在大型语言模型中减少对酷儿代表性的偏见：一种协作代理方法

Tianyi Huang[1]
App-In Club
tonyhrule666@gmail.com

Arya Somasundaram[2]
App-In Club
arysom992@gmail.com

Abstract

大规模语言模型（LLM）经常在代词使用中延续偏见，导致酷儿个体的错误表达或排除。本文解决的是 LLM 输出中代词使用偏见的具体问题，特别是在需要使用包容性语言以准确代表所有身份时，不恰当地使用传统的性别代词（“he”，“she”）。我们引入了一个合作代理管道，通过分析和优化代词使用以促进包容性，来减轻这些偏见。我们的多代理框架包括用于偏见检测和校正的专门代理。使用 Tango 数据集进行的实验评估——这是一个专注于性别代词使用的基准——表明我们的方法显著改善了包容性代词分类，与 GPT-4o 相比，在正确反对不恰当的传统性别代词（$\chi^2 = 38.57, p < 0.0001$）上提高了 32.6 个百分点。这些结果强调了基于代理驱动的框架在增强 AI 生成内容的公平性和包容性方面的潜力，展示了其在减少偏见和促进社会责任 AI 方面的效力。

1 介绍

大型语言模型（LLMs）的进步显著推动了自然语言处理（NLP），使机器能够生成类似人类的文本并以显著的熟练度执行复杂的语言任务 [3, 9] 。然而，LLMs 通常会继承并放大其训练数据中存在的社会偏见，导致边缘群体受到忽视 [4, 2] 。在这些群体中，酷儿群体在人工智能表示中面临独特的挑战，特别是在代词使用和性别认同方面 [16, 6] 。

现有的偏差缓解技术，如数据增强 [17] 、去偏算法和公平意识的机器学习模型 [8, 7] ，主要集中在二元性别和种族等更广泛的人口类别上。这些方法通常无法解决酷儿身份的多样性问题，这涉及性别表达的流动性和多样性以及所用语言的演变 [10, 16, 1] 。诸如“they”、“xe”、“ey”和“fae”这样的代词由非二元性别和跨性别个人使用，但经常在 LLMs 中得不到充分代表或被误解 [5] 。性别误称和排斥性语言会导致对酷儿个体的歧视的延续 [11, 15] 。因此，解决 LLMs 中的酷儿偏见需要采用专门的方法来考虑性别身份和代词使用的复杂性。

在本文中，我们解决了大语言模型（LLM）输出中的偏见代词使用的特定问题，尤其是在需要包容性语言时对传统性别代词的不恰当使用。我们引入了一个协作代理管道，旨在减少代词使用中的偏见，从而改善 AI 生成内容中对酷儿个体的表现。我们的多代理框架包括专门用于偏见检测和优化的代理，重点关注代词的包容性。

2 相关工作

在语言模型中缓解偏见已经通过各种技术得到了研究，其中一些研究集中在特定的人口统计偏见。Bolukbasi 等人（2016 年）进行的一项有影响力的研究调查了词嵌入中的性别偏见，其中刻板印象的关联（例如，“男性”与“程序员”和“女性”与“家庭主妇”）盛行。他们提出了一种通过识别性别特定的子空间并中和它们来减少直接性别偏见 [2] 的方法。虽然这种方法在缓解性别偏见方面标志着一个重要步骤，但它主要涉及二元性别区别，在捕捉非二元和跨性别身份的变异性方面留下了缺口，特别是在像代词使用这类更为流动的语言背景中。

NeurIPS 2024 Workshop on Queer in AI.

另一个值得注意的工作是 Zhao 等人（2018），他们关注共指消解系统中的性别偏见。他们发现这些系统通常通过将职业与特定性别联系起来表现出偏见，在输出结果中反映了社会刻板印象。为了解决这个问题，他们引入了 WinoBias 数据集，这是一个专门设计用来测试共指消解任务中性别偏见的基准数据集 [17] 。然而，尽管这些进展显著，研究的重点依然停留在二元性别类别上，对非二元代词或酷儿个体的独特需求探索有限，而代词的使用会影响酷儿个体的代表性。

Cao 和 Daumé III（2020 年）将注意力转向 LGBTQIA+ 的表现，特别是解决性别包容的共指解析问题。他们的研究识别出语言模型在处理非二元性别代词 [5] 时面临的重大挑战。研究结果证实，标准语言模型在处理非二元性别代词时存在显著困难，常常导致性别误判。然而，他们的工作更多侧重于评估而非缓解方法，因而需要专门的解决方案。

我们的工作在这些研究的基础上，通过解决以二元为中心的方法和基于评估的研究中的局限性进一步扩展。我们引入了一个协作代理管道，其中包含专门的代理来检测和优化代词使用，特别是针对酷儿包容性。这种方法不仅仅是扩展二元偏见校正，而是通过一个多代理框架优化模型输出，积极减轻错误性别识别，旨在解决在语言模型中有效缓解酷儿偏见的重要缺口。

3 方法论

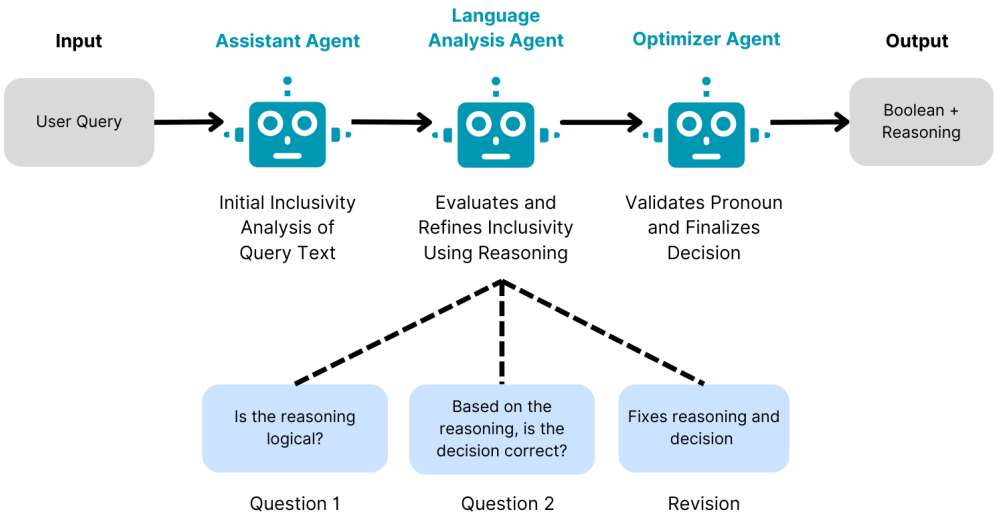

Figure 1: 该图示展示了一个旨在增强用户查询语言包容性的多代理工作流程。过程始于助手代理，它分析查询中的性别中立语言并提供初步的包容性评估。接下来，语言分析代理通过关键推理回顾并完善该评估。最后，优化代理验证并最终确定代词包容性的决定，生成布尔结果和详细的推理过程。这种结构化的方法确保了全面的包容性评估。

3.1 代理及其角色

输入：该过程从用户的问题或任务开始，范围从简单的句子到复杂的查询。输入直接转发到代理，而无需预处理。

助理代理：助理代理确定输入是否包含所有性别的所有人，并就其关于代词使用包容性的决定提供解释。

使用的提示：这是提示：{ 输入 }。指代人物时应使用性别中立的代词。请自行判断代词是否适合句子，以确保包容性。

语言分析代理：该代理分析助手代理的推理是否合理且逻辑一致。它根据推理和输入确定关于包容性的决定是否正确，并在发现差异或错误时对决定和推理进行修改。

提示使用: 这是输入: { input } 。这是一个决定: { choose_statement } 。这里是做出该决定的推理: { reasoning } 。如果代词适合句子，则决定该决定是否正确。代词应包含所有人。

优化代理：优化代理对输入的包容性做出最终决定。它结合了助理代理和语言分析代理的推理来支持其决定，并确保最终的决定与之前代理提供的分析一致。

使用的提示：这是输入：{ input }。这是一个决定：{ choose_statement }。这是做出这个决定的推理：{ reasoning }。如果代词符合句子，请决定该决定是否正确。使用推理最后决定代词是否符合句子。

3.2 设计考虑

- 顺序代理协作：通过依次使用具有特定角色的代理，我们减少了个体偏见影响最终结果的可能性。这种协作方法允许从不同的分析角度进行多次评估、错误纠正和层次推理。
- 关注代词的包容性：这个过程特别针对代词的使用，因为代词是语言中性别偏见的常见来源。确保代词具有包容性是一种在沟通中促进性别中立和包容性的切实方法。
- 通过推理实现透明性：通过要求每个代理提供推理，我们促进了决策过程的透明性。这种方法允许用户理解决策的基础，信任系统，并提供反馈。

3.3 实现细节

为了确保一致性并促进代理之间的沟通，我们使用一个 JSON 模式 [13] 来强制执行结构化输出格式。每个代理的输出包括一个表示包容性的布尔值 *choose_statement* 和一个包含详细解释的字符串 *reasoning* 。

API 调用结构的示例：

```
response = self.client.chat.completions.create(
    model='gpt-4o-2024-08-06',
    messages=messages,
    response_format={
        "type": "json_schema",
        "json_schema": {
            "name": "identifier",
            "strict": True,
            "schema": {
                "type": "object",
                "properties": {
                    "choose_statement": {"type": "boolean"},
                    "reasoning": {"type": "string"}
                },
                "required": ["choose_statement", "reasoning"],
                "additionalProperties": False
            }
        }
    }
)
```

使用严格的模式确保代理遵循预期的输出格式，从而减少错误和误解。

评估指标和实验设置

我们选择 Tango Dataset [14] 作为基准，用以评估我们的多代理框架的有效性。该数据集是专门设计用来评估模型在语言中的性别包容性敏感度，其中包含一些句子，这些句子中传统的性别代词如“he”或“she”可能不合适。相反，鼓励使用性别中立或非二元代词，例如“they”，“xe”，“ey”和“fae”。解决误性别化问题——即身份误分类的经历可能会导致潜在的痛苦和歧视——在构建包容性的 AI 系统中至关重要 [11] 。

示例案例：

- 前件: "Charlotte"
- 先行词类型：女性化
- 代词家族：“ey”

- 句子：“夏洛特是美国演员，ey 以在电影中的角色而闻名。”

在本次评估中，我们从 Tango 数据集中选择了 1,500 个样本，其中每个代词类别分为 250 个实例：

- 传统的性别代词（“he” 和 “she”）：更高的分歧率表明更好的表现，因为模型能够成功识别这些代词在特定语境中可能是不包容的。
- 非二元代词（“they”，“xe”，“ey”，“fae”）：更高的认同率表明更好的表现，因为模型支持使用这些包容性代词。

每个样本通过以下三个流程进行处理以进行比较：

- 代理工作流：整个多代理系统，包括助手代理、语言分析代理和优化代理。
- 双代理管道：一个简化版本，仅包括助理和语言分析代理。
- GPT-4o：基础模型，因其先进的能力和广泛的使用而被选择，提供了一个标准来衡量我们的多代理框架 [12] 的性能。

4 结果

本节分析了 Agent Workflow、Two-Agent Pipeline 以及基线模型 GPT-4o 在使用 Tango 数据集进行包容性代词分类中的表现。评估衡量了每个模型通过适当地同意或不同意各种代词的使用（包括非二元代词如"they," "xe," "ey," 和"fae."），来促进包容性语言的能力。

图 2 展示了多代理系统的工作流程，通过一个一致（上）和一个不一致（下）的示例。在每种情况下，辅助代理首先评估代词的包容性，然后由语言分析代理进行解释，以检查包容性指南。最后，优化代理提供整体决策，或者确认包容性选择，或者建议替代方案以实现更好的性别中立性。

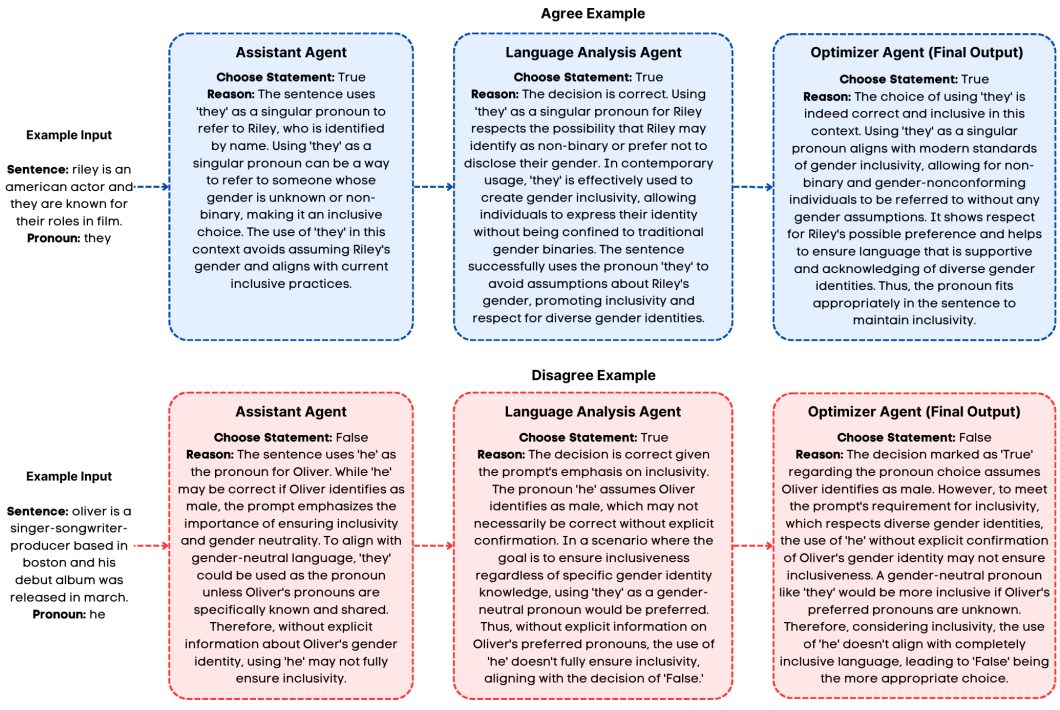

Figure 2: 代理工作流决策流程示例，展示了助手代理、语言分析代理和优化代理处理的同意和不同意情况。

表 1、2 和 3 总结了三个模型中每个代词的分类结果，展示了模型在相应句子中对代词使用的同意或不同意次数。正确响应率是根据 Tango 数据集指南中所期望的模型行为计算的。

Table 1: 完整代理工作流程在 Tango 数据集上的结果

Pronoun	Agree	Disagree	Correct Response Rate %
He	79	171	71.6 (Disagree/Total)
She	100	150	59.6 (Disagree/Total)
They	248	2	99.2 (Agree/Total)
Xe	212	38	86.0 (Agree/Total)
Ey	228	22	92.4 (Agree/Total)
Fae	245	5	98.4 (Agree/Total)

Table 2: 两个代理在 Tango 数据集上的流水线结果

Pronoun	Agree	Disagree	Correct Response Rate %
He	194	56	22.4 (Disagree/Total)
She	162	88	35.2 (Disagree/Total)
They	250	0	100.0 (Agree/Total)
Xe	226	24	90.4 (Agree/Total)
Ey	238	12	95.2 (Agree/Total)
Fae	243	7	97.2 (Agree/Total)

Table 3: GPT-4o 在 Tango 数据集上的结果

Pronoun	Agree	Disagree	Correct Response Rate %
He	149	101	40.4 (Disagree/Total)
She	186	64	25.6 (Disagree/Total)
They	250	0	100.0 (Agree/Total)
Xe	199	51	79.6 (Agree/Total)
Ey	224	26	89.6 (Agree/Total)
Fae	246	4	98.4 (Agree/Total)

4.1 研究结果

1. 传统性别代词（“他”，“她”）：
   (a) 代理工作流程实现了最高的正确响应率。
   (b) GPT-4o 表现适中。
   (c) 双代理流水线显示较低的正确响应率。
   (d) 解释：当更偏好使用包容性语言时，代理工作流程在正确反对传统性别化代词方面表现最佳。
2. 非二元代词（“他们”，“xe”，“ey”，“fae”）：
   (a) 所有模型在使用"they" 和"fae" 时表现出色，正确响应率都超过 97 % 。
   (b) 对于不常见的代词（“xe”，“ey”），双代理流水线略胜一筹。
   (c) 解释：基于代理的模型在识别不太常见的非二元代词方面优于 GPT-4o。
3. 总体正确反应率：
   (a) 代理工作流程：
      i. 传统性别代词：65.6 %
      ii. 非二元代词: 94.0 %
   (b) 双 Agent 流水线：
      i. 传统性别代词：28.8 %
      ii. 非二元代词：95.7 %
   (c) GPT-4o:
      i. 传统性别代词：33.0 %

ii. 非二元代词：91.9 %

(d) 解释：在正确处理传统的性别代词方面，代理工作流表现最佳，而两代理管道在处理非二元性别代词时略胜一筹。

4.2 统计显著性

卡方检验确认了观察到的差异的显著性：

i. 传统的性别代词：
   A. 智能体工作流 vs. GPT-4o:
   B. $\chi^2 = 38.57, p < 0.0001$
   C. 代理工作流与双代理管道：
   D. $\chi^2 = 115.31, p < 0.0001$
   E. 解释：在正确地不同意传统性别化代词方面，代理工作流程明显优于 GPT-4o 和双代理管道。

ii. 非二元代词:
   A. 代理工作流程 vs. GPT-4o：
   B. $\chi^2 = 5.89, p = 0.0152$
   C. 两代理管道与 GPT-4o：
   D. $\chi^2 = 11.93, p = 0.0006$
   E. 代理工作流与双代理流水线：
   F. $\chi^2 = 0.97, p = 0.3246$
   G. 解释：两种基于代理的模型在正确地使用非二元代词方面显著优于 GPT-4o。在这一类别中，代理工作流和双代理流程之间没有显著差异。

4.3 意义

代理工作流程在传统性别代词上取得了 65.6 % 的正确响应率，超过了 GPT-4o 的 32.6 个百分点，以及双代理管道的 36.8 个百分点，这表明它在减少不适当使用“他”和“她”方面的优势。对于非二元性别代词，两种基于代理的模型都表现出色，其中双代理管道取得了 95.7 % 的正确响应率，代理工作流程紧随其后，达到 94.0 % ，均优于 GPT-4o 的 91.9 % 。这些发现证明了多代理框架在处理包容性语言方面的准确性，证明了它们在提高公平性和减少 AI 生成内容中的代词偏见方面的价值。

5 伦理考虑

我们的研究致力于通过解决大型语言模型中对酷儿个体的偏见来推动公平和包容性。通过一个旨在提高代词包容性的多代理框架，我们旨在减少 AI 生成内容中对酷儿身份的误描绘和边缘化现象。

我们认识到，专注于代词的使用并不能捕捉到所有影响酷儿代表性的偏见形式，而且文化和语言的差异可能会影响该框架在不同背景下的效果。确保我们的方法不会无意中引入新的偏见或忽略身份的交叉性方面是很重要的。

透明性和问责制是我们方法论的基础。通过在代理管道的每一步中融入推理，我们使用户能够理解和信任决策过程，创建一个反馈和改进的环境，这对于道德的 AI 开发至关重要。

本研究不涉及高风险数据或模型，所有使用的数据集均为公开且适当注明出处的。我们遵循 AI 研究中的伦理准则，优先尊重多样性并公平对待所有个体。

通过解决语言模型中的特定偏见，我们旨在推动社会责任型人工智能的发展，从而更好地代表人类身份的多样性和复杂性。

本文成功开发并展示了一种协作代理框架，通过改进对酷儿代词的处理来增强大型语言模型的包容性。代理工作流程在正确分类传统性别代词方面比基线 GPT-4o 提高了 32.6 个百分点，并在识别非二元性别代词方面提高了 2.1 个百分点。这一进展展示了创建尊重并反映人类身份多样性的人工智能系统的潜力，促进公平访问，并减少对酷儿个体的污名化。

5.1 局限性与未来工作

尽管该框架有效地减少了代词使用中的偏见，但其范围目前仅限于特定的代词分类，并未解决与酷儿代表相关的其他形式语言偏见，例如上下文语言或隐性偏见。将该框架的验证扩展到包括更广泛的标准化数据集和基准，将增强其有效性的评估，并使其能够与其他偏见缓解技术进行比较。

未来的研究将旨在通过整合上下文推理代理和检索增强生成 (RAG) 来扩展框架，以生成不仅具有包容性而且具有上下文意识的响应。这些增强功能将使模型能够生成更相关和准确的响应，有助于 AI 系统更好地理解性别身份和社会动态。总体而言，这项研究强调了构建积极支持多样身份的包容性 AI 的重要性，推进了 AI 领域和我们对社会公平的更广泛承诺。

References

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[5] Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. *arXiv preprint arXiv:1910.13913*, 2019.

[6] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2018.

[7] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[8] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2. Association for Computational Linguistics, 2019.

[10] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

[11] Kevin A McLemore. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74, 2015.

[12] OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Accessed: 2024-09-21.

[13] OpenAI. Structured outputs. `https://platform.openai.com/docs/guides/structured-outputs`, 2024. Accessed: 2024-10-11.

[14] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. I'm fully who I am: Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266, 2023.

[15] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: A case study with google translate. *Neural Computing and Applications*, 32:6363–6381, 2020.

[16] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

A 附录

A.1 补充资源

- 源代码库：该项目的源代码在 GitHub 上公开可用，地址为 `https://github.com/Tonyhrule/Queer-Bias-LLMs`
- 评估数据集：本研究中使用的评估数据集可以在 `https://github.com/amazon-science/tango` 获得

A.2 计算资源

所有实验都在一台 15 英寸的 MacBook Air (2024) 上进行，具体规格如下：

- 芯片：Apple M3
- 内存: 24 GB
- 操作系统: macOS 14.6.1 (23G93)

由于计算密集型工作是通过 OpenAI API 完成的，本地机器的规格并没有显著影响实验的性能。每个 API 调用处理 Tango 数据集的一个样本大约需要 2 秒。通过我们的多智能体系统和基础的 GPT-4 模型处理 100 个样本，每个模型大约需要总共 6.7 分钟，总计 13.4 分钟。

该项目所需的整体计算资源较少，并且没有实质性的计算限制。除了此处报告的内容外，没有使用额外的计算资源。初步实验和失败的运行没有需要显著的额外计算时间。