迈向更高效的金融语言模型偏见检测

Firas Hadj Kacem[1], **Ahmed Khanfir**[2,1] **和 Mike Papadakis**[1]
[1] SnT，卢森堡大学，卢森堡[2] RIADI，ENSI，马
努巴大学，突尼斯 firashadjkacem@ieee.org,
ahmed.khanfir@ensi-uma.tn 和
michail.papadakis@uni.lu

arXiv:2603.08267v1 [cs.AI] 9 Mar 2026

摘要

金融语言模型中的偏见是阻碍其在现实应用中得到采用的一大障碍。检测此类偏见颇具挑战，因为这需要识别出在改变与决策无关的属性（如人口统计属性）时预测结果会发生变化的输入。现有方法通常依赖于在大型语料库上进行穷尽式变异和成对预测分析，尽管效果良好，但计算成本高昂——尤其对于大型语言模型而言——在持续再训练和发布过程中甚至可能变得不切实际。为降低这一成本，我们对五种金融语言模型中的偏见开展了大规模研究，考察它们在受保护属性方面的偏见倾向是否相似，并探索跨模型引导的偏见检测，以便更早识别揭示偏见的输入。本研究使用约 1.7 万条真实金融新闻句子，通过变异构建了超过 12.5 万组原始—变异输入对。结果表明，所有模型在原子（0.58%–6.05%）和交叉性（0.75%–5.97%）设置下均表现出偏见。此外，我们观察到各模型的偏见揭示输入具有一致的模式，从而能够大幅复用检测结果并降低偏见检测成本。例如，借助从 DistilRoBERTa 输出中提取的属性进行引导，只需使用 20% 的输入对，即可发现 FinMA 多达 73% 的偏见行为。

1 引言

人工智能的快速发展，推动了人们对语言模型在金融新闻分析、风险评估和决策支持等任务中的兴趣日益增长（Xie 等人，2023）。基于 Transformer 的架构近期取得进展，使得专用金融语言模型得以开发，并在金融任务上的表现超越通用语言模型（Yang 等人，2023；Xie 等人，2023；Araci，2019）。然而，这些模型在现实系统中的应用仍然有限。语言模型存在偏见，是其推广应用的主要障碍之一（Guo 等人，2024），因为有偏预测可能导致影响个人或群体的歧视性结果。在金融领域，严格的监管要求进一步放大了此类风险。因此，金融语言模型的偏见检测已成为研究人员和从业者亟待关注的问题。现有研究表明，语言模型可能对性别、种族或身体特征等受保护属性表现出偏见（Kiritchenko 和 Mohammad，2018；Bolukbasi 等人，2016）。然而，以往的大多数研究都聚焦于通用语言模型，或使用小型数据集评估偏见（Asyrofi 等人，2021）。此外，偏见检测技术通常依赖穷尽式测试策略；这类策略成本高昂，难以扩展至多个模型、持续发布的版本或大型语料库。因此，目前缺乏实证证据来说明不同金融语言模型是否呈现相似的偏见模式，以及能否高效识别偏见揭示输入并在不同模型间重复使用。

本研究通过开展一项关于金融语言模型中偏见的大规模实证研究来弥合这些差距，即两个生成式大语言模型（FinMA（Yang 等人，2023）和 FinGPT（Yang 等人，2023））以及三个基于编码器的模型（FinBERT（Araci，2019年）以及经过微调的

DeBERTa-v3（Romero，2024a）和 DistilRoBERTa（Romero，2024b）），在一个由真实金融句子构成的数据集上，这些句子被系统性地变异，以揭示种族、性别和身体等受保护属性上的偏见。我们使用 HInter（Souani 等人，2025）对来自金融情感数据集（Fin-Sen）的 16,969 条金融句子进行单属性（原子）和双属性（交叉性）变异，生成 125,161 个原始样本-变异样本对。

为了（1）区分揭示偏见的输入与其他输入，以及（2）研究不同模型间偏差检测活动的结果能否复用，我们探索了模型间共享的偏见，并使用 Jensen-Shannon距离（JSD）和余弦相似度，量化每个模型中原始句子与变异后句子之间的预测偏移。

我们的结果表明，所研究的模型均存在原子偏见（0.58% 至 6.05%）和交叉偏见（0.75% 至 5.97%），且不同属性的偏见程度各异。这些较低的偏差比例表明，只有极少数输入能够揭示偏见，证实揭示偏见的输入十分罕见，大部分变异和推理开销都付诸东流。这进一步凸显了本研究的主要目标：降低语言模型偏见检测的成本。

更重要的是，我们表明，只需评估一小部分经过筛选的输入对，就能发现相当一部分有偏行为。事实上，本研究纳入的三个轻量级模型共享了超过 94% 的揭示偏见的输入，这为跨模型复用揭示偏见的输入（即在轻量级模型之间复用）最终能够节省成本提供了实证依据。尽管有所不同，生成模型中的揭示偏差的输入仍可根据预测概率偏移加以区分。例如，我们发现大型模型（即 FinMA）中的揭示偏差的输入往往会使原始句子与变异句子之间的预测概率相距更远。事实上，依据轻量模型（即 DistilRoBERTa）的预测分数计算詹森–香农距离，并按距离由大到小对测试输入进行优先排序，能够更早地在其他模型中发现偏见；仅使用 20% 的测试输入，即可暴露 FinMA 73% 的偏见，显著高于随机输入选择所达到的 $\approx 20\%$。这些发现令人鼓舞，表明在跨模型的偏差检测活动之间复用预测结果具有明显优势，尤其是利用轻量（低成本）模型的结果来指导大型（高成本）模型中的偏见检测。

本文的贡献有三方面。首先，我们利用大规模真实金融报表数据集，对五个金融语言模型中的偏见进行了全面的实证分析。其次，我们识别了各模型揭示偏差的输入中的共享模式，为偏见检测工作的可迁移性提供了洞见。第三，我们证明，基于模型导出特征对输入进行优先排序，可以显著加速偏见检测，并降低大型模型上的成本。我们公开了完整代码库以及包含更多细节的设置指令[1]，以实现可复现性并支持未来的同类研究。

2 相关工作

语言模型中的偏见检测 语言模型中的社会偏见通常源于训练数据中的失衡和既有偏见，导致模型反映并再现性别、种族或文化刻板印象（Bender 等人，2021年；Blodgett 等人，2020）。近期研究广泛探讨了此类偏见，并一致发现，这些偏见普遍存在于不同的架构和任务中。既有研究表明，生成式语言模型（Si 等人，2025）和分类语言模型（Kiritchenko 和 Mohammad，2018）都可能编码并放大与敏感属性相关的社会刻板印象。

语言模型中的此类偏见通常通过反事实（受控变异）测试进行研究（Kusner 等人，2017），该方法在输入的受控属性发生变化时评估软件输出的变化（Dwork 等人，2012年）。为此，已提出多个数据集，例如 CrowS-Pairs、BBQ 和 Winogender15，它们提供经过精心整理的、差异极小的句子对输入，且仅改变人口统计属性（Si 等人，2025）。为了不依赖现有数据集并增强偏见研究的真实性，近期工作提出了基于现有真实世界数据的自动化输入生成方法。例如，Asyrofi 等人（2021年）提出了 BiasFinder，这是一种蜕变测试方法，通过将句子变异为预期产生相似预测结果的输入对来生成测试用例，从而揭示更广泛的偏见。沿着

[1] https://github.com/Firas-HadjKacem/FinancialBias.git

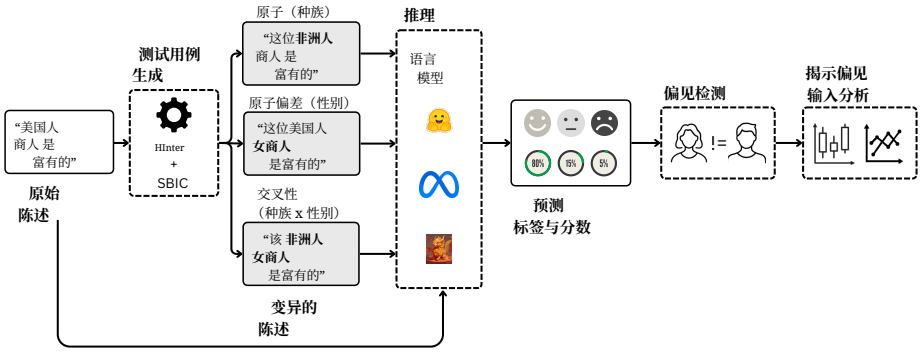


图1：实验工作流程

同一研究思路，Souani 等人（2025）提出了 HInter，该方法在每个输入中引入一个或两个人口统计属性变化，从而不仅能检测目标模型中的原子偏见，还能检测交叉性偏见。在我们的工作中，我们在金融情感数据集（FinSen）（Eaglewhl，2023）上运行 HInter，以从真实金融报表生成测试输入对。由于检测依赖于结果的变化，大多数针对语言模型的研究聚焦于分类任务而非生成任务。这是因为与复杂数据（即文本）相比，模型决策偏差在两个不同的预测标签之间更容易区分。同样，在我们的工作中，我们以分类任务——情感分析——作为研究金融语言模型中偏见的代理。

与以往研究不同，我们考察了揭示偏差的输入对的具体特征，尤其关注语言模型内部及不同语言模型之间的常见模式。本研究为语言模型偏见检测研究作出了贡献，提供了实证证据，表明根据模型导出特征对输入进行优先排序，可以显著加快检测过程，并降低大型模型的使用成本。

金融语言模型 过去几年中，金融语言模型不断涌现，既包括可公开使用的模型（如 FinBERT（Araci，2019年）、FinGPT（Yang 等人，2023）和 FinMA（Xie 等人，2023）），也包括私营公司开发的专有解决方案（如 BloombergGPT（Wu 等人，2023））。这些模型以骨干预训练通用模型为基础，使用大型金融语料进行训练或微调。
尽管这些模型在多项金融下游任务中表现出很高的效率（Siala 等人，2026），但目前仍缺乏对其公平性和偏见的评估，以及在此背景下对不同架构类型的全面评估与比较。

我们的研究是最早明确评估金融语言模型人口统计公平性的研究之一。鉴于此类模型在高风险金融应用中的使用日益广泛，这些偏见可能会影响投资、贷款、招聘等领域的关键决策。

3 方法

如图1所示，我们的实验流程包括四个主要阶段：（1）通过输入变异生成偏见测试用例；（2）情感预测；（3）偏见检测；以及（4）跨模型的偏见揭示输入分析。

3.1 偏见测试用例生成

此步骤旨在生成一个用于偏见检测的测试用例数据集。为此，我们使用 HInter（Souani 等人，2025），这是一种黑盒蜕变模糊测试方法，通过改变输入句子的某些属性（例如性别），同时保持其含义和语法结构不变，对句子进行变异，从而生成多个与原始句子相似但又有所不同的版本。这些替换是借助从社会偏见推理语料库（SBIC）（Autonlp，2021年）中提取的偏见词典进行的，该词典包含人口统计类别相关的词语及其替换项。HInter 聚焦于通常与偏见相关的三个属性轴：性别、种族 （对与族裔、原籍国等相关的术语进行归类）以及身体 （对与外貌相关的术语进行归类）。例如，偏见词典收录了阳性↔阴性代

词（例如，“他” ↔ “她”）、性别化名词（“商人” ↔ “女商人”）、带有性别刻板印象的职业（“首席执行官” ↔ “助理”）、国籍或族裔描述词（“美国人” ↔ “中国人”）、种族术语（“黑人” ↔“白人”）以及一些与身体相关的描述词（“男性” ↔ “自闭症男性”）。利用该词典，我们为每个输入生成了变异版本，采用两种主要的变异类型：

- **原子变异：** 每次在输入句子中更改一个敏感属性。例如，对句子“首席执行官表示他对……充满信心。”应用性别原子变异，会通过将**他**改为**她**，生成如下变异句子“首席执行官表示她对……充满信心。”。
- **交叉变异：** 同时更改输入句子中的两个敏感属性。例如，对句子**“他**带领**美国人** 公司走向成功。”应用性别与种族 交叉变异，会通过分别将**他** 和**美国人** 改为**她** 和**亚洲人**，生成**“她** 带领**亚洲** 公司走向成功。”这些高阶变异旨在揭示仅当两种属性变化同时出现时才可能显现的交叉性偏见。

3.2 模型推理与偏见检测

在此步骤中，我们要求受测语言模型预测所构建数据集（原始样本-变异样本对）中每条陈述的情感。我们收集每一项推理结果——即每个类别的标签和预测概率——并将其与相应的输入和预测模型进行映射，以用于偏见检测和研究分析。情感分析推理的运行方式因目标模型的运作机制而异：

轻量级（分类器）模型，即 FinBERT、DeBERTa-v3（微调）和 DistilRoBERTa（微调），专为情感分析任务设计；它们直接将原始句子作为输入，并通过分类头输出预测的情感，以及三个可能标签（负面、正面和中性）各自的分数（经 Softmax 归一化的 logit 值）。

生成式语言模型，即 FinGPT 和 FinMA，是基于指令的模型，需要特定的提示来处理任何任务，例如情感预测。我们遵循 FinMA 作者提出的方法（Xie 等人，2023）来执行分类任务，采用零样本提示来生成情感标签。具体而言，我们将每个输入句子包装在一个询问情感的提示中，如下所示：

> 分析从一篇财经新闻文章中摘录的这段陈述的情感。请将答案填写为负面、正面或中性。文本：“<句子>”答案：

除了提示词的功能设计外，还需要额外步骤从生成的输出文本中提取标签和分数。事实上，模型并不总是遵循给定的指令，例如，它可能不会在提示词之后直接生成情感，而是在标签之前或中间添加空格或其他容易造成误导的令牌。为了解决这些问题，我们限制词汇表，仅允许那些能够构成有效标签变体开头的令牌 ID（例如，“positive”、“Positive”，以及可能带有空格或标点的形式等）。这确保模型确实以情感标签形式开始回答。同时，这也使我们能够在同一次前向传播中，从对应生成令牌的逻辑值中提取分数和标签。

具体而言，我们解析生成文本，定位与每个情感标签对应的第一个生成令牌，忽略序列结束标记或特殊标记，并将不同的标签变体作为跨度纳入考虑（“Neg”+“ative”）。接下来，我们在找到的标签位置准确提取模型逻辑值。最后，对提取出的、与各情感标签对应的逻辑值进行 Softmax 运算，从而为每种情感分配一个分数（概率）。

每当一个标签出现多种分词形式时，我们会在应用 Softmax 之前使用对数和指数将它们合并。这确保了情感标签的各种形式（例如“negative”和“ Negative”）都能得到公平汇总。根据 Goldberg（2017）的说法，“对数和指数”是在对数空间中累加概率的数值可靠方法。给定标签所有可能分词形式的对数分数 $\{\ell_i\}$，我们计算 $\log\sum e^{\ell_i}$，它等于这些预测结果$_i$之和的对数

分数（概率）。这样我们就能计算所选标签的概率，并在同一解码步骤中将其与其他类别逻辑值进行比较。通过将逻辑值提取与模型固有的生成行为相结合，我们得到一种混合逻辑值提取方法，既避免了词元错位问题，也确保我们能在决策点准确衡量模型对其所选标签的真实置信度。通过确保预测结果的一致性，我们也保证了输出可与分类器模型的输出进行比较。我们采用贪心解码，以确保实验具有确定性和可复现性。

3.3 偏见检测

由于每个原始样本—变异样本对之间唯一的差异是人口统计特征变化，因此我们预期真正无偏的模型对两者会给出相同的情感判断。因此，我们将任何标签变化（例如，情感从“正面”变为“负面”）视为有偏预测，并据此将该输入对视为揭示偏见的输入。通过比较上一步收集的每个原始样本—变异样本对的预测结果，我们为每个模型、变异类型和人口统计属性获得一组揭示偏见的输入。因此，每个模型的偏见检测率可计算为揭示偏见的输入数量占输入总数的比例。

3.4 揭示偏见的输入分析

在此步骤中，我们考察不同模型的揭示偏见的输入中常见的趋势和属性。首先，我们研究所测模型中揭示偏见的输入之间的重叠情况，以及彼此不相交的集合。这一点很重要，因为较大的重叠意味着可以在不同的偏见检测活动之间直接重复使用输入，从而节省推理开销以及寻找偏见输入的计算成本。

此外，我们还将研究扩展至考察变异所导致的整体决策变化，而不只关注标签翻转。为捕捉决策偏移，我们计算原始样本—变异样本对的预测中，各种情感的分数向量之间的差异。根据我们此前对句子概率分布的定义，我们发现，只要输入发生改变，$P(S_{\text{mutated}}) \neq P(S_{\text{orig}})$，即使预测标签保持不变。这些变化有大有小，轻微的变化甚至可能大到足以导致标签翻转。我们的直觉是，这类变化能够揭示输出中未必体现的隐性偏见形式。更重要的是，即使这些变化未揭示偏见（未发生预测标签翻转），它们也可能捕捉到能够区分其他模型中偏见揭示输入的特征，从而实现更具成本效益、由跨模型引导的偏见检测。
To 为衡量输出向量中预示最终偏见的变化，我们使用两种距离度量 :

詹森–香农距离（JSD）： 是詹森–香农散度的平方根。我们选择它，是因为它具有对称性且取值有限（范围为 0 到 1），而库尔贝克–莱布勒散度则不具备这些性质；同时，它保留了两种分布差异程度的信息论含义（Fuglede & Topsøe，2004年）。如果两个向量完全相同，则一对输入的 $JSD = 0$ 为 $P(S_{\text{mutated}}) = P(S_{\text{orig}})$；当二者差异最大时，该值趋近于 1。在偏见检测中，JSD 可作为偏见引发的变化的分级指标：较高的 JSD 表明人口统计特征变异导致了较大的概率偏移。

余弦相似度： 该指标已被广泛用于衡量模型所捕获上下文之间的相似度（Salton 等人，1975）。我们用它来估算模型预测偏移的幅度，方法是衡量变异后的预测分数向量与原始预测分数向量之间的相似度。该指标的取值范围为 0 到 1，其中 1 表示向量完全相同，0 表示向量完全不同。

3.5 跨模型引导偏差检测

为探究模型间观察到的共性与趋势在实践中的实用价值，我们提出了一种跨模型引导偏差检测方法。不同于传统的随机方法，我们根据其他模型的预测结果，优先处理输入测试对集合。为了评估该方法的有效性和成本效益，我们将其与随机方法进行比较，并按输入对推理的工作量逐步统计偏差检测率。
尽管这种方法有助于任何偏见检测活动，我们尤其关注利用轻量、低成本的模型来指导大型高成本模型进行偏见检测的优势。

表1：模型在 FinSen 上的偏见率。按属性划分的测试配对总数：身体 = 22,972；性别 = 105,962；种族 = 121,490；总计 = 125,161。

模型	原子（身体）	交互级（身体）	原子（性别）	交互级（性别）	原子（种族）	交互级（种族）	总计（原子）	总计（交互级）	总隐藏层（交互级）
FinMA	9.23%	7.48%	2.77%	2.25%	3.25%	3.29%	3.99%	3.23%	4.05%
FinGPT	5.39%	2.77%	6.10%	6.55%	6.13%	6.07%	6.05%	5.97%	31.29%
FinBERT	1.89%	1.88%	0.69%	0.88%	0.25%	0.41%	0.58%	0.75%	30.34%
DeBERTa-v3	1.69%	1.67%	0.70%	0.89%	0.30%	0.46%	0.60%	0.75%	29.95%
DistilRoBERTa	1.69%	1.67%	0.70%	0.89%	0.30%	0.46%	0.60%	0.75%	29.95%

我们将分类器（FinBERT、DeBERTa-v3（微调后）、DistilRoBERTa（微调后））称为用作参考的轻量型较小模型，将生成式大语言模型（FinMA、FinGPT）称为用作目标的较大模型。

针对轻量级—大型模型的六种组合，我们采用以下每种输入优先级策略运行（100 次）偏差检测活动：

- 以随机输入排序作为基线，- 按参考模型预测结果中的 JSD 从高到低排列输入，- 按参考模型预测结果中的余弦相似度从低到高排列输入，- 优先处理参考模型标记为存在偏见的输入

4 实验

4.1 实验设置

数据集：为开展本研究，我们使用金融情感数据集（Fin-Sen）（Eaglewhl，2023）中的美国子集，其中包含 16,969 个金融句子。FinSen 提供了涵盖 15 年金融相关文本的真实样本，这些文本摘自财经新闻文章、标题和财务报表。句子长度不一，短则 42 个词，长则 186 个词。我们按照第 3.1 节所述对该数据集进行变异，构建了 125,161 个原始—变异偏见测试用例。我们保留每对样本中被更改属性的相关信息，从而按照变异类型对扩充后的语料进行划分。

金融语言模型：我们在五种金融语言模型上开展研究，包括 2 个多用途生成式大语言模型和 3 个基于编码器的分类器；这些模型均通过金融语料微调获得：

- FinMA：一个基于 LLaMA2、拥有 70 亿参数的生成式金融大语言模型，来自 PIXIU 研究（Xie 等人，2023）。该模型使用金融数据进行了广泛训练，并专门针对情感分类进行了微调，以便在收到提示时生成文本情感标签。

- FinGPT：同样是一款基于 LLaMA2、拥有 7B 参数的生成式大语言模型，由 Yang 等人（2023）开源。通过额外的金融语料库预训练和基于 LoRA 的轻量级微调，该模型适用于金融情感生成。

- FinBERT：一种经典的、基于 BERT（1.1 亿参数）的金融情感模型分类器。该模型先在金融文本上进一步预训练，再针对金融情感分类进行微调（使用 Financial PhraseBank 和 FiQA 等数据集）（Araci，2019年）。

- DeBERTa-v3（微调版，（Romero，2024a））：一种基于 DeBERTa-v3 基础模型（He 等人，2021年）的分类器，针对金融新闻情感进行了微调（≈44M 参数）。
- DistilRoBERTa（微调版，（Romero，2024b））：一种蒸馏版 RoBERTa-base 模型分类器（Liu 等，2019年），在金融新闻情感数据上进行了微调。它有 ≈82M p个参数。
模型在 Hugging Face 上的 ID 链接列于附录 A.1 节的表 4 中。

4.2 结果

4.2.1 金融语言模型中的偏见检测有效性

所研究的五种模型在不同变异（原子和交叉性）以及三种受控属性下，表现出程度各异的偏见行为。表1按模型和受控属性报告了数据集中检测到偏见的配对所占比例。或许

令人惊讶的是，我们发现，简单的轻量级模型，即 FinBERT、DeBERTa-v3 和 Distil-RoBERTa，相较于较大的模型，即 FinMA 和 FinGPT，整体原子偏差比例（$\approx 0.6\%$）和交叉性偏差比例（$\approx 0.75\%$）更低；后两者的原子偏见比例分别为 3.99% 和 6.05%，交叉性偏见比例分别为 3.23% 和 5.97%。这表明，与较大的模型相比，采用小型分类语言模型可作为一种偏见缓解措施。对交叉性偏见的分析显示，其中很大一部分是隐藏的（无法通过单属性变异发现）；FinBERT、DeBERTa-v3 和 Distil-RoBERT 的这一比例为 $\approx 30\%$，FinGPT 为 $\approx 31\%$，FinMA 为 $\approx 4\%$，这凸显了通过使用高阶变异来超越单属性偏见检测的重要性。

4.2.2 模型间的共享偏见

图2中的维恩图展示了所研究的五种模型中，揭示偏见的输入对的重叠部分与不相交集合。尽管存在共同的偏见触发因素，但我们注意到，不同模型之间不存在一组普遍适用的共同偏见揭示输入。不过，我们发现，架构系列相似的模型之间存在明显重叠，尤其是所研究的三种分类器，其偏见有超过94%由相同的输入揭示；而 DeBERTa-v3 与 DistilRoBERTa 的集合则完全重叠。

这些结果令人鼓舞，表明轻量级模型之间可以复用输入，因为无需开展完整的偏见检测活动，就能在一个模型中找出揭示偏见的输入。然而，尽管生成模型的训练方式和容量相似，它们共享的有偏输入仍然很少（9）。再加上它们与轻量级模型之间的重叠微乎其微，这些发现促使我们进一步研究不同模型间揭示偏见的输入所共有的属性。

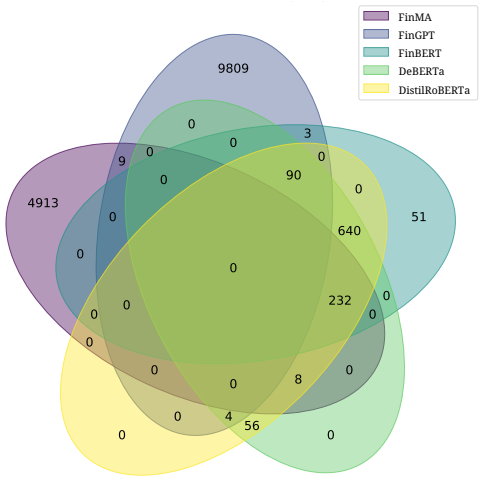


图2：各模型之间的偏见重叠。

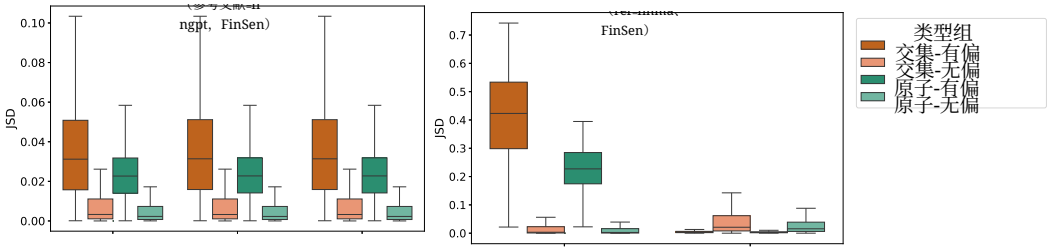
4.2.3 跨模型偏见揭示输入对分析

为了通过在推理前区分揭示偏见的输入对来降低模型偏见检测成本，我们进一步研究了变异在其他模型中造成的决策偏移幅度。具体而言，对于模型*M1*，我们使用第二个模型*M2*获得的预测分数（概率）来表示原始句子及其变异后的对应句子，并计算两者之间的余弦距离和JSD距离。例如，图3a展示了FinMA*(M1)* 的揭示偏见输入对与未揭示偏见输入对之间的差异；该差异是根据各轻量级模型*(M2)*的预测结果计算出的JSD来衡量的。箱线图显示，与未揭示偏见的输入对相比，揭示偏见的输入对往往对应着差异更大的预测结果；交叉性测试用例和原子测试用例的JSD中位数分别为 ≈ 0.031 和 ≈ 0.023，而未揭示偏见的输入对分别为 ≈ 0.003 和 ≈ 0.002 。

We 针对另一个模型时，观察到相似但幅度不同的区分趋势 s
or c计算余弦相似度，而非 JSD。
Wilcoxon p值非常低（低于10^{-32}），且$\hat{A}_{12}$ 值与0.5相差甚远（差值超过0.35，如表2所示），这表明观察到的差异具有显著性，并非偶然发生，从而证实在大多数情况下，可以根据另一个模型的预测偏移有效检测出揭示偏见的输入。

表2：Vargha和Delaney $\hat{A}_{12}$ 效应量，用于衡量 FinMA 揭示偏见输入与未揭示偏见输入之间的预测偏移差异（基于图3中的数值计算）。

$\hat{A}_{12}$	FinBERT	DeBERTa	DistilRoBERTa	FinMA	FinGPT
原子	0,88	0,88	0,88	0,99	0,16
交叉性	0,85	0,85	0.85	0,99	0,18



（a）轻量级模型的预测偏移。　　（b）大型模型的预测偏移。

图 3：基于轻量级模型和大型模型得到的预测概率，比较 FinMA 中揭示偏见与非揭示偏见的输入；比较依据是原始陈述与变异陈述之间的 JSD距离。

表3：比较引导式输入选择与传统（随机）选择在基于 DistilRoBERTa 预测结果检测大模型偏见时的效率和成本效率。我们分别用 BD%、ˆ ˆ∆、p值和 A_{12}, 表示偏见检测率、BD% 相对于随机选择的差值、Wilcoxon p值和 A_{12} 值。

模型 / 工作量	20% BD%	Δ	p值	$\hat{A}_{12}$	40% BD%	Δ	p值	$\hat{A}_{12}$	60% BD%	Δ	p值	$\hat{A}_{12}$	80% BD%	Δ	p值	$\hat{A}_{12}$
优先级排序将 DistilRoBERTa 的偏见揭示输入排在首位。																
FinMA	23,12	**+3,12**	$\mathcal{O}(10^{-18})$	1	42,35	**+2,35**	$\mathcal{O}(10^{-18})$	0,996	61,53	**+1,53**	$\mathcal{O}(10^{-17})$	0,936	80,79	**+0,79**	$\mathcal{O}(10^{-13})$	0,835
FinGPT	20,07	**+0,07**	$\mathcal{O}(10^{-2})$	0,58	40,02	**+0,02**	$\mathcal{O}(10^{-1})$	0,52	60,03	**+0,03**	$\mathcal{O}(10^{-1})$	0,52	80,04	**+0, 04**	$\mathcal{O}$ (10^{-1})	0,52
按照 DistilRoBERTa 预测偏移的 JSD 值从高到低对输入进行优先级排序。																
FinMA	73,01	**+53,01**	$\mathcal{O}(10^{-18})$	1	89,64	**+49,64**	$\mathcal{O}(10^{-18})$	1	95,49	**+35,49**	$\mathcal{O}(10^{-18})$	1	97,5	**+17,5**	$\mathcal{O}(10^{-18})$	1
FinGPT	7,61	**-12,39**	1	0	34,14	**-5,86**	1	0	61,52	**+1,52**	$\mathcal{O}(10^{-18})$	1	82,76	**+2, 76**	$\mathcal{O}(10^{-18})$	1

4.2.4 跨模型引导偏见检测的成本效益

为评估输入对之间观察到的预测偏移重叠和差异（第4.2.2节和第4.2.3节的研究结果）的实际用途，我们比较了引导式选择与传统（随机）选择在效率和成本方面的表现。

由于轻量级模型共享大部分揭示偏差的输入（第4.2.2节的研究结果），用另一个模型引导其中任何一个，都能以极低的成本实现近乎全面的偏见检测（使用 $<$ 20%的输入即可达到95%+ ）。

表3报告了以轻量级模型DistilRoBERTa的预测为引导时，大型模型通过输入对的预测所检测出的偏差比率与投入精力的关系，并将其与随机方法进行比较。

FinMA 和 FinGPT 的正面 ∆ 值均显示，在低成本和高成本情况下，优先选择 Distil-RoBERTa 的偏见揭示输入，相较于随机选择具有优势。这些差异相对较小，这可以由相关模型的偏见揭示输入集之间重叠较少来解释，正如第 4.2.2 节所示。优先选择基于 DistilRoBERTa 预测概率计算出的 JSD 较大的输入对时，相较于随机排序，FinGPT 的改进仍微乎其微，且仅在投入较高时才可观察到。然而，将这一策略应用于 FinMA 时，则显示出明显优势：只需投入 20% 和 40% 的精力，便可分别揭示 FinMA 73,01% 和 89,64% 的偏见。所进行的统计检验和效应量测量证实，所提出的输入优先级策略显著优于随机策略，其 p值的数量级为 10^{-18}，ˆA_{12}值为 ≈ 1。

5 结论

本文对金融语言模型中的人口统计学偏见开展了大规模研究，表明大型模型与轻量级模型在原子设定和交叉性设定下均表现出偏见。这项工作是对金融语言模型偏见检测的首次全面调查。我们进一步证明，通过利用跨模型揭示偏差的输入中的共享模式，可以显著加速偏见检测，从而仅使用一小部分测试输入即可揭示很大一部分偏见行为。由此，我们提供了首个实证证据，支持将跨模型引导的偏见检测作为一种有前景的方向，用以降低偏见审计以及任何下游偏见相关任务（即缓解）的成本，具体方式是向从业者

以显著更低的成本提供关键的揭示偏差的输入。尽管这些结果是在金融领域获得的，但我们的发现可能推广至其他语言模型和应用领域，我们将此留作未来工作。

致谢

本研究全部或部分由卢森堡国家研究基金（FNR）资助，资助编号为 NCER22/IS/16570468/NCER-FT。

参考文献

Dogu Araci。FinBERT：使用预训练语言模型进行金融情感分析。*arXiv* 预印本，2019年。doi：10.48550/arXiv.1908.10063。URL https://arxiv.org/abs/ 1908.10063。

Muhammad Hilmi Asyrofi、Imam Nur Bani Yusuf、Hong Jin Kang、Ferdian Thung、Zhou Yang 和 David Lo。Biasfinder：通过蜕变测试生成揭示情感分析系统中的偏见。*arXiv* 预印本，2021年。doi：10.48550/arXiv.2102.01859。URL https://arxiv.or g/abs/2102.01859。

Auto自然语言处理。社会偏见推理语料库（SBIC）。https://autonlp.ai/datasets/socia l -bias-inference-corpus-(SBIC)，2021年。

Emily M. Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。《随机鹦鹉的危险：语言模型是否可能过于庞大？》载于 *2021*年 *ACM* 公平性、问责制与透明度会议（*FAccT '21*）论文集，第 610–623 页，美国纽约州纽约市，2021年。美国计算机协会。doi：10.1145/3442188.3445922。URL：https://doi.org/10.1145/3442188.3445922。

Su Lin Blodgett、Solon Barocas、Hal Daum´e III 和 Hanna Wallach。《语言（技术）即权力：NLP 中"偏见"的批判性调查》。*arXiv* 预印本，2020 年。doi：10.48550/arXiv.2005.14050。URL：https://arxiv.org/abs/2005.14050。

Tolga Bolukbasi、Kai-Wei Chang、James Y. Zou、Venkatesh Saligrama 和 Adam Kalai。《男性之于计算机程序员，犹如女性之于家庭主妇？消除词嵌入中的偏见》。*arXiv* 预印本，2016年。doi：10.48550/arXiv.1607.06520。网址：https://arxiv.org/abs/1607.06520。

Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。《认知公平性》。载于第 *3* 届理论计算机科学创新会议论文集，ITCS '12，第 214–226 页，纽约，美国，2012年。计算机协会。ISBN 9781450311151。doi：10.1145/2090236.2090255。网址：https://doi.org/10.1145/2090236.2090255。

Eaglewhl。FinSen：金融情感数据集。Kaggle 数据集，2023。网址 https://www.ka ggle.com/datasets/eaglewhl/finsen-financial-sentiment-dataset。访问日期：2026年1月11日。

Bent Fuglede 和 Flemming Topsøe。詹森–香农散度与希尔伯特空间嵌入。*IEEE*信息论国际研讨会，第 31 页，2004年。

Yoav Goldberg。自然语言处理的神经网络方法。Morgan & Claypool，2017。doi：10.2200/S00762ED1V01Y201703HLT037。

Yufei Guo、Muzhe Guo、Juntao Su、Zhou Yang、Mengqiu Zhu、Hongfei Li、Mengyang Qiu 和 ShuoShuo Liu。大型语言模型中的偏见：起源、评估与缓解。*arXiv* 预印本，2024年。doi：10.48550/arXiv.2411.10915。网址 https://arxiv.org/abs/2411.10915。

Pengcheng He、Xiaodong Liu、Jianfeng Gao 和 Weizhu Chen。Deberta：采用解耦注意力的增强解码 BERT。载于国际学习表征会议论文集（*ICLR*），2021年。

Svetlana Kiritchenko 和 Saif M. Mohammad。考察两百个情感分析系统中的性别与种族偏见。*arXiv* 预印本，2018年。doi：10.48550/arXiv.1805.04508。网址：https://arxiv.org/abs/1805.04508。

Matt Kusner、Joshua Loftus、Chris Russell 和 Ricardo Silva。反事实公平性。载于第*31*届神经信息处理系统国际会议论文集，NIPS'17，第 4069–4079 页，美国纽约州雷德胡克，2017。Curran Associates Inc. ISBN 9781510860964。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv*，abs/1907.11692，2019年。

Manuel Romero。DeBERTa-v3-ft-financial-news-sentiment-analysis。Hugging Face 模型，2024a。网址 https://huggingface.co/mrm8488/deberta-v3-ft-financial-new s-sentiment-analysis。修订版本 7430ace。

Manuel Romero. DistilRoBERTa-finetuned-financial-news-sentiment-analysis. Hugging Face 模型，2024b。网址：https://huggingface.co/mrm8488/distilroberta-f inetuned-financial-news-sentiment-analysis。访问日期：2026-01-?

Gerard Salton、Anita Wong 和 Chung-Shu Yang。《用于自动索引的向量空间模型》。《*ACM* 通讯》，18(11):613–620，1975。doi: 10.1145/361219.361220。

Shijing Si、Xiaoming Jiang、Qinliang Su、Zhou Yang、Mengqiu Zhu、Hongfei Li、Mengyang Qiu、Shuo Shuo Liu 和 Lawrence Carin。利用贝叶斯假设检验检测大语言模型的隐性偏见。《科学报告》，15(1):12415，2025。doi: 10.1038/s41598-025-9 5825-x。网址：https://www.nature.com/articles/s41598-025-95825-x。

Walid Siala、Ahmed Khanfir 和 Mike Papadakis。大语言模型新闻情感分析对股价走势预测的影响。*arXiv* 预印本 *arXiv:2602.00086*，2026。

Badr Souani、Ezekiel O. Soremekun、Mike Papadakis、Setsuko Yokoyama、Sudipta Chattopadhyay 和 Yves Le Traon。HInter：揭示大语言模型中隐藏的交叉性偏见。*arXiv* 预印本，2025。doi: 10.48550/arXiv.2503.11962。URL https://arxiv.org/abs/2503 .11962。

Shijie Wu、Ozan Irsoy、Steven Lu、Vadim Dabravolski、Mark Dredze、Sebastian Gehrmann、Prab- hanjan Kambadur、David Rosenberg 和 Gideon Mann。Bloomberggpt：面向金融的大语言模型。*arXiv* 预印本 *arXiv:2303.17564*，2023。

Qianqian Xie、Weiguang Han、Xiao Zhang、Yanzhao Lai、Min Peng、Alejandro Lopez-Lira 和 Jimin Huang。PIXIU：面向金融的大语言模型、指令数据和评估基准。*arXiv* 预印本，2023。网址：https://arxiv.org/abs/2306.05443。

Hongyang Yang、Xiao-Yang Liu和Christina Dan Wang。FinGPT：开源金融大语言模型。*arXiv*预印本，2023。doi: 10.48550/arXiv.2306.06031。URL https: //arxiv.org/abs/2306.06031。

A 附录

A.1 金融语言模型

表 4：HuggingFace 上所研究金融语言模型的链接

模型	HuggingFace ID
FinMA	https://huggingface.co/ChanceFocus/finma-7b-full
FinGPT	https://huggingface.co/oliverwang15/FinGPT_v32_Llama2_情感_指令_LoRA_FT
FinBERT	https://huggingface.co/ProsusAI/finbert
DeBERTa-v3	https://huggingface.co/mrm8488/deberta-v3-ft-financial-新闻情感分析
DistilRoBERTa	https://huggingface.co/mrm8488/distilroberta-finetuned-金融新闻情感分析

A.2 实现细节

我们按照前文第3节所述，将实验工作流实现为Python脚本，并依赖成熟的公开可用库，即：

- transformers 用于模型加载、分词和推理，- bitsandbytes 用于模型量化，- PyTorch 用于推理和 GPU 上的执行操作，- LogitsProcessor 用于词汇表屏蔽和受约束生成
- NumPy 和 pandas 用于若干向量运算。

在整个代码库中还使用了其他库和工具。所有这些对于加载金融模型、生成预测、获取偏见结果，以及通过 JSD 和余弦相似度量化偏见都十分重要。

我们为所有随机操作（例如文本生成）固定了随机种子，以确保结果的可复现性。

我们方法的实现及复现包可在 https://gi thub.com/biasfinllms/iclr-finai-submission 获取

A.3 计算基础设施与资源

测试用例生成：我们在一台标准工作站（Intel Core i7-1165G7 @ 2.8GHz CPU，16GB 内存）上完成了变异生成与数据集预处理，未使用 GPU 加速，仅耗时数小时即完成。

模型推理与分析：为在 125k+测试用例上运行情感模型并执行偏差分析，我们使用了支持 GPU 的计算环境。模型输出的初步原型验证与抽查在单块 NVIDIA T4 GPU（通过 Google Colab 免费层级）上完成。然而，对所有变异对进行全批次推理的计算量更大。为开展这些实验，我们使用了一台配备 3 ×NVIDIA RTX A6000 GPU（每块具有 49 GB 显存）及 504 GB 系统内存的高性能服务器。