arXiv:2301.05154v1 [cs.AI] 12 Jan 2023

Mephisto：一个支持可移植、可复现和迭代式众包的框架

Jack Urbanek 和 Pratik Ringshia

Meta AI jju@meta.com,
tikir@meta.com

摘要

我们介绍 Mephisto，这是一个旨在让研究众包更具可复现性、透明度和协作性的框架。Mephisto 提供了涵盖多种任务设计和数据收集工作流的抽象，并通过简洁的用户体验，将最佳实践设为易于采用的默认选项。在这份白皮书中，我们讨论机器学习研究中数据收集和标注的现状，阐述构建共享框架的动机，以便研究人员能够创建数据收集和标注工具，并将其作为出版物的一部分开源；同时概述了有助于实现这些目标的系统建议要求。随后，我们将逐步介绍 Mephisto 中的解决方案，说明所采用的抽象、围绕用户体验所做的设计决策，并分享实现细节及其与最初动机的契合之处。我们还将讨论当前的局限性，以及为持续实现框架最初目标而开展的未来工作。Mephisto 是一个开源项目[1]，，其文档可在 www.mephisto.ai 查阅。

1 动机

在当今的机器学习（ML）研究领域，我们发现，高质量众包的新方法和最佳实践从被发现到广泛应用之间，总是存在着一道鸿沟。这种脱节让试图收集高质量数据的研究人员和最终只能面对低质量任务与工作实践的工作者都经历不佳。人们认为，尽管有些项目确实致力于在众包中采用最先进的方法，但采用这些方法的人不得不重复实现工作，而不采用的人则会得到更差的结果。不过，这一说法无法直接证实；但研究显示，即使是常用数据集也存在明显的质量问题（Paullada 等，2021；Northcutt、Athalye 和 Mueller，2021）。

我们认为，这主要是因为一种常见做法所致：发表研究论文时，不提供配套数据收集方法的技术实现。2022年11月，我们对机器学习出版物和数据集的公共存储库 PapersWithCode 进行了快速审查，查看引用量最高的 35 个数据集，了解其中哪些提供了代码实现。在这些数据集中，有 31 个在不同深度上介绍了其众包方法论，作为研究的一部分；撰写本文时，其中有 18 篇论文附带了使用 代码，但只有 3 篇为论文中的数据收集和质量保证 部分提供了代码。如果研究人员想扩展这些现有的常用数据集之一，例如出于去偏的考虑，就必须从头重新实现数据收集或标注流程。

这与机器学习研究中的建模领域形成对比，在那里，代码实现已变得更加标准化。根据 PapersWithCode 的趋势数据[2], 8，2015年有代码实现的论文占比为 .5%，而 2022年的论文中这一比例达到 27.75%。这一增长与 TensorFlow 的发布和普及、以及随后 PyTorch 框架的推出和应用密切相关；尽管每年发表的机器学习和人工智能论文数量大幅增加，这一趋势依然成立。2022年的代码仓库中有 66–68% 现已基于 PyTorch，因此很难断言这些框架在推动代码共享增长方面毫无作用。

随着数据驱动的机器学习解决方案日益兴起，以及“大数据”在更广泛领域的普及（Sarker 2021），改进数据收集和标注标准的压力也随之增大。标准的深度模型架构或许能够在标签准确率和训练时间之间权衡（Rolnick et al. 2017），但随着我们开始在大型数据集上遭遇缩放定律（Kaplan et al. 2020），提升数据集质量便再次成为改进模型的一条途径。虽然这通常通过剪枝来实现（Sorscher et al. 2022），但也可以在数据漏斗的起始阶段进行处理。训练时间和成本也日益成为核心问题，因为大型模型可能需要在高能耗硬件上训练数月。

如果数据质量是一个已知 问题，我们就必须探究它为何尚未得到解决。尽管标注服务提供商通常会尝试构建工具，以改进常见情形下的标注，但这些工具无法涵盖研究前沿的情形。Paullada 等人（2021年）、Vaughan（2018年）以及 Sheng 和 Zhang（2019年）等作品对这一问题进行了研究和分析，但可以理解的是，它们并未进一步为研究人员提供改进后的工具。此外，改变并非易事。对于大多数研究实验室而言，采用更好的数据收集方法并不简单，尤其是在已经建立了现有系统，或无力投入时间和资金来构建替代方案的情况下。

[1] https://www.github.org/facebookresearch/mephisto

[2] https://paperswithcode.com/trends

为从长远角度满足这一需求，我们希望通过易于发布的共享工具，为高效的社区协作收集高质量数据提供途径。这套工具就是 Mephisto。在数据漏斗的起始阶段，让最佳实践易于分享、完善和使用，能够影响收集成本、训练成本、公平性以及模型准确率。

该项目名为 Mephisto，意指最初名为 Mechanical Turk 的“人机象棋”自动机的继任者[3] 。Mephisto 用远程操控取代了原版“人机”操作者，从而对其进行了改进；我们的框架也旨在进一步抽象传统众包的复杂性，使其远离使用者，并希望在这一过程中改善工作者和研究人员的体验。

2 相关工作

数据标注当然不是一个被忽视的领域，不过，该领域的出版物所涉及的数据收集方法大致可分为几类。针对特定领域的论文会为某项具体任务设计数据收集方案，并可能说明相关流程。综述则试图将这些数据收集工作中的经验归纳为更广泛的主题。工具和平台则尝试通过清晰的代码实现来解决问题的一部分。

2.1 新颖方法

在整个领域中，研究人员用于收集数据集的具体技术层出不穷。以下列举的内容并不全面，Vaughan（2018年）对此类工作的梳理要完整得多。

针对复杂领域内任务，已有研究开发了专用接口（Torralba、Russell 和 Yuen，2010年；Chang 等，2015年）。

模型在环设置已用于评估（Xu 等，2020年）、样本高效数据收集（Settles，2012年）和实时服务（Lasecki 等，2017年；Kamar，2016年）。离线模型评估也得到应用（Ribeiro、Singh 和 Guestrin，2016年；Li、Weston 和 Roller，2019年）。

研究者提出了多种确保数据质量的技术，不再局限于标注者间一致性，而是拓展到更先进的方法（Ghosh、Kale 和 McAfee，2011年）。一些研究还专门分析了数据质量与预算之间的权衡（Khetan 和 Oh，2016年）。

不过，从顶尖数据集可以看出，发表这些方法并不一定会同时提供配套代码（尽管上述部分研究使用了 Mephisto）。

2.2 众包综述

为了从纷繁的建议和方法中提炼出共通的见解，针对该领域的综述会尝试收集、归类并评估各种技术。这些综述通常会建议他人如何纳入自己的收集内容（Vaughan 2018；Sheng 和 Zhang 2019）。可以理解的是，它们并未进一步为研究人员提供全面的工具。

在许多情况下，综述将众包视为数据生命周期的重要组成部分，并尝试阐明特定领域的复杂性及后续步骤（Ashmore、Calinescu 和 Paterson 2021；Sambasivan 等人 2021）。一些研究概述了开发高质量数据集的完整工作流（Hutchinson 等人（2021））。这些见解无疑很有价值，但在构建工具时，很少有论文直接提及这些研究，而其中最终发布可操作代码资源的更是寥寥无几。

一篇综述（Neves 和 Seva，2020年）尝试寻找并记录所有现有的标注工具，不过这份清单显然并不全面。

2.3 完整代码解决方案

通常，发布众包代码的研究会将其作为论文的一部分，用于解决范围明确的问题，例如 Li 等人（2021）如何设置组合式对话任务，或者让研究的实验流程更轻松一些，以满足特定需求；Gureckis 等人（2016）尝试将与 Mechanical Turk 平台交互的复杂性抽象化，便是一个例子。

Mephisto 也属于后一类研究。我们从 ParlAI-MTurk（Miller 等人（2017））发展而来；该项目旨在让基于对话的研究更容易开展。我们意识到自己可能会成为又一个过时的标准[4] ，因此尝试打造一个足够通用的平台，以支持整合上述任何研究成果，并希望通过提供实用工具，帮助研究人员围绕这些工具展开交流。

3 项目目标

借助 Mephisto，我们希望解决当前众包工作在编写、使用和分发方面不够便捷的核心问题。为此，我们概述了研究标注平台应遵循的核心价值和要素，以便评估我们的进展。

3.1 分发、复现与扩展

在研究环境中，分发、复现和扩展这一流程的每个步骤，都关系到一项工作能否推动领域发展。分发将相关技术交到其他研究人员手中，分享内容可以从方法论一直涵盖到完整的代码配置。复现让新参与者能够尝试相关工作，而复现的门槛通常与能够分享的流程内容多少成反比。扩展是复现之后的下一步，旨在推动初始工作发展为新的研究。

为了促进工作的分发、复现和扩展，标注平台应使研究项目相关代码的分发变得轻松，并确保新读者能够轻松地自行配置、运行和修改这些代码。

3.2 灵活性

由于研究目标不断变化，任何旨在满足研究多样化需求的平台，都应广泛涵盖各种功能。

[3] https://en.wikipedia.org/wiki/Mephisto (自动机)

[4] 对于熟悉 https://xkcd.com/927/ 这一趋势的人来说，

然而，这种复杂性与平台的易用性背道而驰，可能会提高使用门槛，降低研究人员的接受度。理想的平台应找到兼顾这两种需求的方法。

抽象化众包与实现为了适应新的场景和情况，考察众包任务的核心要素并将其归纳为一个连贯的数据模型，是很有价值的。基于这些构件，可以构建多个互操作实现，从而对专门的收集系统实现高度控制。对于普通用户，可以提供控制选项较为简单的基础实现，让他们尽可能少依赖背景知识就能快速上手。

这种方法使在该平台上创建的工作能够避开（Gureckis 等人（2016））中所述工作的缺陷；后者虽然有助于用户入门，却会将他们绑定到特定的众包平台。

采用默认最佳实践的钩子在平台的各个组件中，研究人员应当能够精确控制各项设置，以运行特定任务。这包括工作者与任务的配对、收集管道和工作流、自动化审核工具，以及其他需要考虑的方面。为了避免让新用户不知所措，每个方面都应默认提供最佳实践，使新用户能够受益于共享的当前最佳方法。

这避免了 Hutchinson 等人（2021）和 Miller 等人（2017）所面临的问题：前者一开始灵活性过高（因为所有内容都必须由你亲自实现），而后者则将所有任务都强行纳入《对话数据收集最佳实践与技术》。

3.3 数据质量方面的考量

任何旨在促进众包的项目都必须将数据质量列为优先事项，因为无论使用和共享有多方便，如果产出的数据质量低下，就没有多大用处。在这方面，我们提出了一些重要考量，其中许多都基于这样一个现实：研究工作往往受到时间限制，要经过多轮修改和迭代，才能做出有价值的成果。

用户界面/用户体验与工作者生活质量 任务的用户界面（UI）设计是影响任务结果质量的重要因素（Finnerty 等人（2013））。设计良好的任务可以包含明确的标准、示例和格式说明，并降低认知负荷等（Wu 和 Quinn（2017））。理想的框架应能帮助忙碌的研究人员落实其中一些最佳实践。

工作者也更愿意重返那些注重良好用户体验（UX）的任务。任务设计和可用性都应尽量减少用户的挫败感。任务架构还应具备较强的容错能力。至少，在出现错误情况时，应清楚地告知用户；更进一步，还可以触发某种警报，以便研究人员及时、妥善地作出响应。

整合反馈渠道也是发现并改进可能出现的设计盲点的有效方式。我们认为，这些盲点是常态，而非例外。理想的众包框架应为研究人员提供反馈机制，用以弥补他们可能存在的各种疏漏。在实际应用中，这意味着工作者可以通过某种反馈渠道轻松地向研究人员传达意见。为工作者提供向研究人员分享反馈的机会，有助于提升工作者体验。（Bragg、Mausam 和 Weld，2018年）。

质量保证实践 标准的质量保证实践，例如工作者资格审核、黄金标准标注、标注者间一致性等，应易于查找和启用，而无需查阅文献。鼓励研究人员在设计任务时优先考虑这些要素，相较于事后才添加这些要素的任务，能够带来更高的数据质量。此外，试点和与工作者沟通等常见工作流可能至关重要，因此也应易于开展。

工作者多样性与代表性 收集数据集时，确保尽可能广泛且具有代表性的贡献者群体参与数据处理，是数据质量的一个重要方面。这往往受限于提供人群的公司的工具，有时相关信息还会被视为私人信息。尽管如此，优秀的众包平台应提供相应工具，鼓励研究人员通过简化入门引导、设置每位工作者的任务上限，以及使用多个众包平台等最佳实践来扩大来源人群。至少，平台应能够报告来源人群的一些指标，并可与数据卡（Pushkarna、Zaldivar 和 Kjartansson，2022年）等工具集成。

4 当前实现

遵循上一节所述的价值观和原则，我们设计了 Mephisto，其基础是一组抽象、若干初始实现，以及一些最佳实践要素，旨在提升任务质量和研究人员体验。本节旨在从技术角度概述 Mephisto 当前的运作方式。

4.1 抽象

我们首先介绍 Mephisto 的底层抽象，这些抽象旨在将众包的复杂性拆解为各个组件，以此构建架构。介绍完数据模型后，要了解 Mephisto 的其余部分就几乎和**A**rchitect、**B**lueprint、**C**rowdProvider、**D**atabase[5] 一样简单。

任务 中 包含什么？数据模型 为了理解众包，我们定义了若干概念，用以表示数据模型的底层要素。

首先是任务（Task），可将其视为一组彼此直接相关、需要完成的工作，例如“为 50,000 张图像标注不同的分割掩码”。

低于此层级的是任务运行（TaskRun），可以将其视为您可能运行过的一项单独任务。在上面的 50,000 张图像中，您可能希望为前 1,000 张图像标注 3 个可能的掩码标签，以开展试点。这将是一个合适的

[5]我们保证，这并非有意设计的初始化序列

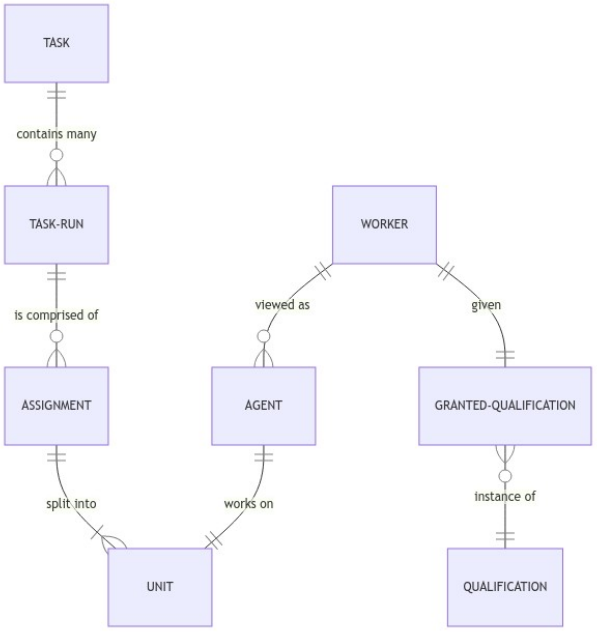

图 1：Mephisto 众包数据模型概览

任务运行（TaskRun）。（一个任务（Task）可能只有一次任务运行，但通常会有多次任务运行。）

在一个任务运行中，可以包含许多分配任务，每个分配任务都可以视为一个需要完成的独立事项。它的粒度通常对应于你展示给工作者的内容，例如“为这 5 张图像标注分割掩码”。

一个分配任务可以拆分为多个工作单元，代表某个个体在任务中所做的贡献。有些分配任务可能只有一个工作单元；不过，例如，如果你希望某个分配任务被标注两次以检查标注者间一致性，就可以设置两个工作单元；又或者在对话任务中，需要两名工作者同时相互交流，也可以设置两个工作单元。

对于实际执行工作的人，我们使用工作者来记录每个个体为你完成的所有任务及其全部历史记录。

为了区分工作者的完整历史记录，我们还引入了代理，它可以视为工作单元与工作者的配对，用来表示该工作者为特定工作单元所做的工作。

这些是支撑 Mephisto 的底层数据模型组件。借助这些术语，我们可以开始梳理标注任务中其余流程。

托管任务：架构器 架构器包含用于设置服务器的脚本，这些服务器负责运行 Mephisto 中的任务。它们使研究人员能够将 Mephisto 与不同的云配置结合使用。为此，它们涵盖任务期间的服务器生命周期，因此应实现用于准备、部署和关闭服务器的方法。它们还定义了外部工作者连接到 Mephisto后端所需的接口。

任务：蓝图 蓝图是 Mephisto 中不同任务的核心，旨在同时涵盖某项任务中常见的任务流程，以及允许用户自定义和扩展该任务的配置设置。它们定义特定任务的输入和输出，以及整体任务接口。这种特定抽象需要几个重要组件，具体如下：

- AgentState 定义了收集工作单元期间将保存的数据格式。
- TaskRunner 定义了执行任务所需的所有后端逻辑。
- TaskBuilder 定义了任务开始前需要构建的所有资源。通常包括作为任务一部分托管的前端。
- SharedTaskState 可用于保存 TaskRun 中所有工作单元共享的实时状态信息，通常在分配工作时会用到。

工作者：众包提供者 CrowdProviders 使 Mephisto 能够连接不同的众包群体来处理任务。它们与抽象的工作者、代理和工作单元密切交互，具体如下：

- <众包>工作者构成了工作者的长期身份，也是一个接口，Mephisto 可通过它集成与提供者 API 对接的工作者专属功能。这些功能可能包括封禁、授予资格以及直接通信。
- <众包>单元是用于对接工作发布远程钩子或类似机制的接口。它们需要跟踪外部状态，还应提供向提供者注册和使工作请求过期的接口。
- <众包>代理负责连接工作者与单个工作单元，并且必须实现用于检查其远程状态以及将工作标记为已完成或已拒绝的方法。

结果存储：数据库数据库使 Mephisto 能够存储结果，无论你使用何种服务器配置。为此，我们提供了 MephistoDB 抽象，其中列出了运行 Mephisto 所需实现的所有数据库调用。

4.2 Mephisto 架构

在实践中，Mephisto 能够处理 Blueprint、Architect、CrowdProvider 和 Database 的任意配置，并协调它们各自在所组成的 TaskRun 中的初始化、部署、监控和关闭。在这样的实时任务运行过程中，它还会报告指标并保存部分结果。我们的一个关键目标是，让用户希望自定义的任何“业务逻辑”都能在 抽象 中找到明确的扩展点，这样大多数用户就不必处理这些接口如何协调的复杂性。数据收集完成后，Mephisto 会提供一些工具，方便用户与存储在 MephistoDB 中的数据交互并进行探索。

4.3 蓝图

我们的目标是，在大多数情况下，Mephisto 的大多数用户都能依赖现有的蓝图，而不必从头编写新的蓝图。为此，我们提供了一些实用的实现方案，涵盖了广泛的使用场景。

- ReactStaticBlueprint 是一种设置方式，用户可以提供任意使用 React 编写的简单数据收集前端应用程序，并将其视为单轮任务。简而言之：工作者会收到一些数据，远程处理这些数据，然后返回结果。
- RemoteProcedureBlueprint 支持更复杂的设置，前端应用程序能够直接查询任务设置期间指定的某个后端。这样便可以执行在工作者端无法完成的处理，例如在循环中运行模型。

- StaticHTMLBlueprint 是进入 Mephisto 最简单的入门途径，因为它接受研究人员可能比 React 更熟悉的标准 .html 文件。不过，它的功能不如其他选项丰富。

除此之外，ParlAIChatBlueprint 是实时任务的一个很好的示例，它具有明确且高度可配置的流程，适用于以对话为重点的工作。

质量保证混入模块 我们提供了一些用于质量保证的混入模块，可用于通过 ReactStaticBlueprint 或 RemoteProcedureBlueprint 运行的任何内容。将这些混入模块纳入 Blueprint 后，该 Blueprint 就会启用指定功能，并知道如何处理这些功能。

- OnboardingRequired 混入模块允许研究人员为以前没有完成过该任务的工作者设置单独的流程，让他们了解任务要求。
- UseGoldUnit 混入模块提供了一套熟悉的流程，用于提供已知良好示例，并定期以此评估工作者，作为质量检查。
- ScreenTaskRequired 混入模块允许研究人员指定工作者为某项工作完成的第一个实际工作单元（通常是易于验证的工作单元），从而使研究人员能够在分配更多工作之前进行自动化分析和验证。

研究人员可以利用这些方法的基本构件，将严谨的质量保证流程融入任务中，而无需在任务底层的验证措施之上构建复杂机制。当然，仍需进行若干轮初步试点和调整，以确保验证器经过良好校准。

4.4 众包提供商

目前我们已实现的主要众包提供商是 MTurkProvider 和 MockProvider。前者支持直接与亚马逊土耳其机器人平台进行交互，而后者支持在本地测试任务，或允许人们以“模拟”指定工作者的方式访问。添加更多众包提供商的工作仍在进行中。

4.5 架构器

撰写本文时，目前可用的架构器包括 HerokuArchitect 和 LocalArchitect。前者允许使用 Heroku 云服务作为提供商来启动任务。后者允许在运行 Mephisto 的机器上托管，这适用于本地测试，或从同一局域网中的研究参与者处收集数据。我们还有一种基于 EC2 的架构器，使用时需要在 AWS 注册域名。

4.6 前端软件包

启动众包任务的另一个环节是设计任务的用户界面。Mephisto 允许研究人员启动用户界面采用纯浏览器 HTML 实现的任务；对于更高级的情况，也可以使用 React JavaScript 库实现。

对于 React 实现，Mephisto 提供了一个名为 mephisto-task 的 npm（Node 包管理器）软件包。该软件包使研究人员能够通过前端代码，无缝对接 Mephisto 后端蓝图。它提供后端传来的任务数据、用于处理 CrowdProvider 任务提交的回调，以及可用于按条件显示不同视图（例如任务预览、入门引导、错误、提交内容等）的布尔标志。

该软件包还提供三个 React Hooks：useMephistoTask、useMephistoLiveTask 和 useMephistoRemoteProcedureTask。后两个可分别用于更高级的任务，例如聊天机器人或模型在环任务。我们尤其认为，模型在环有望通过实时增强工作者的表现、减少繁琐且重复的工作、改善感知到的用户体验，以及提供实时验证和反馈来提高任务结果的质量，不过这一领域仍需更多研究。模型在环方法还可根据先前的任务动态生成后续任务，从而在任务运行已经开始时，仍能灵活控制接下来要启动哪些任务。

Mephisto GitHub 仓库中的 examples/ 文件夹提供了采用简单设置的任务模板示例，以及包含聊天机器人和模型在环功能的高级设置示例。

4.7 工人反馈

借助 Mephisto 可扩展的架构，创建插件也很简单。我们通过 npm 软件包 mephisto-worker-addons 提供两个官方插件，帮助改善任务的工作者体验。具体来说，我们提供 Feedback 和 Tips 这两个 React 组件。

反馈组件允许工作者在完成任务时向研究人员提出建议。这些建议可以包括问题、发现的错误，

也可以是积极的认可。这个与研究人员沟通的渠道有助于改善工作者的情绪并提高任务质量。研究人员还可以选择给提交有价值反馈的人小费或奖金。说明这一奖励方案也能建立有效的激励机制，鼓励人们提供提示。

Tips组件允许工作者创建共享的常见问题式维基，其他工作者可以从中受益。该组件内置了内容审核机制，因为提交的Tips需要经过研究人员批准后，其他工作者才能看到。除了实用之外，这些示例还展示了如何扩展 Mephisto，以满足定制化的研究需求。

4.8 审核工具

Mephisto还包含一个基于Python的命令行界面（CLI）工具，允许用户审核任务结果，或更广泛地审核任意数据。该命令接受一个输入数据源和一个“审核模板”，并启动本地服务器，以便浏览数据。

对于任意数据，只需将输入文件通过管道传入即可：

```
cat input.jsonl | mephisto review --json
my-review-interface --stdout
```

如果要专门用于Mephisto任务运行，则可以使用--db标志：

```
mephisto review --db task-name
my-review-interface --stdout
```

为了便于审核，我们提供了一个基于 create-react-app 的 React 模板，它实现了模块化的渲染架构。研究人员只需将自定义渲染器实现为单个 React 组件，即可轻松定义“数据项”的呈现方式。模板开箱即用地附带了几个默认渲染器，例如用于文本密集型任务的JSON渲染器和词云渲染器。

任务运行完成并积累数据集后，研究人员可以将结果与基于 Mephisto 的审核和可视化工具一同作为出版物的一部分分享。Ego4D 项目（Grauman等人，2021年）使用了 Mephisto 的基础审核工具，分享了其收集的 3,000 小时自我中心视频[6]。

4.9 工人资格

Mephisto 提供了一种简单的设置方式，可出于任何原因为工作者添加标记；你可以为任何工作者创建并分配任意资格。我们发现，这在设置允许名单和屏蔽名单、根据记录的技能查询或筛选工作者，以及创建复杂任务流程（例如，参与一个角色会导致其无法参与另一个角色的流程）时很有用。

5 未来工作

Mephisto 是一个不断演进的系统，我们会继续秉持本文档所列的价值观，对其进行迭代和开发。

随着我们发现新的强大众包方法，我们计划将其作为顶层功能纳入 Mephisto。我们也希望为平台上的所有人提供便捷的方式，以构建新的钩子和功能，并与其他任务开发者分享。此外，我们希望扩展 Mephisto 开箱即用所支持的现有工具集。最后，我们计划提升平台的可移植性，使其尽可能兼容更多服务提供商、更多托管方案和更多任务。

我们还希望继续从任务设计的维度进行建设——让社区贡献的设计和任务模板更易于共享和使用，在用户界面和用户体验的最佳实践出现时及时采用，并尝试新的机制来改善工作者体验，例如游戏化。

不过，即使采取了这些措施，我们也只是触及了当今最佳实践实施工作的表面，更不用说适应未来的最佳实践了。我们希望这项工作能够成为未来工作进一步发展的基础。我们的路线图可在 GitHub 项目页面上查看，我们也欢迎大家就项目的发展方向提供反馈。

5.1 贡献

Mephisto 是一个开源项目，我们重视用户的贡献。我们欢迎任何人加入我们，帮助实现这一愿景：让众包变得简单、可复现，并内置最佳实践，详见我们的 GitHub[7]。觉得我们做错了什么，或者遗漏了人们应该立即使用的技术？太好了！请提交 issue，或者更进一步，发起一个拉取请求。

伦理声明

Mephisto 是一款采用宽松许可证的软件，可用于众包。Mephisto 平台旨在改进标注方法，并促进协作，以解决数据收集问题；但要实现这些目标，它显然仍在持续完善中。平台不会直接将这些目标设为对用户的约束，因此，虽然我们努力将目前公认的最佳实践设为默认选项，但用户可以覆盖这些选项。

因此，欠薪或对工作者的不当对待、收集有偏数据集以及数据许可问题仍可能出现。使用 Mephisto 的研究人员仍须尽到应有的审慎义务，确保了解其数据收集所适用的最新最佳方法。

致谢

我们感谢所有为 Mephisto 作出贡献的公众贡献者，以及 ParlAI 团队和其他早期试点用户。

参考文献

Ashmore, R.；Calinescu, R.；以及 Paterson, C. 2021. 确保机器学习生命周期：期望、方法与挑战。*ACM*计算机调查，54(5)。

[6] https://ego4d-data.org/docs/viz/

[7] https://www.github.org/facebookresearch/mephisto

Bragg, J.；Mausam；以及 Weld, D. S. 2018. Sprout：由众包驱动的众包任务设计。载于第 *31* 届 *ACM* 用户界面软件与技术年度研讨会论文集，UIST '18，165–176。美国纽约州纽约市：计算机协会。ISBN 9781450359481。Chang, A. X.；Funkhouser, T. A.；Guibas, L. J.；Hanrahan, P.；Huang, Q.；Li, Z.；Savarese, S.；Savva, M.；Song, S.；Su, H.；Xiao, J.；Yi, L.；以及 Yu, F. 2015. ShapeNet：信息丰富的三维模型库。*CoRR*，abs/1512.03012。Finnerty, A.；Kucherbaev, P.；Tranquillini, S.；以及 Convertino, G. 2013. 简单为上：众包中的奖励与任务设计。载于 *SIGCHI* 意大利分会双年会议论文集，CHItaly '13。美国纽约州纽约市：计算机协会。ISBN 9781450320610。Ghosh, A.；Kale, S.；以及 McAfee, P. 2011. 谁来监管监管者？对用户生成内容进行滥用检测的众包。载于第 *12* 届 *ACM* 电子商务会议论文集，EC '11，167–176。美国纽约州纽约市：计算机协会。ISBN 9781450302616。Grauman, K.；Westbury, A.；Byrne, E.；Chavis, Z.；Furnari, A.；Girdhar, R.；Hamburger, J.；Jiang, H.；Liu, M.；Liu, X.；等。2021. Ego4D：在全球范围内采集的 3,000 小时自我中心视频。*CoRR*，abs/2110.07058。Gureckis, T.；Martin, J.；McDonnell, J.；Rich, A.；Markant, D.；Coenen, A.；Halpern, D.；Hamrick, J.；以及 Chan, P. 2016. psiTurk：用于在线开展可复制行为实验的开源框架。《行为研究方法》，48(3)：829–842。出版商版权：© 2015，Psychonomic Society, Inc. Hutchinson, B.；Smart, A.；Hanna, A.；Denton, E.；Greer, C.；Kjartansson, O.；Barnes, P.；以及 Mitchell, M. 2021. 迈向机器学习数据集问责：软件工程与基础设施领域的实践。载于*2021* 年 *ACM* 公平性、问责制与透明度会议论文集，FAccT '21，560–575。美国纽约州纽约市：计算机协会。ISBN 9781450383097。Kamar, E. 2016. 混合智能的发展方向：以人类智能补充人工智能系统。载于第二十五届国际人工智能联合会议论文集，IJCAI'16，4070–4073。AAAI 出版社。ISBN 9781577357704。Kaplan, J.；McCandlish, S.；Henighan, T.；Brown, T. B.；Chess, B.；Child, R.；Gray, S.；Radford, A.；Wu, J.；以及 Amodei, D. 2020. 神经语言模型的缩放定律。Khetan, A.；以及 Oh, S. 2016. 广义 Dawid-Skene 模型下的可靠众包。*CoRR*，abs/1602.03481。Lasecki, W. S.；Miller, C. D.；Naim, I.；Kushalnagar, R.；Sadilek, A.；Gildea, D.；以及 Bigham, J. P. 2017. Scribe：深度融合人类与机器智能，实现实时语音字幕。*Commun. ACM*，60(9)：93–100。

Li, M.; Weston, J.; and Roller, S. 2019. ACUTE-EVAL：通过优化问题和多轮比较改进对话评估。*CoRR*, abs/1909.03087. Li, Y.; Arnold, J.; Yan, F.; Shi, W.; and Yu, Z. 2021. LEGO-Eval：一个通过众包评估对话系统的开源工具包。*CoRR*, abs/2105.01992. Miller, A. H.; Feng, W.; Fisch, A.; Lu, J.; Batra, D.; Bordes, A.; Parikh, D.; and Weston, J. 2017. ParlAI：一个对话研究软件平台。Neves, M.; and Seva, J. 2020. Annotationsaurus：一个可搜索的标注工具目录。*CoRR*, abs/2010.06251. Northcutt, C. G.; Athalye, A.; and Mueller, J. 2021. 测试集中的普遍标签错误会破坏机器学习基准。Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. 数据及其（不）满：机器学习研究中数据集开发与使用综述。*Patterns*, 2(11): 100336. Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. 数据卡：面向负责任人工智能的有目的、透明的数据集文档。*arXivpreprint arXiv:2204.01075*. Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "我为什么应该相信你？"：解释任意分类器的预测。*CoRR*, abs/1602.04938. Rolnick, D.; Veit, A.; Belongie, S.; and Shavit, N. 2017. 深度学习能够抵御大规模标签噪声。Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "每个人都想做模型工作，而不是数据工作"：高风险人工智能中的数据级联。载于*2021*年计算机系统中的人为因素会议论文集，CHI '21。美国纽约州纽约：计算机协会。国际标准书号 9781450380966. Sarker, I. 2021. 机器学习：算法、现实世界应用与研究方向。*SNComputer Sci-ence*, 2. Settles, B. 2012. 主动学习。人工智能与机器学习综合讲座, 6(1): 1–114. Sheng, V. S.; and Zhang, J. 2019. 众包机器学习：过往研究与未来方向简述。*AAAI*人工智能大会论文集, 33(01): 9837–9843. Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; and Morcos, A. S. 2022. 超越神经网络缩放定律：通过数据剪枝打破幂律缩放。Torralba, A.; Russell, B. C.; and Yuen, J. 2010. LabelMe：在线图像标注与应用。通过开发一款公开可用的工具，使用户能够利用互联网快速、轻松地标注图像，作者得以收集大量详尽的图像描述。Vaughan, J. W. 2018. 更好地利用众包：众包如何推动机器学习研究。机器学习研究期刊, 18(193): 1–46. Wu, M.-H.; and Quinn, A. J. 2017. 混淆众包：亚马逊土耳其机器人上的任务指令质量。载于*HCOMP*。

Xu、J.、Ju、D.、Li、M.、Boureau、Y.、Weston、J. 和 Dinan、E. 2020年。《开放域聊天机器人安全指南》。*CoRR*, abs/2010.07079。