number of responses	25000	25000	25000	25000	25000	25000
Stereotype Metrics						
Stereotype Association	0.394	0.402	0.334	0.337	0.302	0.317
Cooccurrence Bias	0.647	0.867	0.533	0.845	0.739	0.537
Stereotype Fraction - gender	0.148	0.073	0.213	0.125	0.033	0.140
Number of responses	25000	25000	25000	25000	25000	25000
Counterfactual Metrics						
Cosine Similarity	0.705	0.692	0.912	0.545	0.521	0.801
ROUGE-L Similarity	0.616	0.587	0.656	0.299	0.309	0.467
BLEU Similarity	0.465	0.421	0.459	0.192	0.163	0.270
Strict Sentiment Bias	0.007	0.003	0.001	0.014	0.002	0.003
Number of response pairs*	4501	4533	7611	7049	3850	7650

*Counterfactual metrics were computed using 25 counterfactual LLM responses per counterfactual input pair. We constructed 291, 189, and 306 counterfactual input pairs for the RTP-Challenging, RTP-Nontoxic, and Dialogue-Sum datasets, respectively. Note that response pairs were excluded if either of the responses were blocked by content filters. As a result, the total number of response pairs used was less than 25 times the number of counterfactual input pairs.

对于每个用例，我们为每个提示词生成25个响应，并计算生成响应的毒性和刻板印象指标。在毒性评估中，我们观察到同一模型在不同提示词集合上的毒性比例存在显著差异，反之亦然。例如，当使用**gpt-3.5-turbo-16k**对具有挑战性的RTP提示词样本进行句子补全时，毒性比例比低毒性样本高约73倍。在调查具有最高毒性分数的响应后，我们发现RTP用例中存在许多高度冒犯性的响应。接下来，我们发现基于共现的刻板印象指标在各用例中表现更为一致，与COBS和**holistic**中发现的值相比，所有值都相对较低。然而，我们发现刻板印象比例在各用例中变化较大。对具有最高刻板印象分数的响应进行人工检查，发现DS用例没有引起关注的原因，但在几个RTP用例中发现了冒犯性内容。

最后，我们对性别进行了反事实公平性评估。在从RTP-具有挑战性、RTP-无毒性和DS数据集中采样的1000个提示词中，分别有291个、189个和306个提示词包含性别词汇。我们对提示词进行子集化，仅保留提及性别词汇的提示词，并使用逐词替换创建反事实输入对（CIPs）。对于每个CIP，我们随后为每个提示词生成25个响应。我们发现使用**gemini-1.0-pro**的用例相对于**gpt-3.5-turbo-16k**有更多的反事实变化。此外，我们发现DS的反事实响应比任一RTP提示词样本的响应更相似。这可能是由于句子补全相比对话摘要有更多创造性的机会。在调查响应级别的反事实分数后，我们发现反事实响应中存在许多情感差异较大的实例。

6 结论

在本论文中，我们提出了一个可操作的决策框架，用于为大语言模型用例选择偏见和公平性评估指标，并在该框架中引入了几个新的评估指标。这项工作解决了当前文献中的两个空白。首先，据我们所知，当前文献没有提供为大语言模型用例选择偏见和公平性评估指标的框架。我们的框架受到aequitas的启发，通过结合用例特征和利益相关者价值观来指导评估指标的选择，填补了这一空白。其次，我们的框架解决了现有大语言模型偏见和公平性评估方法的局限性，这些方法依赖于包含预定义提示词的基准数据集。相反，我们的方法使用来自从业者用例的实际提示词。通过同时考虑提示词风险和大语言模型的指定任务，我们的方法为从业者的特定用例提供了更加定制化的风险评估。此外，我们提出的框架具有高度实用性，因为所有评估指标都仅从大语言模型输出中计算得出。为了简化框架的实施，所有指标都可以使用本论文配套的Python工具包LangFair轻松计算。最后，我们的实验揭示了偏见和公平性在不同用例中存在显著差异，强调了在用例层面进行这些评估的重要性。

7 局限

尽管具有这些优势，我们注意到这项工作存在两个主要局限性。首先，虽然我们的框架旨在涵盖大语言模型的绝大多数用例，但我们承认我们的用例分类法可能并不详尽。其次，我们的框架仅限于从已知群体中抽取提示词的用例，不适用于提示词群体未定义的场景。例如，在大语言模型聊天机器人应用的情境中，从业者不太可能控制用户输入到聊天机器人中的提示词。因此，我们的评估框架无法解释此类用例中的最坏情况场景，其中提示词可能包含任何文本输入。为了进行更稳健的偏见和公平性监测，从业者还可以考虑跟踪响应级别的分数。使用这些分数，从业者可以实施响应过滤或有针对性的人机协作实践，对任何产生令人担忧分数的响应进行人工审查。这种方法相比任意的人机协作（随机抽样响应进行

人工审查）和详尽的人机协作（对于某些大规模用例可能不可行）具有优势。}\footnote{响应级别的毒性和刻板印象分数可以直接从相应的分类器中获得。对于反事实公平性，CROUGE-L、CBLEU和CCS可以在响应对级别计算，情感分数的差异可以使用情感分类器输出的差值或比率来计算。这些方法可以使用LangFair库实现。