arXiv:2303.02884v1 [cs.HC] 6 Mar 2023

模型草图绘制：聚焦早期机器学习模型设计中的概念

Michelle S. Lam
斯坦福大学
斯坦福，CA, USA
mlam4@cs.stanford.edu

Zixian Ma
斯坦福大学
斯坦福，CA, USA
zixianma@cs.stanford.edu

Anne Li
斯坦福大学
斯坦福，CA, USA
anne24@stanford.edu

Izequiel Freitas
斯坦福大学
斯坦福，CA, USA
ifreitas@stanford.edu

Dakuo Wang*
东北大学
波士顿，MA, USA
d.wang@northeastern.edu

James A. Landay
斯坦福大学
斯坦福，CA, USA
landay@stanford.edu

Michael S. Bernstein
斯坦福大学
斯坦福，CA, USA
msb@cs.stanford.edu

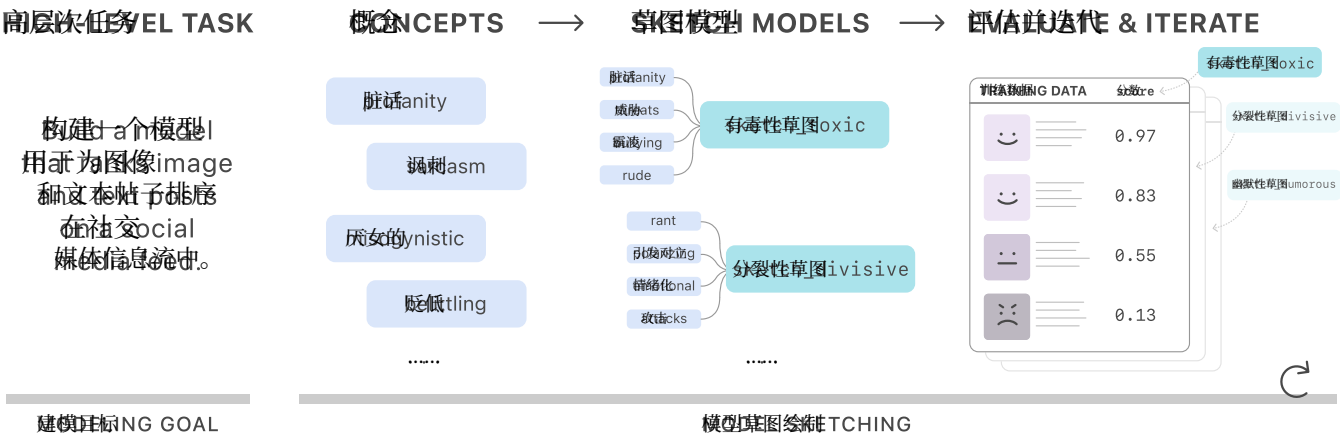

图 1：模型草图工作流程的概念图。模型草图绘制只需几分钟即可让机器学习从业者围绕高层概念编写早期功能模型。我们在 ModelSketchBook 中实例化了这一方法；ModelSketchBook 是一个用于计算笔记本的 Python 软件包。概念：用户指定与决策任务相关的概念。我们的系统将每个概念转换为一个功能模型，用于对未见样本在该概念上的表现进行评分。草图模型：用户训练一个草图模型，该模型聚合各个概念，以预测整体建模目标的真实评分。评估与迭代：用户在训练数据上评估草图模型的预测结果，以理解其行为和不足之处。随后，他们会对概念和草图模型进行迭代，以探索替代方法，使模型更好地符合其目标。

抽象的

机器学习从业者常常会陷入模型架构和性能指标等底层技术细节。早期的模型开发是否可以转而聚焦于更高层次的问题，例如模型应该关注哪些因素？受设计领域草图绘制实践的启发——这种实践将想法提炼为最简表示——我们提出了模型草图绘制：一种技术框架，用于快速、迭代地编写机器学习模型决策逻辑的函数近似。模型草图绘制将从业者的注意力重新聚焦于组合模型预期要推理的、由人类理解的高层次概念

*作者在斯坦福以人为本人工智能研究所访问期间完成了这项工作。

允许免费制作本作品的全部或部分内容的数字或纸质副本，供个人或课堂使用，但不得为营利或商业利益制作或分发副本，且副本必须附有本声明及首页的完整引文。本作品中由作者以外的其他人拥有版权的部分必须遵守其版权规定。注明出处后，可以摘录。若要以其他方式复制、重新出版、发布到服务器或向列表重新分发，须事先获得特别许可和/或支付费用。请联系 permissions@acm.org 申请许可。
*CHI '23，2023年4月23日至28日，*德国汉堡© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9421-5/23/04...$15.00
https://doi.org/10.1145/3544548.3581290

（例如，在内容审核任务中识别脏话、种族主义或讽刺），并采用零样本概念实例化。在一项涉及 17 位机器学习从业者的评估中，模型草图绘制将思考重心从实现转向更高层次的探索，促使参与者迭代更多样的模型设计，并帮助他们发现问题表述中的不足——而所用时间仅为通常构建模型所需时间的一小部分。

CCS 概念

• **以人为中心的计算** → **人机交互 (HCI)**；交互式系统与工具；• **计算方法** → 人工智能；机器学习。

ACM 参考格式：
Michelle S. Lam、Zixian Ma、Anne Li、Izequiel Freitas、Dakuo Wang、James A. Landay 和 Michael S. Bernstein。2023年。模型草图设计：在机器学习模型设计早期聚焦概念。载于 *2023*年*CHI*计算机系统中的人为因素会议论文集（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约市，共24页。https://doi.org/10.1145/3544548.3581290

1 引言

设想你是一名机器学习从业者，任务是在某个社交媒体网站上构建一个模型，用来标记仇恨图像迷因。当你着手构建模型并不断迭代时，你要解决的问题很快就变成了："如果我们用这些示例进行训练并采用这些超参数，Transformer模型会表现如何？"一旦进入这种模式，就很难跳出来。我们怎样才能帮助你在早期开展模型开发工作时，转而专注于这样的问题："模型在判断哪些内容具有仇恨性时，应该关注哪些因素——例如暴力、种族主义、性别歧视和脏话？"机器学习工具能否帮助我们在早期将重点放在如何执行解决方案之外，而是首先关注应当探索哪类解决方案？

当前的机器学习（ML）工具和方法助长了一种隧道式思维，将注意力集中在技术实现上 [29]。这些工具会把我们带入一连串低层级任务——收集数据集、提取特征、规划数据流水线、权衡模型架构、确定性能指标——所有这些都是为了得出模型的第一个版本。这种隧道式思维很危险，因为它会将我们锁定在单一的问题表述上，而这种表述往往会在后续迭代中延续下去 [45]。与此相反，在流程早期，最重要的恰恰是对问题表述本身进行迭代；此时，诸如"检测仇恨帖子"这类高层次、定义尚不明确的决策目标，正被转译为众多具体问题定义中的一种 [27, 45, 46]。问题领域越是价值导向、伦理挑战越大，机器学习从业者在模型设计过程中就越需要谨慎，确保转译得当。遗憾的是，这项转译工作往往被埋没，也得不到足够重视 [68]。

为了在工作中抵御类似的技术隧道效应，设计师长期以来一直采用一种草图绘制实践，有意将高层次设计可能性的自由探索置于首位。Buxton 将草图绘制定义为任何一种设计过程，其产出都具有快速、丰富、模糊且细节最简的特点 [6]。草图将一个想法提炼为最简表示，把注意力集中在最重要的设计方向要素上。通过这种方式，草图成为从业者理解初步设计的宝贵认知工具 [16]。机器学习模型开发的恰当最简表示是什么？交互式机器学习方法（如 Wekinator、Create ML 和 Teachable Machine [2, 17, 21]）以及基于提示的方法（使用大语言模型 [5, 63]和 Stable Diffusion [53]）可以实现快速的模型开发，但它们都围绕输入和输出展开，在很大程度上将中间的决策逻辑抽象掉了。对于问题构建任务，我们需要直接迭代的恰恰是决策逻辑：模型应考虑哪些因素来作出决策？如今的工具尚无法让机器学习从业者绘制模型的决策逻辑草图，以便快速解答这类关键的模型设计问题。

本文介绍了模型草图绘制，这是一种技术框架，使机器学习从业者能够创建富有表现力、类似草图的机器学习模型版本。为了能够直接迭代模型的核心逻辑，我们将概念——即与决策任务相关的人类可理解的理念或因素——作为机器学习草图的最简表示。例如，在仇恨图像迷因检测任务中，概念可以是"脏话""暴力"和"性别歧视"等具体因素，也可以是"讽刺"和"幽默"等更抽象的因素。尽管以往研究探索过机器学习中的面向概念的抽象 [32, 33, 44, 52], ，我们的方法有一个关键区别：我们不仅以概念为中心，还借助零样本方法，实现任意概念的快速实例化与迭代。概念与灵活、快速创作的结合，开启了新的交互方式，让用户能够直接绘制模型逻辑草图。我们通过一个名为*ModelSketchBook*的 Python 软件包实现了模型草图构建框架。该软件包利用 GPT-3 和 CLIP 等预训练模型在零样本学习方面的最新进展 [5, 51]，使用户能够将开放式概念描述转化为可用的构建模块。在计算笔记本环境中加载我们的工具后，用户可以指定"性别歧视"或"脏话"等概念，并提供若干图像或文本示例。ModelSketchBook 借助预训练模型为每个输入评分，判断其可能存在多大程度的性别歧视，或包含脏话的可能性有多高；随后，用户将多个概念组合起来，训练一个整体的草图模型。通过交互式地创作概念、将其纳入草图模型并检查由此产生的建模成果，用户可以灵活探索不同的建模方法及其影响。

我们通过演示案例研究和实地评估，展示模型草图绘制框架的泛化能力。我们提供了涵盖政治候选人调查、创意灵感和审稿人偏见审计等主题的案例研究（图 2）。随后，在一项实地评估中，我们邀请 17 位机器学习从业者撰写草图，以判定是否应从社交媒体网站移除源自仇恨表情包挑战赛 [31]的互联网迷因。这项任务在技术难度方面要求很高（涉及相互作用的文本和图像模态），包含主观判断（提供足够的解释灵活性，以构建广泛的模型设计空间），也涉及伦理权衡（需要考虑多个相互竞争的观点）。

模型草图绘制成功地促使参与者从技术思维转向更高层次、以设计为导向的思维，开始思考模型应该采集和不应该采集哪些概念

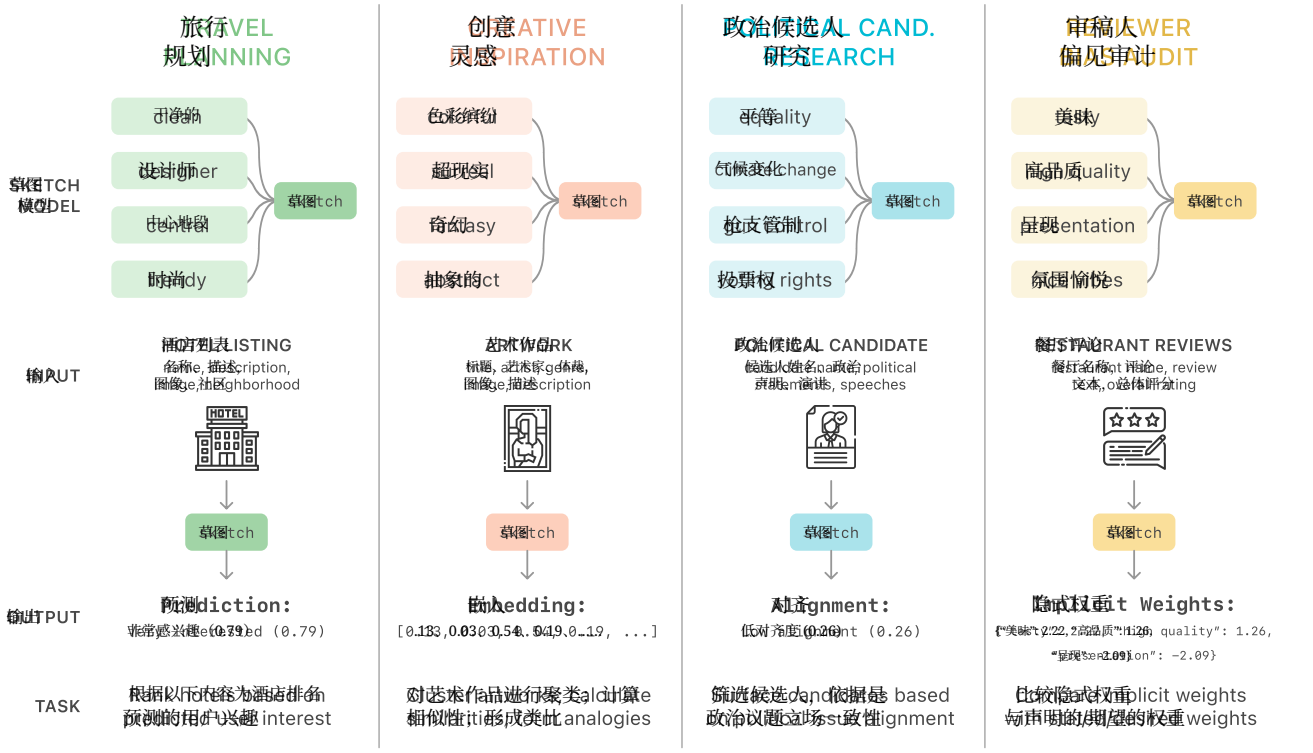

图 2：模型草图绘制案例研究：(1) 旅行规划，一项个性化推荐任务；(2) 创意灵感，一项无监督聚类任务；(3) 政治候选人研究，一项个性化搜索任务；以及 (4) 餐厅评论，一项审稿人偏见审计任务。

在接触模型草图绘制之前，参与者最初在为这项任务提出建模方案时，关注的是技术细节（例如卷积神经网络和 Transformer 等模型类型，以及 F1 和 AUC 等性能指标）。使用 ModelSketchBook 后，参与者转而从更高层次的设计和问题表述决策角度描述自己的方法。参与者通过反复尝试不同概念，弥补早期草图模型暴露出的不足：所有参与者都扩展了最初的概念集合，最终涵盖了从“种族主义”和“伊斯兰恐惧症”到“讽刺”“反讽”和“争议性”等概念；如果采用传统方法开发模型，这些概念可能不会被纳入考量。借助这种富有表现力、轻量化的交互方式，参与者将注意力从建模延伸至其他方面，并发现了数据集、标注方法和问题表述中的不足；他们指出了标签不一致、类别不平衡、某些类型的伤害缺乏代表性，以及文本与图像语义之间的交互细微之处等问题。参与者最初估计，为这项任务构建模型需要几天到几个月的时间；但使用我们的工具 20 至 30 分钟后，绝大多数参与者（17 人中有 14 人）对自己创建的模型感到满意。

我们认为，模型草图绘制是一个重要步骤，它让人们摆脱事后修复失效模型的做法，转而从一开始就帮助机器学习从业者成为更优秀的模型设计者。总而言之，本文做出了以下贡献：

- **模型草图绘制**。我们提出了一个用于构建零样本概念构件的框架，这些构件可以灵活组合成组合模型，使机器学习从业者能够快速迭代概念，为模型的决策逻辑绘制草图。
- **ModelSketchBook 工具**。我们通过推出一个开源 Python 软件包来实现我们的模型草图设计框架，用户可以在计算笔记本中交互式地构建概念并绘制模型草图。

- **与机器学习从业者共同评估模型草图绘制**。我们与 17 位机器学习从业者开展了实地评估，他们使用我们的工具为一项具有挑战性的仇恨模因检测任务构建模型。我们发现，机器学习从业者能够成功地从技术实现细节转向更高层次的决策逻辑。这种认知转变有助于他们更广泛地探索模型设计空间，并快速发现数据集、标注方法和问题表述中的不足。

2 相关工作

我们综述了设计和人机交互领域中关于草图绘制与原型设计的研究、机器学习从业者的工作实践，以及当前

早期模型设计的方法，这些内容共同促成了我们的模型草绘方法。

2.1 设计与人机交互中的草图绘制和原型设计

设计与人机交互领域长期以来都推崇草图绘制这一有价值的思考工具 [6, 16]。成功的草图绘制或原型设计工具有一个重要特点：它能让认知过程重新聚焦于高层次的迭代，使实践者得以进行“设计思维，而非实现调试” [23]。草图与原型相似，二者都是设计概念的具体呈现，但草图的用途不同：它旨在探索设计方法并提出问题，而非给出答案或解决方案 [6]。因此，草图适用于设计过程的早期构思阶段，而原型则用于后期阶段。

什么是草图的特征？Buxton 将草图描述为快速、丰富、含义模糊且细节最简 [6]。快速和丰富这两个特征相辅相成：快速周转的方法能让从业者探索并行的设计路径，这种做法有助于增加迭代次数、避免有害的设计固着，并带来更多样、质量更高的设计 [13, 59]。草图是强大的工具，因为它们能让从业者快速迭代，测试截然不同的高层次想法。

含义模糊和细节最简这两个特征也密切相关：草图的许多价值，既在于将想法提炼成风格化的最简表示这一认知任务 [39]，也在于保留足够的模糊性，从而向自己和他人暗示多种解读方式（进而指向多种未来方向） [20]。用户界面设计的早期研究表明，过于详尽的原型制作方法会妨碍这项宝贵的认知工作 [38]，也不会显著提升反馈质量 [61]；事实上，较简单的原型与更好的设计成果相关 [65]。

最后，草图还应具有表现力：它必须提供足够的灵活性和控制力，使实践者能够传达自己的想法并不断迭代完善 [6, 38]。传统的徒手草图富有表现力，因为设计师可以在具体或抽象的视觉和文本表征上进行绘制，以完善自己的想法，让他人能够感知其中的感知特征和功能关系 [58]。

由于人工智能的能力具有不确定性，且输出复杂，机器学习领域给传统原型设计方法带来了独特挑战 [66, 67]。然而，我们发现，基于概念的抽象尤其适用于机器学习模型编写中的草图式方法。关于数据标注中概念演化的先前研究表明，许多决策目标具有可变性和主观性，但可以通过反复进行概念引导与完善来明确这些目标 [7, 34]。研究人员甚至发现，对于法官、医生和管理者等现实生活中的决策者，基于少量人类可解释特征的简单统计推导规则，其表现显著优于人类专家 [30]。概念与人类对决策任务进行推理的方式十分契合，而统计聚合概念的方法效果出乎意料地好，甚至超出了原型设计的范畴。

2.2 机器学习实践者：工作实践与认知重点

在我们的研究中，我们交替使用“机器学习从业者”和“模型开发者”这两个术语，指代主要从事机器学习模型设计、开发和迭代工作的用户 [43, 68]。近期文献将数据科学和机器学习项目工作流描述为一个多阶段的协作生命周期：领域专家专注于高层次的问题定义和结果解读，而机器学习从业者则将认知重点放在底层实现细节上 [50, 62, 68]。然而，当前工作流的一个主要问题是，机器学习从业者为衔接高层次目标与底层实现而开展的大量转译工作（即通过问题表述、特征提取和特征工程实现的工作）往往不为人所见 [41, 43, 47, 68]。因此，机器学习团队会失去关于特征本质与含义，以及建模决策所依据的假设和推理的重要知识 [25, 41],；如果机器学习从业者需要重新审视这些决策，以遏制有害的模型行为，这一点尤其成问题。统计模型编写中也存在类似情况，因此有人呼吁开发高级库，以“[表达] 概念假设，从而为模型实现提供初始基础”，并开展双向概念建模，以“[共同编写] 概念模型与模型实现” [29]。我们的研究旨在机器学习模型开发领域直接践行这一愿景：让机器学习从业者能够通过基于高层次概念的抽象来实例化模型，并将其衔接至函数式模型实现。

2.3 早期阶段的机器学习模型设计

交互式机器学习（IML）领域长期以来一直关注如何让人类参与模型开发过程。针对构建模型过程缓慢且费力的问题，交互式机器学习领域的研究表明，通过支持交互式模型训练与改进，用户能够更快地开发模型，并使模型始终符合自己的目标 [1, 15, 19, 48]。在大多数交互式机器学习系统中，交互采取带标签示例和示范的形式，例如 CueFlik 中的正面和负面图像示例 [18] ，以及 Wekinator 中配对的输入手势和输出动作 [17]。

在交互式机器学习的基础上，交互式机器教学（IMT）相关工作强调以教学作为隐喻，聚焦人类作为教师而非仅仅作为标签提供者的角色 [52, 57]。通过这一隐喻，交互式机器教学凸显了新的挑战，例如需要超越数据标注、更加丰富的教学策略。沿着这一方向开展的一项绿野仙踪研究探索了用户如何向模型传授知识；该工作发现，用户会按照概念、概念之间的关系，以及组合和应用这些概念的规则来分解知识 [44]。然而，由于无法实际实例化和评估其概念，用户难以形成关于机器学习模型如何工作的心智模型，这使得他们尤其难以知道如何表述微妙、抽象知识形式。

尽管模型草图绘制与交互式机器学习和交互式机器教学的许多动机相同，但其目标略有不同——探索模糊且尚未明确的问题表述中的模型设计空间（问题表述）

范式	重点	范式转变	改进机器学习模型的策略	设计需求
交互式机器学习	交互式	机器学习模型不必孤立存在—— 通过人与交互 并提供反馈，我们 能够实现用户对齐，并快速构建模型 创作	更好的反馈回路	速度
交互式机器教学	教学	机器学习不只是学习——我们还需要关注 关于教学：哪些人正在开发机器学习模型 以及他们如何分享自己的知识	更好的教学	表达性
模型草图绘制	草图绘制	机器学习模型不必是高保真的—— 事实上，低保真、短暂的 模型更有助于 帮助我们探索模型设计的 多重宇宙	更好地探索设计空间	细节极简、速度快、例—— 表现力

表 1：交互式机器学习和交互式机器教学提出的关键思想，以及模型草图绘制的关键思想总结。这三种范式相辅相成，但强调了不同的视角和策略。

而非执行明确的建模构想（模型实现）。这意味着，除了交互式机器学习和交互式机器教学所要求的速度和表达性之外，模型草图绘制还需要对模型决策逻辑的最简表示，以促进有效的设计空间探索（表 1）。为了支持模型草图绘制，我们需要提供一种机制，让用户能够直接迭代控制模型行为的决策原语。因此，我们不能使用将输入不透明地映射到输出的黑箱方法，例如基于示例的交互式机器学习方法和基于提示的模型创作。概念是一种理想的表示方式，既能支持对决策策略进行迭代，也符合人们偏好的知识分解与表达方式 [44]。

草图绘制这一隐喻也启发了我们如何实例化这些基于概念的表示。以往的交互式机器学习和交互式机器教学研究探索了通过示例、示范、预定义特征和形式知识库来实现概念实例化 [4, 17, 18, 52]。然而，由于草图旨在探索一个事先未知的广阔模型设计空间，用户需要能够灵活地创建任意概念，因此我们不能依赖静态特征或知识库。草图绘制还面临一项额外挑战：将用户的注意力转移到不同任务上（例如，切换到标注示例、开发特征流水线或训练新的辅助模型）会带来很高的成本，因为这可能会使用户脱离草图绘制的思维状态，陷入技术隧道效应。因此，我们采用基于零样本提示的建模方法 [28,63]，以灵活而快速地实例化概念，同时不转移用户的注意力。

3 模型草图绘制

在本节中，我们介绍模型草图绘制框架，该框架帮助机器学习从业者在机器学习模型开发的早期阶段超越技术实现，探索概念性模型设计空间。

3.1 模型草图绘制框架

模型草图绘制旨在促进高层描述与任何特定底层实现之间的转译和解释工作 [27, 45]。这项工作为实现某一高层建模目标的众多可行策略勾勒出路径，而每种策略都会固化一套人类价值观和伦理影响，并将其编码进模型中。我们通过引入概念这一中间表示，弥合高层建模目标与底层实现之间的差距——概念是人类能够理解、且与决策任务相关的想法或因素。通过凸显目标与实现之间这一转译步骤，我们旨在为机器学习从业者的推理提供支撑，并帮助他们更有效地勾勒和探索并行解决路径。

3.1.1 数据前提条件。 我们只需要少量表格形式的数据（例如电子表格、逗号分隔值、Pandas 数据框），其中每一行代表一个数据示例，每一列代表一个数据字段。这些数据字段可以包含文本、图像或数值数据。我们建议数据集至少包含 40 至 50 行，以辅助模型开发过程中的意义建构，同时也允许划分训练数据和测试数据。如果用户提供真实标签，我们的系统可以帮助围绕模型准确率进行意义建构；不过，此类标签并非必需。

3.1.2 概念：构建决策的功能性构件。 概念是我们将草图模型这一更大、更具雄心的任务分解为更易于管理且人类可理解的组成部分的方式。我们通过将数据示例映射为一个分数，把概念转化为功能性组件；该分数表明该示例与指定概念的相关性，或其体现该概念的程度（图 3）。例如，“色彩缤纷”这一概念可以为一组图像生成分数，以大致表明它们有多么色彩缤纷；“脏话”这一概念可以为一组文本评论生成分数，以估计其中脏话的程度。因此，概念与特征提取器具有相似性，但它们并不局限于较低层次的客观属性，而且所需的生成时间和精力要少得多，因为创建一个新概念只需几秒钟。创建后，概念可以应用于任何任意的数据示例，无论其是否已见过。

我们基于零样本建模的进展来实例化概念。零样本建模描述的是一种无需任何训练示例即可执行任务的机器学习建模方法。要创建一个概念，用户需要提供一个输入概念词（简短的词语或短语）以及其数据集中的一个数据字段（列名称）。随后，我们的框架会构建一个零样本模型，为数据集中的每个项目输出该概念的数值分数。对于我们的概念，我们实现了两种方法：1）通过使用大语言模型（LLM）提示进行零样本文本分类，使用

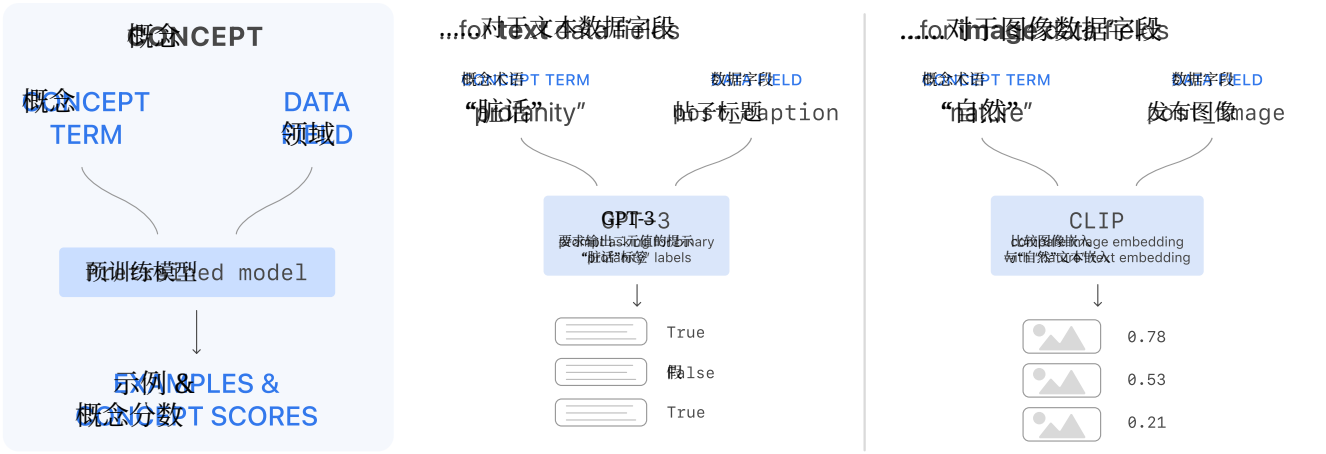

图 3：用户可以将人类可理解的概念转化为功能构件。一个概念以概念术语和数据字段作为输入。随后，预训练零样本模型会为每个示例生成一个分数，表明该示例与指定概念的相关程度。我们会根据数据字段的类型调整底层方法。对于文本，我们使用 GPT-3 提示进行二元标签预测。对于图像，我们使用 CLIP 嵌入进行连续分数预测。

GPT-3；2）通过使用基于 CLIP 的预训练图像转文本模型，依据嵌入相似度进行零样本图像分类。我们将在第 3.2.2 节中更深入地介绍我们的技术方法。虽然我们在工作中使用了这两种零样本方法，但我们的模型草图框架兼容任何当前或未来的零样本建模方法。

3.1.3 草图模型：将概念组合起来，以构想决策结果。我们框架的第二个组成部分涉及创建草图模型，即聚合由概念产生的信号的轻量级机器学习模型（图4）。草图模型旨在完成与其要模拟的完整模型相同的任务，因此它期望接收相同的输入，并产生相同类别的输出。

一旦用户创建了一组与决策任务相关的组成概念，他们便可选择希望组合到草图模型中的概念。我们提供了一组初始聚合器，可将概念分数结合起来生成单一的草图模型分数：线性回归、逻辑回归、决策树分类、随机森林分类和多层感知机（MLP）分类。我们并不要求使用复杂的深度学习模型来整合概念，而是借鉴了先前的研究；该研究表明，具有少量人类可解释特征的简单线性模型可以实现超越人类决策者并媲美复杂预测模型的性能 [30]。不过，我们并不局限于这组初始聚合器。在创建草图模型时，我们使用训练数据集中提供的真实标签来训练这些聚合器；此后每次运行草图模型时，都会在新的输入数据上评估训练好的模型。如果数据集没有标签，用户可以选择使用线性回归聚合器为概念提供手动权重。

借助草图模型这一抽象层，机器学习从业者可以集中精力尝试不同的概念组合，创建能够采集不同类别的决策价值的草图。例如，在编写仇恨言论检测模型时，机器学习从业者可以尝试侧重于种族相关或性别相关仇恨言论的草图，也可以尝试明确考虑交叉性的方法。他们可以试用相对宽松的草图，只添加与煽动暴力言论相关的概念；也可以尝试更严格地处理脏话，在草图中添加脏话和侮辱性称呼相关的概念。草图模型最具风格化的形式完全由概念构成，但也能兼容通过自定义特征提取器或现有机器学习模型生成的、更传统的特征。

3.1.4 从草图模型到生产模型。虽然草图模型最初是在模型开发早期阶段用于思考和迭代的工具，但我们设想，它们也可以为完整生产模型的设计提供参考。

首先，完整模型设计的一个关键部分是决定如何将模型拆分成更小的部分。成功的生产机器学习系统有一个显著特点：无论是在信息流排序和内容审核等用例中，还是在 Facebook、LinkedIn 和 Reddit 等平台上，它们都由较小、目标更明确的机器学习模型组成，这些模型的输出会合成为一个分数，通常采用加权线性组合 [5, 12, 14]。换句话说，机器学习从业者在行业中开发的子模型是我们所说的概念的更复杂版本，但通常仍可归结为人类能够理解的决策因素，并通过聚合过程合成为一个预测。

其次，据一些估算，模型开发中相当大的一部分（高达 80%）实际上都耗费在数据清理和整理上，而机器学习从业者也积极参与数据的塑造、整理和设计 [43]。模型草图绘制可以帮助完善 数据需求，通过梳理模型必须推理的概念，明确任务需要哪类数据。通过在不同数据字段（例如，

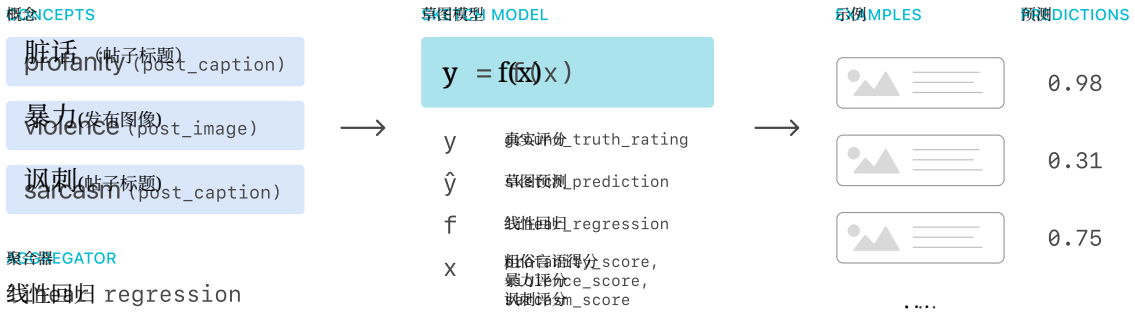

图4：草图模型将多个概念结合起来，为决策任务创建一个整体模型。首先，用户从自己编写的概念中进行选择，并决定采用何种聚合方法。我们使用概念分数、指定的聚合器以及用户的真实评分（例如，对帖子仇恨程度的评分）训练草图模型。随后，草图模型可以为任意示例生成预测（例如，估计帖子仇恨程度的分数）。

不同的文本或图像字段）上测试概念，机器学习从业者或许能迅速了解哪些形式的数据最能体现不同概念，并最终确定建模任务需要哪些数据字段以及哪些类型的数据示例。

3.2 ModelSketchBook 工具

为了实现我们的模型草图设计框架，我们推出了 ModelSketchBook API，[1] 这是一个 Python 软件包，可让机器学习从业者创建自己的概念和草图。

3.2.1 环境与设置。我们的 API 可以在任何 Python 环境中加载，但针对 Jupyter 或 Colab 等计算笔记本进行了优化。我们选择让 API 设计适配笔记本环境，因为许多机器学习从业者已经在此开展大量探索性建模工作 [54, 68]。

```
导入模型_草图_本，简称 msb

# 设置草图本sb =msb.create_模型_草图本（目标="检测社交媒体上的仇恨表情包。"，数据集={ # 传入要使用的数据集"训练集": df_训练集，"测试集": df_测试集，}，模式 ={ # 指定数据字段的类型"文本": InputType.Text，"图像_网址": InputType.Image，"总体_评分": InputType.GroundTruth，}，)
```

用户在笔记本中将数据集加载为 Pandas 数据框并导入 ModelSketchBook 软件包后，会创建一个草图本对象。草图本包含用户在本次会话中将要编写的所有数据集、概念和草图，并保存模型预测、性能指标以及概念和草图历史记录等其他元数据。

3.2.2 创建概念：幕后机制。如前所述，我们专门支持由 GPT-3 驱动的文本概念，以及由 CLIP 驱动的图像概念。为了实现零样本文本分类，我们编写提示词，列出一批供模型分类的示例，并要求模型为每个示例判定一个二元标签（"<概念术语>"或"不是 <概念术语>"）。GPT-3 等大语言模型旨在根据提供的输入提示词续写文本，因此，模型给出的合理续写能大致反映出基于模型已有知识、可能合理的二元标签。下面给出一个示例提示词。

```
判断以下评论是"鼓舞人心"还是"并不鼓舞人心"。1. 你能做到！！
2. 唉，你真是最差劲的。3. 我太累了 ……

评论结果：
```

随后，GPT-3 模型会返回类似这样的文本补全响应，为示例提供标签。我们会解析并对该 GPT-3 响应进行后处理，以获取二元标签。

```
1. 有激励作用 2. 没有激
励作用 3. 没有激励作用
```

为了执行零样本图像分类，我们使用 CLIP（对比式语言-图像预训练）[51]，这是一种在（图像、文本）对上训练的神经网络，可将图像和文本表示在共享的嵌入空间中。给定一个包含图像的数据字段和一个概念术语，我们使用 CLIP 在该共享的图像-文本嵌入空间中对图像和概念术语进行编码。然后，我们计算并返回嵌入后的图像与概念术语之间的余弦相似度。因此，输出分数近似反映了给定图像与指定概念术语的相关程度。我们还允许用户对这些相似度分数进行归一化和重新校准，使其符合用户对概念的内部理解；此外，用户还可以应用自定义阈值，将连续的概念分数二值化。

3.2.3 使用应用程序接口创建概念。设置好草图本后，用户便可自由创作概念。虽然底层也提供了原始应用程序接口函数，但为方便起见，我们通过 ipywidgets 提供用户界面元素。[2] 当用户调用创建

[1] https://github.com/斯坦福HCI/ModelSketchBook

[2] http://ipywidgets.readthedocs.io

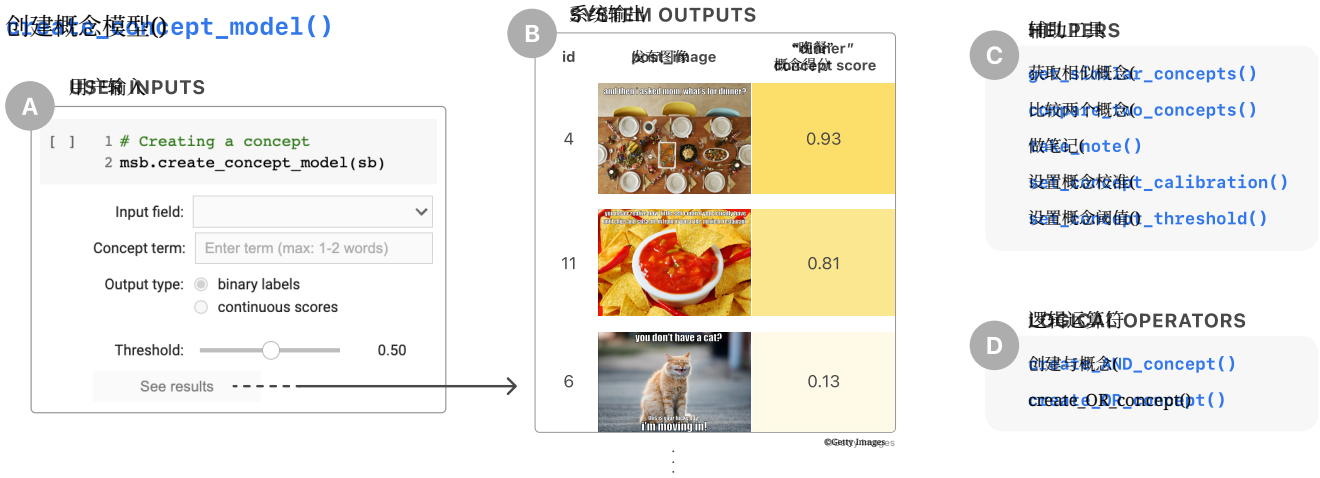

图 5：在计算笔记本中使用 ModelSketchBook 创建概念。(A) create_concept_model() 函数接受用户输入的数据输入字段、概念术语及其他设置。(B) 随后，应用程序接口会输出一个数据框，显示按概念得分排序的所有训练示例。(C) 我们提供用于微调、理解现有概念以及头脑风暴新概念的辅助函数。(D) 我们还支持逻辑与和或运算符，以创建复合概念。

_concept_model() 时，会出现一组小部件，供用户指定输入字段、概念术语，以及其他细节，例如概念得分的输出类型（图 5A）。

用户输入这些详细信息后，即可运行该单元格；几秒钟后，应用程序接口会返回数据框可视化，其中按顺序显示训练数据集中的示例及其概念得分（图 5B）。用户可以浏览概念评分结果，判断模型对该概念的理解是否与自己的理解充分对齐。如果不对齐，可以尝试其他概念术语或其他输入字段，以进一步改善对齐效果。应用程序接口还提供其他功能，用户可以使用自定义阈值将连续得分二值化，将连续得分自动归一化到 0-1 的得分范围，或根据指定的最低和最高得分值对连续得分进行自定义校准。

我们提供了一些函数，帮助用户在头脑风暴时评估不同概念之间的权衡。头脑风暴辅助函数会为给定词语提供同义词或反义词，在用户希望完善概念时激发灵感。同时，概念比较辅助函数允许用户指定概念术语，并比较这些术语与输入数据及真实评分之间的相关性。为了支持用户进行意义建构，我们还提供笔记功能，让他们记录有关概念和草图的草稿笔记，以供日后参考。

最后，为了进一步控制和完善概念，我们允许用户使用逻辑与和或运算符将现有概念连接起来，以创建复合概念。通过调用 create_AND_concept()，用户可以对任意数量的二元概念执行逻辑与运算（同样，也可以使用 create_OR_concept() 函数，通过或关系连接任意概念）。这使用户能够表达更复杂的概念，例如一个可以以多种方式呈现的单一概念，或一个同时需要图像和文本信号的多模态概念。

3.2.4 使用应用程序接口创建草图。用户编写若干概念后，就可以调用 create_sketch_model() 创建草图模型，通过聚合这些概念来构建模型。该函数会显示小部件，供用户选择概念和草图模型所采用的聚合方法（默认：线性回归）（图 6A）。

执行单元格后，应用程序接口会创建新的草图模型，并返回数据框可视化，其中显示训练示例的所有输入字段、所有相关概念分数，以及草图模型生成的总体预测（图 6B）。此视图还包含真实评分，以及草图预测与这些真实评分之间的差异。用户可以选择按不同列排序，以帮助理解模型输出。草图模型输出还会根据训练示例的真实标签，显示任务在回归和分类设定下的传统性能指标；我们展示平均绝对误差（MAE）、分类准确率、F1 分数、精确率和召回率。

最后，应用程序接口提供了帮助用户评估其草图模型的函数。compare_sketches() 函数通过在类似于图 11 的图表中并排显示性能指标，使用户能够比较指定草图的性能指标。test_sketch() 函数允许用户在带标签或不带标签的验证集（如果提供）上测试草图模型，从而了解其对更广泛示例的泛化能力。用户选择草图模型和数据集后，该函数会生成类似于图 6B 的数据框，但仅在存在真实值评分和误差时将其包括在内。有关实例化草图本、概念和草图模型所需的三个函数调用的概述，请参阅附录 A。

3.2.5 实现。 ModelSketchBook API 使用 Python 编写。我们使用 OpenAI 的 API 调用 GPT-3，执行基于零样本提示的文本分类，并使用开发时最先进水平的 GPT-3 变体 text-davinci-002

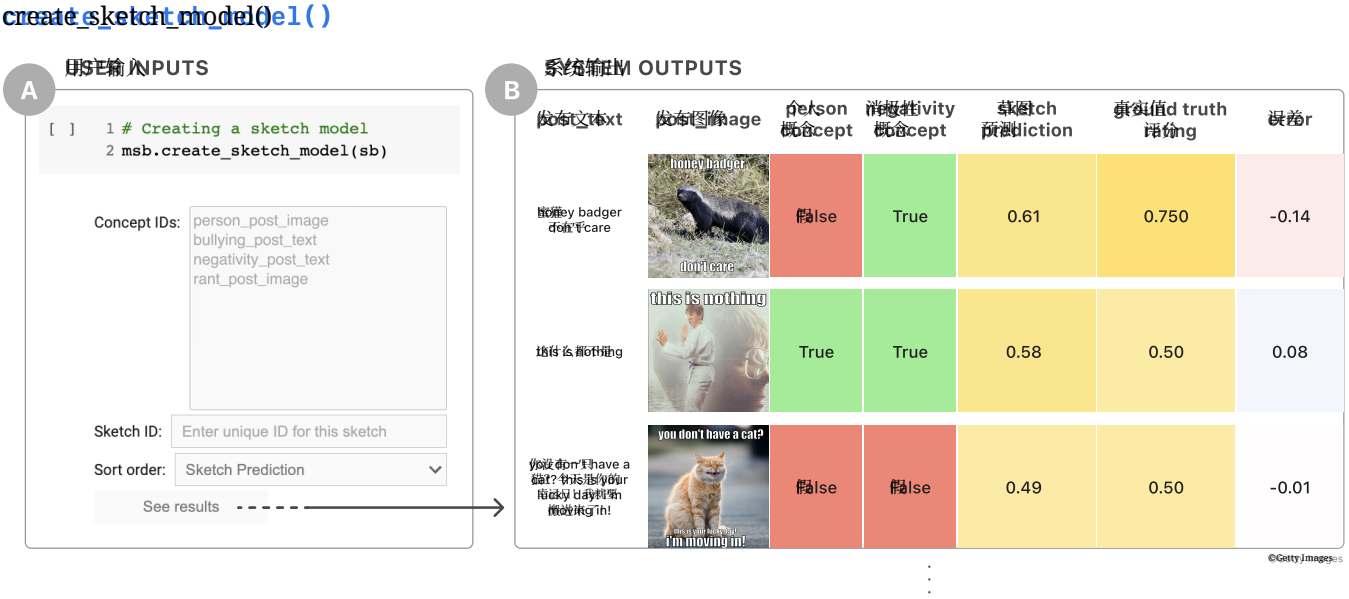

图 6：在计算笔记本中使用 ModelSketchBook 编写草图。(A) create_sketch_model() 函数接受用户输入，以指定要聚合的概念；还可使用可选参数设置聚合方法和排序列。(B) 随后，应用程序接口会输出一个数据框，其中显示了按指定列排序的所有训练示例。此视图展示了所有输入字段、所有概念分数以及草图模型的预测结果。此外，它还显示真实评分，以及草图预测与真实评分之间的误差，以帮助进行误差分析。

模型。我们使用 OpenCLIP [26]，（CLIP 的开源版本）执行零样本图文对齐；我们使用在 laion400m_e32 数据集上预训练的视觉Transformer（ViT）模型（ViT-B-32-quickgelu）。我们使用 Scikit-learn [49] 模型实现来构建草图模型聚合器。所有训练和推理均在运行笔记本的本地机器上进行（即，如果用户运行的是本地 Juypter 笔记本，则在用户自己的机器上进行；如果运行的是 Colab 笔记本，则在 Google云服务器上进行）。在计算笔记本环境中加载时，输入界面元素使用 ipywidgets 软件包进行渲染。输出可视化结果使用带有自定义 CSS 样式的 Pandas 数据框进行渲染。

4 演示案例研究

为了展示模型草绘方法可应用于多种机器学习任务，我们提供了由我们编写的示例案例研究，以演示推荐、无监督聚类、搜索和偏见审计应用（图 2；详情见附录表 5）。

旅行规划——个性化推荐。 在旅行规划的示例中，针对某位特定用户，我们的目标是推荐最符合其偏好的 Airbnb 房源。在这里，我们绘制的草图可以捕捉用户对 Airbnb 的品味要点：根据图像，我们表达了偏好的审美风格（“干净的”“明亮”“设计师风格”）和负面标记（“过时”“丑陋”）；根据文本描述，我们传达了社区的理想属性（“氛围愉悦”“中心地段”“时尚”）。草图绘制过程帮助我们发现建模设置中的不足，例如客户评价在客观评估清洁度等因素方面可能具有的价值，以及在旅行应用场景中，邻近程度和特定实际地理位置信息的重要性。

创意灵感——无监督嵌入与聚类。 模型草图绘制也可用于无监督学习任务，让用户创建定制的类草图嵌入空间，以执行项目相似度、聚类和类比检索任务。基于博物馆艺术品数据集，我们针对与大胆视觉风格（“色彩缤纷”“超现实”“奇幻”“抽象的”）、社会正义（“社会正义主题”“种族主题”“女权主义”“酷儿主题”）和政治运动（“政治主题”“革命主题”“劳工运动主题”）相关的文本和图像字段，创作了草图模型。随后，我们将这些草图模型应用于新的艺术作品，根据概念得分检索它们的“嵌入”。接着，我们可以对艺术作品进行聚类，并通过呈现不同体裁中的类比艺术作品来获取创意灵感：选择一件绘画体裁的艺术作品，并指定雕塑作为目标体裁后，我们便能从雕塑体裁中检索出与原始艺术作品最相似的类比作品。

政治候选人调查——个性化搜索。 我们的方法也能支持与搜索相关的任务。在这个示例中，我们协助一位用户在选举前调查政治候选人，以了解哪些候选人在他们最关心的议题上与其立场最为一致。我们参考了第三方候选人信息来源，包括政纲、政治声明和

演讲，并撰写了能够捕捉相关议题的草图，涵盖环境（“气候变化”“环境”“环保”）、枪支管制（“枪支管制”“反对枪支”）以及其他公民权利（“平等权利”“平等”“投票权”“支持堕胎权”）。模型草图绘制尤其有助于反复探索候选人细致入微的立场，而这些立场并未在标准投票指南中被明确列为核心议题。

餐厅评论——审稿人偏见审计。我们还可以将模型草图绘制应用于审计任务：从决策标准出发，反向推导并揭示人类或算法决策者的潜在权重。在这个例子中，我们协助一位美食评论员评估自身可能存在的偏差。根据这位评论员自我报告的决策因素（按重要性排序：味道、质量、呈现、氛围、服务和价格），我们创建了相应的概念。随后，我们编写草图模型来汇总这些概念，并以该评论员对餐厅的总体评分作为真实值。我们的模型推断出“美味”和“质量”这两个概念具有较高的正系数，这与该评论员最重要的两个决策因素相符。然而，模型推断出“呈现”和“氛围”这两个概念具有较高的负系数，而它们对应的是该评论员接下来最重要的因素。这让评论员意识到，这些因素对其总体评分的影响比其原本预期的要小。

5 评估

我们的研究旨在评估让经验丰富的机器学习从业者采用模型草图绘制方法完成现实的模型创作任务所产生的影响。由于我们的目标是改变机器学习从业者进行模型创作时所采用的认知框架，因此评估围绕以下主要研究问题展开：

研究问题1： 参与模型草图绘制会如何影响机器学习从业者在模型开发过程中的认知重点？ 采用这种方法后，他们是否能够成功地从技术导向和实现导向的思维转向更高层次、以概念为中心的思维？

RQ2： 模型草图工作流程会为机器学习模型和机器学习从业者带来哪些结果？ 当机器学习从业者转向更高层次的思考时，他们会生成什么类型的模型？他们如何预判建模差距，并在模型设计中考虑价值权衡？

为回答这些研究问题，我们针对具有丰富机器学习经验的参与者设计了一项现场评估，让他们使用 ModelSketchBook API 创作模型，以检测发布在社交平台上的仇恨表情包。

5.1 研究设计

5.1.1 研究形式。我们的评估包括一个为时一小时的研究环节。我们提前向参与者发送了一项简短的 10 分钟标注任务，请他们在研究环节开始前完成，以便在模型创作任务中使用他们个人给出的真实评分。我们请他们为 40 个示例提供标签；其中 20 个示例在研究环节中用作训练集，另外 20 个留作测试集。在一小时的研究环节中，前 15 至 20 分钟用于知情同意、设置 Colab 笔记本，以及 ModelSketchBook API 的教程和视频演示。随后，参与者有大约 30 分钟的时间，在 Colab 笔记本中使用 ModelSketchBook API 完成模型创作任务。最后的 10 至 15 分钟则用于简短访谈，了解他们的体验，并填写研究结束后的调查问卷。完整的书面作答题目、调查问题和访谈脚本收录于附录 D-F 节。附录 C 节介绍了我们的定性分析方法。

5.1.2 仇恨表情包检测任务。 我们的数据集源自 Meta AI 于2020年发起的仇恨表情包挑战赛 [31]，旨在引起人们对识别多模态仇恨言论这一复杂任务的关注。该数据集包含超过10,000个基于现实世界中的表情包的多模态表情包示例（结合文本和图像）。我们为参与者的标注任务提供的评分方案为：0.00 = 无害（保留）； 0.25 = 略有问题（可能存在争议，但保留）； 0.50 = 阈值（问题严重到需要移除）； 0.75 = 有问题（移除）； 1.00 = 问题严重（移除）。我们的表述不同于最初的仇恨表情包挑战赛；我们围绕内容移除而非仇恨言论来设定标签，使建模任务围绕用户偏好的决策逻辑展开，而不是围绕集中式定义展开。此外，我们将数据集筛选为在原始数据集中被标注为“非仇恨性”的示例，以减少参与者接触令人不适内容的机会，并倾向于纳入更加模棱两可、处于灰色地带的案例，在这些案例中，用户可能会进行不同的价值权衡。

我们选择这一任务进行用户评估，是因为我们希望在一个困难且贴近现实的情境中评估模型草图绘制，而这一任务代表了机器学习从业者正在积极应对的一项开放挑战。此外，我们希望任务具有一定的主观性，不存在明确的对错答案，从而为参与者提供足够广阔的模型设计空间。围绕多模态仇恨言论和内容审核所作的决策，通常涉及艰难的价值权衡。

5.1.3 任务前后书面回答。 在介绍模型草图绘制之前，我们请参与者根据当前掌握的知识，写下他们针对仇恨表情包任务的建模方法。参与者观看教程和演示后、开始使用 ModelSketchBook API 之前，我们请他们集思广益，列出至少三个可能会用于该任务的初始概念。任务阶段结束时，我们再次请参与者描述他们计划采用的仇恨表情包任务建模方法；他们可以假设自己能够使用类似 ModelSketchBook API 的工具，但并非必须使用此类工具。我们还请参与者写下他们在参与该任务后获得的任何学习心得，并指出在自己创作的草图模型中，哪一个是他们认为的“最佳草图”。

5.2 参与者招募

我们招募了在机器学习和 IPython 笔记本方面具有丰富经验的参与者。我们向计算机科学和人工智能相关群组的大学邮件列表发送电子邮件进行招募，并将招募材料发送给在人工智能行业工作的个人联系人，请他们与自己的社交网络分享。我们从回答“相当

多"或"非常多"（5 点李克特量表中最高的两个选项）的参与者中进行筛选，以表示他们使用 IPython 笔记本的经验以及机器学习方面的经验。共有 17 名用户参与了我们的研究（参与者的人口统计信息请参见附录 G 节）。参与者完成为时一小时的研究环节后，我们向每人提供了一张价值 60 美元的亚马逊礼品卡作为报酬。所有研究均通过视频会议远程进行。

6 结果

使用我们的 ModelSketchBook API，参与者成功地为仇恨表情包任务创建了平均 12.2（SD=4.7）个概念，以及 4.1（SD =2.0）种基于这些概念构建的不同草图模型。图7展示了参与者创建的概念和草图模型样本，表3汇总了所有参与者的概念。

研究中创建的概念里，53.6% 为文本概念，32.4% 为图像概念，14.0% 为涉及逻辑运算符的复合概念。每位参与者平均创建了 6.53 个文本概念（SD=3.00）、4.19 个图像概念（SD =2.14）和 2.07 个复合概念（SD=1.53）。在复合概念中，62.1% 使用与运算符，37.9% 使用或运算符；79.3% 仅组合文本概念，其余 20.7% 则同时组合文本和图像概念。研究参与者共创建了 136 个不同的概念。就个人而言，每位参与者平均创建了 6.29 个独有概念（SD=3.77），这些概念没有其他参与者创建过。

我们的建模任务旨在让参与者拥有较大的主观判断空间，从而提供足够广阔的模型设计空间。事实确实如此：参与者在如何为给定数据集标注标签方面存在显著差异。综合各位参与者和各个数据示例来看，标签的平均标准差（取值范围为 0–1）为 0.23。对于这些相同分数的二值化版本，这意味着 75% 的示例中至少有一位参与者的判断不同于多数标签，37.5% 的示例中则有四分之一或更多的参与者持不同意见。



图7：参与者草图示例。为便于说明，我们展示了两位研究参与者的第一版草图、最佳草图和其他概念。参与者探索了广泛的概念，并在最初的草图模型基础上进行了拓展。

6.1 模型创作认知过程的变化

针对研究问题1，模型草图绘制帮助机器学习从业者将思维从技术性、面向实现的因素转向更高层次的模型概念性、面向设计的因素

开发。

6.1.1 计划建模方法的变化。在任务前的建模计划中，参与者往往着重关注实现细节，例如具体的建模方法或架构（如Transformer模型、卷积神经网络、监督模型、线性分类器）、机器学习中的惯例（如训练-测试集划分和性能指标），以及以数据为中心的特征提取计划（汇总见表2）。例如，P13 将其方法描述为"采用两种模型：[一种]基于情感，另一种是简单的卷积神经网络，然后用一个简单的全连接层将这两个模型结合起来"；P8 则计划"创建数据分布均匀的测试集/验证集/训练集"，然后"训练某个模型，并检查它在留出测试集上的表现。"

然而，接触模型草图绘制后，参与者表示，他们会花时间进行头脑风暴和筛选

与决策任务相关的概念，以及他们会通过检查草图模型输出对其概念进行迭代。P15描述道："花几天时间试验各种概念，便能产生一个相当有效且更易于调试的模型。"参与者还进一步提到，他们希望考虑不同的伤害子类，并说明了如何将基于概念的方法与更传统的建模方法相结合。例如，尽管P8仍倾向于使用更复杂的非线性模型，但他们希望利用ModelSketchBook中的概念来增强其数据点。参与者还描述了在模型草图绘制期间其思维方式的变化，我们在附录第I节中对此进行了定性总结。

研究阶段	建模计划主题	引言
模型创建前草图绘制	特定的模型架构或技术	“我会使用卷积神经网络模型 [...] 可能会结合 Transformer与卷积神经网络一起追踪长期依赖关系，并帮助将文本数据置于上下文中。”（P1），“两种模型：一种基于情感，另一种是简单的卷积神经网络，而且然后用一个简单的全连接层将两个模型结合起来。” (P13)
	在训练集和验证集上进行评估	“确保将数据集划分为训练集、测试集和验证切片。尝试使用多种建模技术，同时在验证集上进行评估。” (P3)
	标准性能指标	“衡量各种指标（准确率、偏差、方差等）。”（P1），“确定用于评估任何训练模型的合适指标。可选择准确率、F1分数，以及其他指标，例如 AUC 或 AuPR，具体取决于相对标签权重。”（P3）
	构建源自数据的低级特征	“可以使用帖子和用户交互中的特征。可以使用一般的用户信息以及帖子信息。”（P4），“构建能够体现直觉的特征，用于表示与移除相关的事物；例如主题元数据、语言选择、点赞/点踩/争议性。” (P5)
在模型草图绘制之后	花时间进行头脑风暴并筛选概念	“我可能会谨慎地采用特征工程，以构建一个可解释模型。花几天时间琢磨概念，可能会构建出一个相当有效的模型，而该模型会更容易调试。”（P15）
	通过检查草图模型输出，迭代概念草图模型输出	“[我会]查看聚合器提供的错误标注（以及正确标注）示例，并进行头脑集思广益，寻找可能导致这一结果的更具体特征。”（P5）
	根据以下内容构思概念不同的伤害类别	“我认为，过滤仇恨言论（种族/性别/宗教）和政治言论会是不错的第一步。有些图像露骨/暴力的帖子可能无法被这些类别捕捉到，所以我会专门为这些内容添加更多过滤器。”（P2）
	结合方法 a 基于概念的 with 传统方法	“还可以将这些标签作为数据点的注释，用于更复杂的模型（非线性）。看来，保留原始数据仍然有好处，所以不会只依赖 CLIP/GPT 注释。”（P8），“我可能还是会手动收集数据，然后而不是使用预训练模型并从头搭建流水线，我会使用 ModelSketchBook API。”（P11）

表2：参与者在研究前后的建模计划。参与者的计划从较偏技术、侧重实现，转向更注重流程和价值，并选择在未来的建模中采用模型草图绘制。

6.1.2 概念演化。 了解参与者思维方式的一个窗口，是他们在使用ModelSketchBook期间概念随时间演化的方式。我们观察到，参与者平均出现了8.5次（SD=2.7）概念创新，我们将其定义为不同概念构想的数量；其中平均有5.0次（SD=2.8）概念创新不在参与者最初头脑风暴得出的列表中。概念演化往往源于参与者发现其草图模型中的误差：P7指出，一个结合种族主义和性概念的草图模型在幽默帖子上出现了假阳性，这促使其探索幽默相关概念。这些进展往往相互衔接：P4最初从一个效果良好的脏话概念开始，随后他们头脑风暴了仇恨的其他维度，并决定探索与种族主义相关的概念；接着，他们发现该数据集中与种族和族裔相关的仇恨往往针对穆斯林，因此产生了探索与伊斯兰恐惧症相关概念的想法。参与者平均表现出3.2次（SD=2.3）概念执行，我们将其定义为：所创建的概念中，属于同一概念构想延续的概念数量（例如，从基于图像的种族主义概念切换到基于文本的种族主义概念，或将概念术语从“种族主义”改为“种族”）。参与者更经常改变模态，但很少修改概念术语。随后，参与者平均对4.2个（SD=2.6）概念构想进行了概念放弃；我们将这一现象定义为：最终未被用于用户自行指定的“最佳”草图模型的概念构想数量。这种放弃也是草图绘制中有价值的一部分——识别并舍弃那些看起来行不通的路径。例如，P14 尝试了一个“有害”的基于图像的概念，却发现这会产生噪声结果；他们认为这是因为该概念过于模糊，与数据集中的图像关联不紧密，于是决定探索与图像关联更紧密、更具体的概念。因此，参与者在创作概念时，似乎会采用多种不同的生成、迭代和反思模式。

6.2 建模成果

关于 RQ2，我们观察到，模型开发过程中关注重点的高层次认知转变为模型本身带来了切实益处——从草图模型所探索的概念广度，到参与者在建模设置中发现的不足。

6.2.1 识别建模差距。模型草图绘制帮助参与者获得了模型本身以外的见解。参与者发现，数据集所提供的样本未能充分呈现若干重要的伤害类别，例如恐同和恐跨；他们还注意到，与当下的网络迷因相比，数据集中的迷因显得不够真实，因此基于这些数据构建的模型可能会因分布偏移而在实际应用中失效。参与者还发现了标注方法的局限性，例如类别不平衡和评分不一致。他们也加深了对任务本身的理解：参与者发现，对于这些迷因而言，文本往往更有意义，因为图像通常相当无害；他们也开始认识到，在某些情况下，多模态

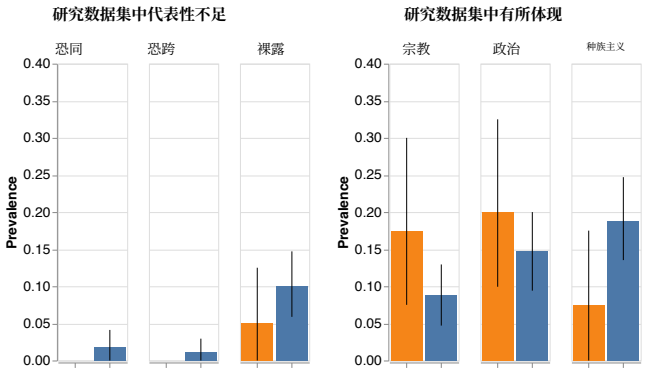

图 8：参与者根据使用研究数据集进行草图绘制的过程指出代表性不足的概念（左），在完整数据集中通常也代表性不足。原本流行的概念（右）在完整数据集中仍占有相当大的比例。

主题	参与者概念示例	数量
族裔、种族	种族主义、种族主义的、种族	34
危险、暴力	暴力、恐怖主义、伤害	28
情绪	吐槽、可怕、愤怒	27
歧视	仇恨、仇恨言论、歧视	22
政治	政治、政治、假新闻	16
冒犯性内容	冒犯性内容、粗俗、淫秽的	15
宗教	宗教、宗教、穆斯林	14
性别	性别歧视的、性别歧视、性别	10
色情内容	裸露、性相关、性	9
幽默	有趣、幽默、笑话	8
网络挑衅	网络挑衅、迷因、低俗迷因	7
交叉性	肤色与性别，伊斯兰恐惧症 | 性别歧视	7

表3：参与者针对由3名或以上参与者涵盖的主题所提出的概念汇总。我们列出了每个主题最常见的概念术语，以及属于该主题的参与者概念总数。

方法可能带来更多益处。更详细的总结请见附录 H 节。

研究结束后，我们寻求定量证据，以支持参与者提出的建模差距。

数据集代表性：首先，我们探究参与者认为在其 40 个样本的数据集中代表性不足的概念，在完整的 8.5k 个样本训练数据集中是否确实代表性不足。由于完整数据集没有针对这些概念进行标注，我们团队的一名成员对 40 个样本的研究数据集以及完整训练集中的随机抽取的 2% 样本（170 个示例）进行了人工标注，标注的三个概念由参与者指出，分别是“恐同”“恐跨”和“裸露”。作为对照，我们还对三个由参与者提出、流行率较高的概念进行了人工标注——“宗教”“政治”和“种族主义”；这些概念在研究数据集中的流行率分别为 17.5%、20.0% 和 7.5%。我们发现，在研究数据集中流行率分别为 0%、0% 和 5% 的恐同、恐跨和裸露，在完整训练数据集中也确实代表性不足，其流行率分别为 1.8%、1.2% 和 10%（图 8）。总体而言，我们发现，用于模型草图绘制的小数据集中的概念代表性，可以为完整数据集中的代表性提供有用的预警信号。

类别不平衡：接着，我们分析了参与者标签中类别不平衡的程度。在研究过程中，有四位参与者提出了类别不平衡的问题；这些参与者的类别不平衡程度确实都很高，而且都偏向负标签（平均负标签占=79.3%）。其中两位参与者的类别不平衡程度在所有参与者中最高（负标签占比分别为 92.5% 和 87.5 %）。此外，除提出这一问题的参与者外，还有三位参与者的负标签占比超过 75%。因此，类别不平衡似乎确实是这项任务中普遍存在的挑战。

我们很欣慰地发现，参与者能够在收集大规模数据集或实现完整模型之前，迅速识别出潜在的建模差距。

6.2.2 概念多样性。 参与者尝试了广泛多样的概念术语——涵盖与种族和族裔相关的概念，以及与政治、宗教和暴力相关的概念（按主题汇总于表3）。值得注意的是，所有参与者都提出了他们在最初头脑风暴中未曾考虑过的概念；提出概念并绘制草图的过程帮助他们发现了值得纳入的重要且相关的概念。将参与者各自的概念映射到我们所确定的12个共同主题上后，我们发现，参与者创建的概念平均涵盖其中的5.65个主题（SD=1.90）。反过来，每个主题平均由8.0名参与者覆盖（SD =3.10）。因此，我们看到，参与者能够在草图绘制环节探索多样化的概念集合，而且许多概念主题都由不同参与者从多个角度进行了研究（图9）。

6.2.3 所需时间。最后，我们注意到，参与者只用了完成该任务的完整模型所需时间的一小部分，就实现了这些积极的建模成果。参与者对于执行其任务前建模计划所需时间的估计差异极大，从半天到数周或数月不等。然而，即使考虑最乐观的数小时估计（由17名参与者中的3名提供），我们的模型草绘方法仍能在不到30分钟内交付功能模型原型。除了模型草图绘制的速度之外，我们还在附录J节中详细记录了参与者体验成果。

6.3 概念性能基准测试

虽然我们在案例研究演示和实地评估中发现，概念足以支持用户的草图建模过程，但我们仍希望将零样本概念分数与人工标注进行比较，以评估其准确率。我们选取了研究参与者中实例化频率最高的概念，这些概念均由至少四名参与者独立使用。其中包括四个基于文本的概念（“宗教”“政治”“种族主义”和“暴力”）以及三个基于图像的概念

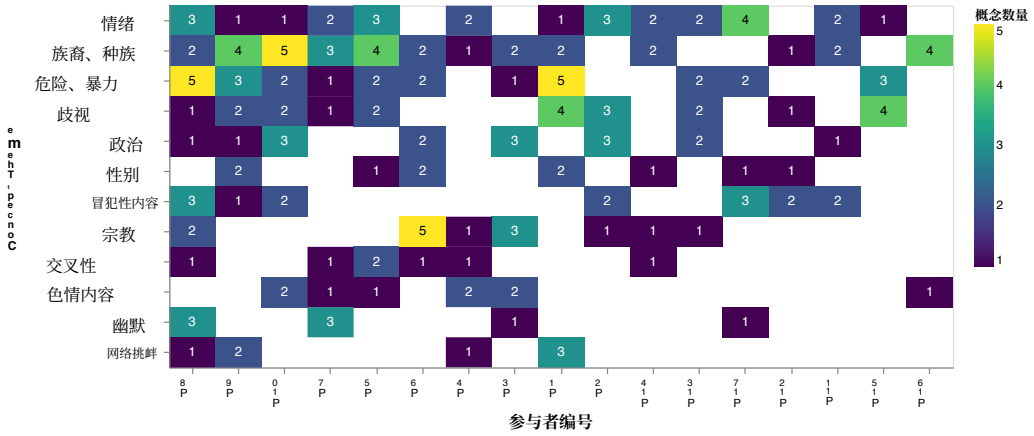

图9：从参与者（x 轴）到概念主题（y 轴）的热力图矩阵显示，参与者与主题之间存在显著差异，这表明每位参与者都探索了多种不同的概念。概念主题按独立参与者数量从上到下排序，参与者则按独立概念主题数量从左到右排序。

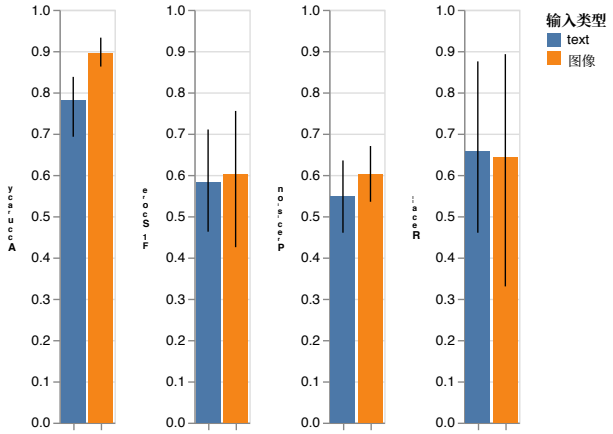

图10：根据人工标注，参与者的文本和图像概念的性能较高。

	概念类型	Text	图像
标注指标	平均流行率	19.4%	13.3%
	Fleiss κ系数	0.68	0.60
概念性能	准确率	0.78	0.90
	F1 分数	0.58	0.60
	召回率	0.66	0.64
	精确率	0.55	0.60

表 4：性能估计基于在数据集中足够常见且评定者间一致性较高的概念。

（“种族主义”“裸露”和“暴力”）。我们团队的四名成员分别对研究中向参与者展示的全部 40 个项目中的这些概念进行了评分。根据这些人工标签，参与者数据集中的基于图像的概念流行率过低，无法得出可靠估计，平均流行率为 4.2%（“裸露”概念的示例数为 0）。[3] 因此，我们对研究参与者中实例化频率次高的基于图像的概念（“穆斯林”“历史”“政治”和“动物”）重复了相同的流程。

如图10和表4所示，文本和图像概念的性能水平相近。两组概念的评定者间一致性都相当高（以 Fleiss κ系数衡量），且根据这些人工标签，其流行率也足够高。对于基于文本的概念，我们发现平均准确率为 0.78（SD=0.09），平均 F1 分数为 0.58（SD=0.14），召回率高于精确率（召回率=0.66，精确率=0.55）。对于基于图像的概念，我们发现平均准确率为 0.90（SD=0.04），平均 F1 分数为 0.60（SD=0.20），同样是召回率高于精确率（召回率=0.64，精确率=0.60）。因此，对于一种无需收集数据、标注概念或花费时间训练的方法，零样本概念分数为用户提供了一个合理的起点，以便进一步迭代和完善。

6.4 草图模型性能基准测试

虽然模型草图绘制的目标并非生成完整规模模型，但我们希望了解草图模型能否达到与完整规模模型相当的合理功能水平。

6.4.1 研究参与者的草图模型。首先，我们考察了研究参与者编写的草图模型。我们将所有参与者自行确定的“最佳草图模型”与两种

[3]这与参与者在第6.2.1节提出的同一问题有关：仇恨表情包数据集需要更充分地呈现概念。

种基线进行了比较：1）第一版草图，即参与者使用 ModelSketchBook API 编写的第一份草图；2）零样本，即使用单个 GPT-3 提示来预测每个示例是否“仇恨”的基线。对于每种模型变体，我们根据参与者自己的真实标签，在其包含 20 个示例的测试集上衡量性能。我们发现，参与者的最佳草图在分类指标（F1、精确率、召回率）上优于两种基线，但在任务的回归版本中，最佳草图的性能略低于基线；该结果以平均绝对误差（MAE）衡量（图 11）。回顾用户积极迭代所依据的 20 个示例训练集，参与者的最佳草图在所有指标上都明显优于第一版草图和零样本基线。

这些结果表明，参与者能够通过迭代改进草图模型，使其与建模目标更加对齐；与细粒度回归任务相比，草图模型在分类任务上的泛化能力可能更强。鉴于草图模型的性能指标从训练集到测试集有所下降，参与者在开发草图模型时可能对训练数据集过拟合了。与此同时，正如预期，零样本基线模型在训练集和测试集上的性能整体较差，但保持一致；这恰恰是因为该模型没有机会对训练集过拟合。这两种极端情况符合机器学习中经典的偏差—方差权衡。我们的草图模型结果与先前文献相呼应，显示出交互式机器学习中普遍存在的过拟合倾向这一挑战 [11, 64], ，尤其是因为我们没有采取措施明确引导用户避免过拟合行为。最后，我们注意到，正如第6.2.1节所述，由于类别不平衡，这项任务的分类指标整体较低。许多参与者的标签明显偏向负类（“非仇恨性”），因此最终生成的模型总是预测负类（这种行为会使召回率、精确率和 F1 分数均为 0.0）。

6.4.2 探究任务难度对草图模型性能的影响。随后，我们进一步探究了在类别不平衡不那么严重、训练示例数量远超 20 个的情况下，草图模型的性能。我们准备了另外三个任务：IMDB 电影评论情感分析任务（最简单：每个示例都有评分，且类别均衡）[60]；评论有害性任务（较难：由多位标注者评分，且类别分布偏斜）[35]；以及原创仇恨表情包任务（最难：类别均衡，但任务需要细致辨析且为多模态任务）[31]。对于每项任务，我们都将草图模型与已发表的最先进水平（SOTA）模型进行比较，采用与 SOTA 模型相同的核心性能指标以及训练集／测试集数据。有关基准测试和我们的草图模型的更多详情，请参见附录 K 节。对于 IMDB 任务，SOTA 模型报告的准确率为 0.93，而我们的草图模型达到 0.91，仅略低一些。在评论有害性任务中（得分范围为 0–4），SOTA 模型报告的平均绝对误差为 0.90，而我们的草图模型的平均绝对误差为 1.39。在仇恨表情包任务中，SOTA 模型报告的曲线下面积为 0.845；该挑战赛刚发布时，SOTA 结果为 0.714；我们的草图模型达到 0.584。

因此，模型草图绘制在简单和中等难度任务上表现出色或尚可，但在第三项难度显著更高的任务上，性能则较弱。

6.5 失效模式

虽然我们的模型草绘方法实现了改变模型开发者认知框架的目标，并为他们开发的模型带来了诸多益处，但我们也观察到了几种失效模式。首先，尽管参与者提出了一些富有潜力的概念构想，但他们并不总能通过我们基于零样本模型的方法将其实现。参与者往往高度依赖他们尝试的第一个概念术语所返回的结果。如果该术语未能返回令人满意的结果，他们通常会放弃这一概念构想，因为他们认为模型不够复杂，无法理解该概念。通过提供脚手架式支持，帮助参与者判断究竟是模型缺乏对某一概念的理解，还是仅仅需要使用不同的术语，将会对他们有所帮助。

其次，我们发现，一些参与者会根据他们认为零样本模型能够理解的内容，大幅调整其概念术语，尤其倾向于使用具体、明确的概念术语。我们观察到，其中一些参与者最终拓展了思路，并发现更宽泛的术语也能取得良好效果。然而，问题仍然存在：由于我们的概念依赖于预训练模型，这些模型可能成为限制因素（第 7.2.1 节将进一步讨论）。用户可能会将自己的思考局限在预训练模型已经擅长的领域，而不是更具创意地思考哪些决策因素可能具有价值。

最后，我们注意到，由于我们的方法强调采用“小数据”方式来帮助用户保持专注，一些参与者忽略了构建稳健的通用模型这一总体目标。参与者有时会陷入过拟合，仔细审视草图模型出错的案例，并用与更宏观的决策任务无关、非常具体的概念来修补误差。若能借助工具评估草图模型的泛化能力，参与者将会受益。我们认为，在草图绘制情境中，精选数据集值得这样的取舍，因为它能鼓励机器学习从业者避免细致的数据分析，转而关注宏观概念。

7 讨论

我们介绍了模型草图绘制，将这一方法应用于一款面向机器学习从业者设计的工具，并证明该方法能够成功地将注意力转向高层次的模型设计问题，而这些问题对于在早期阶段做出正确决策至关重要。在此，我们将讨论模型草图绘制带来的机遇，以及其局限性和未来研究方向。

7.1 模型草图绘制的机遇

我们设想，将草图绘制实践引入机器学习将带来多种令人振奋的影响。

7.1.1 扩展受草图启发的机器学习模型开发方法的适用范围。尽管我们的工作支持使用逻辑运算符表示复合概念，以增加多模态性并表达更细致入微的概念，但这些概念编写起来更加繁琐，而且可能会施加过强的基于规则的僵化性。未来的工作可以探索多层次的模型草图绘制，以表达复杂概念。例如，可以

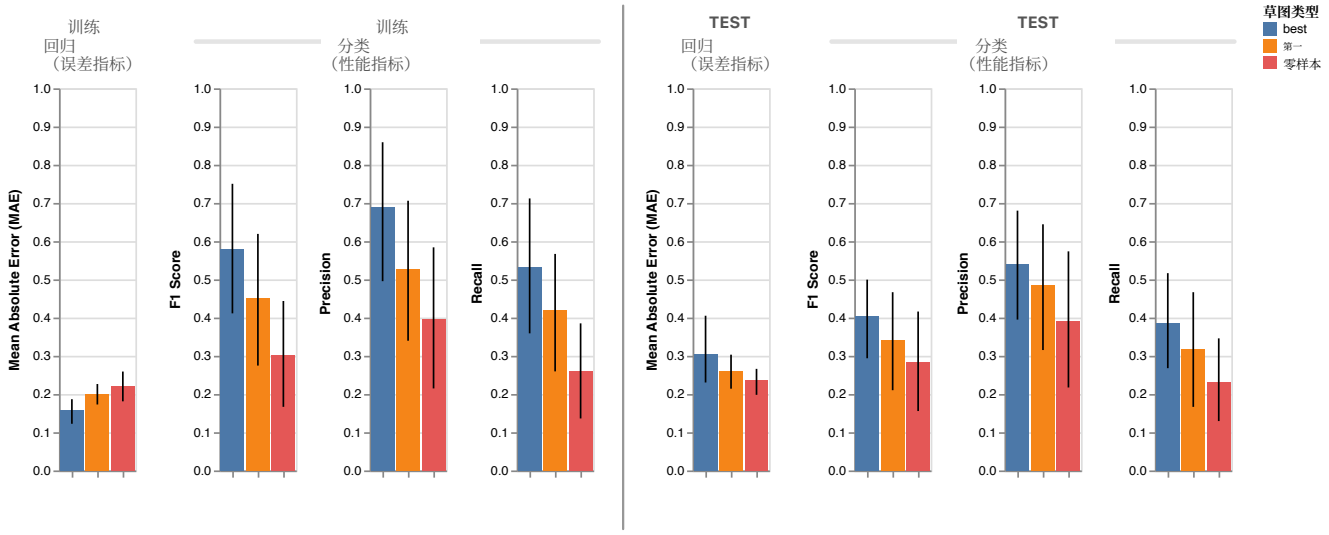

图 11：对于训练集（左），与第一版草图和零样本模型这两个基线相比，我们观察到参与者自行指定的"最佳"草图在各项指标（平均绝对误差、F1 分数、精确率和召回率）上始终表现更好。对于测试集（右），我们观察到"最佳"草图的分类性能更好（F1 分数、精确率和召回率），但平均绝对误差更高。

通过嵌套层级的模型草图绘制来表达复杂概念。比如，种族主义这一概念极其多面；种族主义有多种形式，也可能以许多不同方式被体验。因此，模型设计者不能仅仅接受给定的笼统定义，而必须谨慎明确其特定问题表述和模型部署环境所要求的、关于种族主义的细致概念。通过嵌套草图模型，机器学习从业者可以根据需要反复"放大"，逐步细化概念和子概念。

7.1.2 将草图绘制应用于模型相关任务。参与者在评估中发现的问题不仅存在于模型本身，也存在于更广泛的建模生态系统中，包括问题表述、数据和标注方法。与建模工作相比，数据工作往往被忽视且不受重视 [24, 55]。模型草绘方法可以扩展，以协助完成其中一些任务。例如，在数据标注任务中，一个主要挑战是通过迭代逐步完善标注策略 [34]。模型草图绘制可以迭代模拟不同标注策略的结果，揭示需要进一步细化概念的领域，使机器学习从业者在投入高昂的人工标注工作之前就能发现这些问题。

7.1.3 赋能非技术用户参与机器学习原型设计。目前，缺乏机器学习经验的利益相关者（无论是最终用户、领域专家、设计师还是产品经理）几乎无法主导机器学习模型设计决策，而且在很大程度上只能依赖机器学习专家向他们传达的信息 [37, 50]。模型草图绘制可以让非机器学习专家构建功能性草图模型，并将其领域专长融入其中。鉴于机器学习从业者通常缺乏特定部署领域的专业知识，这将是极有价值的贡献。我们在用户评估中的几位参与者明确提出，模型草图绘制对非技术用户会非常有用。

7.2 局限性与未来工作

7.2.1 超越预训练模型和零样本建模的局限。我们采用基于预训练模型的零样本建模技术方案，其局限之一在于，这些模型可能不够灵活，无法支持广泛的建模任务。由于 GPT-3 和 CLIP 模型是在互联网文本和图像数据上训练的，它们对世界的认知有限，无法为训练数据中代表性不足的领域提供高质量的概念分数。未来的研究可以探索其他专用预训练模型或替代零样本建模方法的有效性，以克服这些不足。

虽然为了保留类草图交互，我们将研究限制在零样本概念实例化，但未来的工作可以采用混合方法，以便通过示例来规范概念。正如概念泛化基准测试所示，零样本概念在草图绘制任务中的平均性能尚可，但准确性并不稳定，可能需要更多的人工控制。对于值得投入额外时间进行完善的情况，用户或许可以通过少样本建模或对预训练模型进行微调，利用带注释的示例来优化概念。

7.2.2 确保正确使用和解读草图模型。我们必须谨慎地向用户传达工具的预期用途。由于我们的方法连接了人类可理解的概念与模型预测，信息不足的用户可能会尝试将我们的方法用于可解释性：这是我们的工具并非为之设计的用途。我们的草绘方法以准确率换取速度和灵活性，旨在辅助早期模型设计探索。与此同时，可解释性用例对概念

和模型准确率的要求要高得多。模型草图绘制工具必须明确区分这些用途。通过限制数据集规模，并将工具定位于新颖的模型创作，而非对生产模型进行事后分析，我们也能引导用户避开这些不受支持的用例。

此外，与任何基于大语言模型及其他预训练模型构建的技术一样，草图模型也会受到这些模型中固有偏差的影响[36, 56]。机器学习从业者了解这些风险至关重要，因此我们鼓励采用模型草图绘制的其他人明确告知用户，这些底层模型可能会表现出有害偏差，并且这些偏差可能会体现在草图模型中。尽管这些偏差本身是负面的，但有一个积极的副作用：在模型开发过程早期遇到这些偏差，会促使机器学习从业者思考同类偏差可能如何影响他们的完整规模模型。早期遭遇此类失败可能促使模型开发者优先开展公平性相关工作；而我们的方法让他们能够尝试不同策略——无论是调整建模方法、上游数据整理，还是下游分数处理和偏差检查——以减轻他们遇到的有害偏差。

7.2.3 从草图到完整规模模型的衔接。草图模型是早期探索，而非模型的最终版本，但仍需为完整规模模型的行为提供有用信号。如第3.1.4节所述，草图模型可以通过指导如何拆解大型建模任务以及细化数据需求方面的决策，为生产模型的设计提供参考。我们针对最先进模型开展的技术基准测试表明，模型草图绘制在某些任务上能够取得相当不错的性能，但在难度更高的任务上可能难以泛化。即使性能指标达到上限，模型草图绘制过程仍有助于促进对高层次建模方法的反思和迭代。未来工作的重要方向包括支持模型草图与其生产部署之间的衔接，并为草图与完整规模性能指标之间的映射提供更有力的保证。

8 结论

在一个由高保真技术实现和工程投入主导的领域中，我们主张通过模型草图绘制，以低保真、快速且富有表现力的方式探索模型的决策逻辑。正如终端用户编程和面向人工智能系统的示范编程长期以来致力于让用户专注于系统应该做什么 [9, 10]，而不是如何实现，我们也希望让模型作者在面对问题表述任务时，专注于自己的模型应该推理哪些内容，而不是过早深入探究模型是如何实现的。我们不必等上几周或几个月，直到模型实现完成后才在真实世界中进行评估，并发现问题表述中编码了有害偏差；我们希望机器学习从业者能在短短几天内实例化多个高层次的模型构想，并与用户一起测试。如果他们发现自己的建模方法在多个方面容易产生偏差，就可以通过多轮草图模型试验来尝试消除这些偏差；只有在完成这一早期阶段的迭代之后，他们才会继续推进完整生产模型的实现。借助模型草图绘制，机器学习从业者可以摆脱被动修补实现问题的模式，转而主动探索富有创意的模型设计，从一开始就打造出更好的模型。

致谢

我们感谢匿名审稿人，以及 Ranjay Krishna、Matthew Jörke、Mitchell Gordon、Lindsay Popowski 和曹汉成对本文提出的宝贵反馈。本研究部分由 IBM 资助；IBM 是斯坦福以人为本人工智能研究院（HAI）的创始成员。Michelle S. Lam 获得了斯坦福以人为本人工智能研究所和斯坦福大学工程学院布朗媒体创新研究所的支持。

参考文献

[1] Saleema Amershi、Maya Cakmak、William Bradley Knox 和 Todd Kulesza。2014年。《权力归于人民：人类在交互式机器学习中的作用》。《人工智能杂志》 35, 4 2014 105 120 10 1609 35 4 2513（12月），–。[2] 2022 https://doi.org/。/aimag.v i。苹果公司。。Create ML。 https://developer.apple.com/machine-learning/create- ml/[3] Phil Aquilina。2022年。打造更好的版主工具。https://www.reddit.com/ r/ RedditEng/comments/uly8s4/building_better_moderator_tools/ [4] Michael Brooks、Saleema Amershi、Bongshin Lee、Steven M. Drucker、Ashish Kapoor 和 Patrice Simard。2015年。FeatureInsight：为文本分类中的错误驱动特征构思提供可视化支持。载于《2015年IEEE可视分析科学与技术会议（VAST）》。105–112。 https://doi.org/10.1109/VAST.2015. 7347637[5] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，1877–1901。https://proceedings. neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [6] Bill Buxton。2010年。《勾勒用户体验：让设计恰到好处》。Morgan Kaufmann。[7] Joseph Chee Chang、Saleema Amershi 和 Ece Kamar。2017年。Revolt：用于标注机器学习数据集的协作式众包。载于《2017年CHI计算机系统中的人为因素会议论文集》（美国科罗拉多州丹佛）《CHI '17》。计算机协会，美国纽约州纽约市， 2334 2346 10 1145 3025453 3026044 –。https://doi.org/。/。Kathy [8] 2006 Charmaz。。《建构扎根理论：定性分析实用指南》。Sage。[9] Allen Cypher。1991年。EAGER：通过示例编程完成重复性任务。载于《SIGCHI计算系统人为因素会议论文集》（美国路易斯安那州新奥尔良）《CHI '91》。计算机协会，美国纽约州纽约市， 33 39 10 1145 108844 108850 –。https://doi.org/。/。[10] Allen Cypher、Daniel C. Halbert、David Kurlander、Henry Lieberman、David Maulsby、Brad A. Myers 和 Alan Turransky（编）。1993年。《看我怎么做：示范式编程》。麻省理工学院出版社，美国马萨诸塞州剑桥。[11] Pedram Daee、Tomi Peltola、Aki Vehtari 和 Samuel Kaski。2018年。用于避免预测交互式知识获取中过拟合的用户建模。载于《第23届智能用户界面国际会议》（日本东京）《IUI '18》。计算机协会，美国纽约州纽约市，305–310。https: //doi.org/ 10 1145 3172944 3172989 。/。Siddharth [12] 2020 Dangi。。了解停留时间以改善 LinkedIn 信息流排序。 https://engineering.linkedin.com/blog/2020/understanding-feed-dwell- time[13] Steven P. Dow、Alana Glassco、Jonathan Kass、Melissa Schwarz、Daniel L. Schwartz 和 Scott R. Klemmer。2011年。并行原型设计带来更好的设计成果、更多的发散思维和更强的自我效能。《ACM人机交互汇刊》17卷4期，文章18（2011年12月），24页。 https://doi.org/10.1145/1879831. 1879836[14] Dean Eckles。2022年。算法透明度与算法排序效果评估。https://doi.org/ 10 31235 8 6 。/osf.io/c za [15] 2003 Jerry Alan Fails 和 Dan R. Olsen。。交互式机器学习。载于《第8届智能用户界面国际会议论文集》（迈阿密，

美国佛罗里达州）（IUI '03）。计算机协会，美国纽约州纽约市，39–45 10 1145 604045 604056 –。https://doi.org/。/。Daniel Fallman。2003。设计导向的人机交互。载于SIGCHI计算系统人为因素会议论文集（美国佛罗里达州劳德代尔堡）（CHI '03）。计算机协会，美国纽约州纽约市，225–232 10 1145 642611 642652 –。https://doi.org/。/。Rebecca Fiebrink、Dan Trueman 和 Perry R. Cook。2009。用于交互式即时机器学习的元乐器。载于国际音乐表达新界面会议（宾夕法尼亚州匹兹堡）（NIME '09）。[18] James Fogarty、Desney Tan、Ashish Kapoor 和 Simon Winder。2008年。CueFlik：图像搜索中的交互式概念学习。载于SIGCHI计算系统人为因素会议论文集（意大利佛罗伦萨）（CHI '08）。计算机协会，美国纽约州纽约市，29–38。https://doi. org/ 10 1145 1357054 1357061 。/。Jules Françoise、Baptiste Caramiaux 和 Téo Sanchez。2021年。Marcelle：编排交互式机器学习工作流与界面。载于第34届ACM用户界面软件与技术研讨会（美国线上活动）（UIST '21）。计算机协会，美国纽约州纽约市，39–53。https://doi.org/ 10 1145 3472749 3474734 。/。Vinod Goel。。《思想草图》。麻省理工学院出版社。[21] 谷歌。2022年。Teachable Machine。https://teachablemachine.withgoogle.com/ [22] Mitchell L. Gordon、Michelle S. Lam、Joon Sung Park、Kayur Patel、Jeff Hancock、Tatsunori Hashimoto 和 Michael S. Bernstein。2022年。陪审团学习：将不同意见融入机器学习模型。载于《2022年CHI计算系统中的人为因素会议论文集》（美国路易斯安那州新奥尔良）（CHI '22）。计算机协会，美国纽约州纽约市，文章 115, 19 10 1145 3491102 3502004 页。https://doi.org/。/。Björn Hartmann、Scott R. Klemmer、Michael Bernstein、Leith Abdulla、Brandon Burr、Avi Robinson-Mosher 和 Jennifer Gee。2006年。通过集成设计、测试与分析实现反思式实体原型设计。载于第19届ACM用户界面软件与技术年度研讨会论文集（瑞士蒙特勒）（UIST '06）。计算机协会，美国纽约州纽约市，299–308 10 1145 1166253 1166300 –。https://doi.org/。/。Fred Hohman、Kanit Wongsuphasawat、Mary Beth Kery 和 Kayur Patel。2020年。理解并可视化机器学习中的数据迭代。载于《2020年CHI计算机系统中的人为因素会议论文集》（美国夏威夷州檀香山）（CHI '20）。计算机协会，美国纽约州纽约市，1–13 10 1145 3313831 3376177 –。https://doi.org/。/。Youyang Hou 和 Dakuo Wang。。与非营利组织合作：公民数据黑客马拉松中的协作分析与中介角色。《ACM人机交互会议论文集》1卷，计算机支持的协同工作，文章53（2017年12月），16页。https://doi.org/10.1145/3134688[26] Gabriel Ilharco、Mitchell Wortsman、Ross Wightman、Cade Gordon、Nicholas Carlini、Rohan Taori、Achal Dave、Vaishaal Shankar、Hongseok Namkoong、John Miller、Hannaneh Hajishirzi、Ali Farhadi 和 Ludwig Schmidt。2021年。OpenCLIP。https://doi.org/10.5281/zenodo.5143773 如果您使用此软件，请按以下格式引用：[27] Abigail Z. Jacobs 和 Hanna Wallach。2021年。测量与公平性。载于《2021年ACM公平性、问责制与透明度会议论文集》（加拿大线上活动）（FAccT '21）。计算机协会，美国纽约州纽约市，375–385。https://doi.org/10.1145/3442188.3445901[28] Ellen Jiang、Kristen Olson、Edwin Toh、Alejandra Molina、Aaron Donsbach、Michael Terry 和 Carrie J Cai。2022年。PromptMaker：利用大语言模型进行基于提示的原型设计。载于《2022年CHI计算机系统中的人为因素会议扩展摘要集》（美国路易斯安那州新奥尔良）（CHI EA '22）。计算机协会，美国纽约州纽约市，文章35，8页。https://doi.org/ 10 1145 3491101 3503564 。/。Eunice Jun、Melissa Birchfield、Nicole De Moura、Jeffrey Heer 和 René Just。2022年。假设形式化：实证发现、软件局限与设计启示。《ACM人机交互汇刊》29卷1期，文章6（2022年1月），28页。https://doi.org/10.1145/3476980[30]Jongbin Jung、Connor Concannon、Ravi Shroff、Sharad Goel 和 Daniel G. Goldstein。2020年。指导专家分类的简单规则。《英国皇家统计学会杂志：A辑（社会统计学）》183卷3期（2020年），771–800。https://doi.org/10.1111/rssa.12576 arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12576[31] Douwe Kiela、Hamed Firooz、Aravind Mohan、Vedanuj Goswami、Amanpreet Singh、Pratik Ringshia 和 Davide Testuggine。2020年。《仇恨表情包挑战赛：检测多模态表情包中的仇恨言论》。载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，2611–2624。https://proceedings. neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf[32] Been Kim、Martin Wattenberg、Justin Gilmer、Carrie Cai、James Wexler、Fernanda Viegas 和 Rory sayres。2018年。超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。载于第35届国际机器学习会议论文集（《机器学习研究论文集》，第80卷），Jennifer Dy 和 Andreas Krause（编）。PMLR，2668–2677。https://proceedings.mlr.press/v80/kim18d.html

[33] Pang Wei Koh、Thao Nguyen、Yew Siang Tang、Stephen Mussmann、Emma Pier- son、Been Kim 和 Percy Liang。2020年。概念瓶颈模型。载于第*37*届国际机器学习会议论文集（机器学习研究论文集，第*119*卷），Hal Daumé III 和 Aarti Singh（编）。PMLR，5338–5348。https://proceedings.mlr.press/v119/koh20a.html [34] Todd Kulesza、Saleema Amershi、Rich Caruana、Danyel Fisher 和 Denis Charles。2014年。用于促进机器学习中概念演化的结构化标注。载于*SIGCHI*计算系统中的人为因素会议论文集（加拿大安大略省多伦多）（*CHI '14*）。计算机协会，纽约州纽约市，美国，3075–3084。https://doi.org/10.1145/2556288.2557238[35] Deepak Kumar、Patrick Gage Kelley、Sunny Consolvo、Joshua Mason、Elie Bursztein、Zakir Durumeric、Kurt Thomas 和 Michael Bailey。2021年。为多元视角设计有害内容分类。载于第十七届可用隐私与安全研讨会（*SOUPS 2021*）。USENIX协会，299–318。https://www.usenix.org/conference/soups2021/presentation/kumar [36] Keita Kurita、Nidhi Vyas、Ayush Pareek、Alan W Black 和 Yulia Tsvetkov。2019年。测量上下文化词表示中的偏差。载于首届自然语言处理中的性别偏见研讨会论文集。计算语言学协会，意大利佛罗伦萨，166–172。https://doi.org/10.18653/v1/W19-3823[37] Michelle S. Lam、Mitchell L. Gordon、Danaë Metaxa、Jeffrey T. Hancock、James A. Landay 和 Michael S. Bernstein。2022年。终端用户审计：赋能社区主导大规模有害算法行为调查的系统。*ACM*人机交互会议论文集6，CSCW2，文章512（2022年11月），34 10 1145 3555625 页。https://doi.org/ . / James A. Landay 和 Brad A. Myers。1995。面向用户界面设计早期阶段的交互式草图绘制。载于*SIGCHI*计算系统中的人为因素会议论文集（美国科罗拉多州丹佛）（*CHI '95*）。ACM出版社/Addison-Wesley出版公司，美国，43–50。https://doi.org/10.1145/ 223904.223910[39] Youn-Kyung Lim、Erik Stolterman 和 Josh Tenenberg。2008年。原型剖析：作为过滤器的原型，作为设计理念具象表现的原型。*ACM*人机交互汇刊15，2，文章7（2008年7月），27页。https://doi.org/ 10 1145 1375761 1375762 . / . Andrew L. Maas、Raymond E. Daly、Peter T. Pham、Dan Huang、Andrew Y. Ng 和 Christopher Potts。2011年。学习用于情感分析的词向量。载于计算语言学协会第*49*届年会：人类语言技术论文集。计算语言学协会，美国俄勒冈州波特兰，142–150 11 1015 – 。https://aclanthology.org/P - Yaoli Mao、Dakuo Wang、Michael Muller、Kush R. Varshney、Ioana Baldini、Casey Dugan 和 Aleksandra Mojsilović。2019年。数据科学家如何与领域专家开展科学合作：寻找正确答案还是提出正确问题？*ACM*人机交互会议论文集3，GROUP，文章237 2019 23 10 1145 3361118 （12月），页。https://doi.org/ . / Nora McDonald、Sarita Schoenebeck 和 Andrea Forte。2019。定性研究中的可靠性与评定者间信度：面向计算机支持的协同工作与人机交互实践的规范与指南。*ACM*人机交互会议论文集3，计算机支持的协同工作，文章72（11月2019 23 10 1145 3359174），页。https://doi.org/ . / Michael Muller、Ingrid Lange、Dakuo Wang、David Piorkowski、Jason Tsay、Q.Vera Liao、Casey Dugan 和 Thomas Erickson。2019年。数据科学工作者如何处理数据：发现、采集、整理、设计、创建。载于*2019*年*CHI*计算系统中的人为因素会议论文集（英国苏格兰格拉斯哥）（*CHI '19*）。计算机协会，纽约州纽约市，美国，1–15 10 1145 3290605 3300356 – 。https://doi.org/ . / . Felicia Ng、Jina Suh 和 Gonzalo Ramos。2020。理解并支持机器教学中的知识分解。载于*2020*年*ACM*交互系统设计会议论文集（荷兰埃因霍温）（*DIS '20*）。计算机协会，纽约州纽约市，美国，1183–1194。https://doi.org/ 10 1145 3357236 3395454 . / . Samir Passi 和 Solon Barocas。2019。问题表述与公平性。载于公平性、问责制与透明度会议论文集（美国佐治亚州亚特兰大）（*FAT* '19*）。计算机协会，纽约州纽约市，美国，39–48 10 1145 3287560 3287567 – 。https://doi.org/ . / . Samir Passi 和 Steven Jackson。2017。数据视觉：学习透过算法抽象来看事物。载于*2017*年*ACM*计算机支持的协同工作与社会计算会议论文集（美国俄勒冈州波特兰）（*CSCW '17*）。计算机协会，纽约州纽约市，美国，2436–2447。https://doi.org/ 10 1145 2998181 2998331 . / . Samir Passi 和 Steven J. Jackson。2018。数据科学中的信任：企业数据科学项目中的协作、转译与问责制。*ACM*人机交互会议论文集2，计算机支持的协同工作，文章136（2018年11月），28页。https://doi. org/ 10 1145 3274405 . / Kayur Patel、Naomi Bancroft、Steven M. Drucker、James Fogarty、Amy J. Ko 和 James Landay。2010年。Gestalt：为机器学习中的实现与分析提供综合支持。载于第*23*届*ACM*用户界面软件与技术年会论文集（美国纽约州纽约市）（*UIST '10*）。计算机协会，纽约州纽约市，美国，37–46。https: // doi.org/10.1145/1866029.1866038

[49] F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。2011年。Scikit-learn：Python中的机器学习。机器学习研究期刊12（2011年），2825–2830。[50] David Piorkowski、Soya Park、April Yi Wang、Dakuo Wang、Michael Muller 和 Felix Portnoy。2021年。人工智能开发者如何克服多学科团队中的沟通挑战：案例研究。*ACM*人机交互会议论文集5，CSCW1，文章131（2021年4月），25页。https://doi.org/10.1145/3449205

[51] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021年。《从自然语言监督中学习可迁移的视觉模型》。https://doi.org/ 10 48550 2103 00020 。/ARXIV。

[52] Gonzalo Ramos、Christopher Meek、Patrice Simard、Jina Suh 和 Soroush Ghorashi。2020年。《交互式机器教学：以人为中心的机器学习模型构建方法》。人机交互35, 5-6 (2020年), 413–451。[53] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022年。《使用潜在扩散模型进行高分辨率图像合成》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*）。10684–10695。[54] Adam Rule、Aurélien Tabard 和 James D. Hollan。2018年。《计算笔记本中的探索与解释》。载于*2018*年*CHI*人机交互系统中的人为因素会议论文集（加拿大魁北克省蒙特利尔），*CHI '18*。计算机协会，美国纽约州纽约市，1–12。https: //doi.org/10.1145/3173574.3173606

[55] Nithya Sambasivan、Shivani Kapania、Hannah Highfill、Diana Akrong、Praveen Paritosh 和 Lora M Aroyo。2021年。"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*CHI*计算机系统中的人为因素会议论文集（日本横滨）（*CHI '21*）。计算机协会，美国纽约州纽约市，文章39，15页。https://doi.org/10.1145/3411764.3445518

[56] Emily Sheng、Kai-Wei Chang、Premkumar Natarajan 和 Nanyun Peng。2019年。《女性担任保姆：论语言生成中的偏差》。载于*EMNLP/IJCNLP(1)*。3405–3410。https://doi.org/10.18653/v1/D19-1339[57] Patrice Simard、Saleema Amershi、Max Chickering、Alicia Edelman Pelton、Soroush Ghorashi、Chris Meek、Gonzalo Ramos、Jina Suh、Johan Verwey、Mo Wang 和 John Wernsing。2017年。机器教学：构建机器学习系统的新范式。技术报告 MSR-TR-2017- 26。 https://www.microsoft.com/en-us/research/publication/machine-teaching-new-paradigm-building-machine-learning-systems/ [58] Masaki Suwa 和 Barbara Tversky。1996年。建筑师从草图中看到了什么：对设计工具的启示。载于计算系统中的人为因素会议伴会（加拿大不列颠哥伦比亚省温哥华）*(CHI'96)*。计算机协会，纽约州纽约市，美国，191–192。https://doi.org/10. 1145/257089.257255[59] Maryam Tohidi、William Buxton、Ronald Baecker 和 Abigail Sellen。2006年。获得恰当的设计，并把设计做好。载于*SIGCHI*计算系统中的人为因素会议论文集（加拿大魁北克省蒙特利尔）*(CHI '06)*。计算机协会，纽约州纽约市，美国，1243–1252。https://doi.org/ 10 1145 1124772 1124960 。/。Leandro von Werra。。distilbert-imdb。https://huggingface.co/lvwerra/ distilbert-imdb[61] Miriam Walker、Leila Takayama 和 James A. Landay。2002年。高保真还是低保真，纸上还是电脑？测试网页原型时如何选择属性。人因与工效学学会年会论文集46, 5（2002年），661–665。https://doi.org/10.1177/154193120204600513 arXiv:https://doi.org/ 10 1177 154193120204600513 。/ Dakuo Wang、Justin D. Weisz、Michael Muller、Parikshit Ram、Werner Geyer、Casey Dugan、Yla Tausczik、Horst Samulowitz 和 Alexander Gray。2019年。数据科学中的人机协作：探索数据科学家对自动化人工智能的看法。*ACM*人机交互会议论文集3, 计算机支持的协同工作，文章 211（2019年11月），24 10 1145 3359313 页。https://doi.org/。/ Tongshuang Wu、Michael Terry 和 Carrie Jun Cai。2022年。AI链：通过串联大语言模型提示词，实现透明且可控的人机交互。载于*2022*年*CHI*计算系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）*(CHI '22)*。计算机协会，纽约州纽约市，美国，文章 385，共 22 页。https://doi.org/10.1145/3491102. 3517582 [64] Tongshuang Wu、Daniel S. Weld 和 Jeffrey Heer。2019年。交互式机器学习中的局部决策陷阱：对情感分析中特征选择的调查。*ACM*人机交互汇刊26, 4，文章 24（6月 2019 27 10 1145 3319616），页。https://doi.org/。/ Maria C. Yang。。一项关于原型、设计活动和设计成果的研究。设计研究 26, 6 2005 649 669 10 1016 2005 04 005 （），–。https://doi.org/。/j.destud.。。Qian Yang、Justin Cranshaw、Saleema Amershi、Shamsi T. Iqbal 和 Jaime Teevan。2019年。NLP草图设计：探索如何借助语言智能设计正确事物的案例研究。载于*2019*年*CHI*计算系统中的人为因素会议论文集（苏格兰格拉斯哥）*(CHI '19)*。计算机协会，纽约州纽约市，美国，1–12。https://doi.org/10.1145/ 3290605.3300415[67] Qian Yang、Aaron Steinfeld、Carolyn Rosé 和 John Zimmerman。2020年。重新审视人机交互为何以及如何在设计上独具挑战性。载于*2020*年*CHI*计算系统中的人为因素会议论文集（美国夏威夷州檀香山）*(CHI '20)*。计算机协会，纽约州纽约市，美国，1 13 10 1145 3313831 3376301 –。[68] https://doi.org/。/。Amy X. Zhang、Michael Muller 和 Dakuo Wang。。数据科学工作者如何协作？角色、工作流与工具。*ACM*人机交互会议论文集4, CSCW1，文章 22（2020年5月），23页。https://doi.org/10.1145/ 3392826[69] Ron Zhu。2020年。利用外部标签和域内预训练增强多模态Transformer：仇恨表情包挑战赛获胜方案。https://doi. org/10.48550/ARXIV.2012.08290

ModelSketchBook 流程概览

这里，我们总结用户在计算笔记本中设置其第一版草图模型时执行的三个命令。首先，用户按照第3.2.1节之前所示的方式设置其草图本。他们只需执行一次此操作即可加载数据，并且可以使用这一本草图本创建任意数量的概念和草图。

```
导入模型_草图_本，命名为 msb

# 设置草图本sb = msb.create_模型_草图本（目标="在社交媒体上检测仇恨表情包。"，数据集={ # 传入要使用的数据集"train": df_train，"test": df_test，}，模式 ={ # 指定数据字段的类型"text": InputType.Text，"img_url": InputType.Image，"overall_rating": InputType.GroundTruth，}）
```

接下来，他们可以通过调用如下所示的 create_concept_model() 函数，创建任意数量的概念。函数参数也可以通过用户界面控件以等效方式指定，该函数会返回一个单元格输出，其中包含概念结果的数据框可视化（图 5）。

```
msb. 创建_概念_模型（输入_字段="post_image"，# 用于输入的列概念_术语="卡通"，# 概念描述）
```

最后，用户可以通过调用如下所示的 create_sketch_model() 创建任意数量的草图；同样，用户也可以使用小部件指定相同的函数参数，而该函数会显示一个单元格输出，其中包含草图模型结果的数据框可视化（图 6）。

```
msb. 创建_草图_模型（概念=["卡通"、"吐槽"]，# 汇总草图_标识符="草图_版本_1"、# 草图名称）
```

B 模型草图绘制案例研究

表5总结了每个示例模型草图绘制应用的任务和数据集，并列出了相关概念和示例草图。

	旅行规划	创意灵感	政治候选人调查	评审者偏见审计
Task	协助旅行者选择一家 Airbnb，前往纽约市旅行。	帮助用户发掘创意联系，连接艺术家与艺术作品。	协助选民寻找一位候选人，参加2022年加利福尼亚州州长选举在选举中与他们立场一致的人就最重要的议题而言。	协助美食评论员检查他们在评论咖啡馆时自身的偏差以及在社交媒体上评论餐厅时的偏差。
Data	纽约市的 Airbnb 房源。包括每条列表的名称、图像、描述和社区概览。	博物馆艺术品。包括艺术家的名称、标题、类型、风格、体裁、媒介、地理区域、图像和每件艺术品的文本描述工作。	汇总的候选人信息 from 政治声明 and 由非营利组织 CalMatters 发布的演讲，一个无党派的第三方团体。	Instagram 美食评论。包括咖啡馆/餐厅的名称、帖子日期、帖子文本、点赞数、评论数和总体评分。
概念	时尚、中心地段、干净的、昏暗肮脏的、整洁的，设计的，设计师，明亮，过时、自然阳光、宽敞、丑陋、可做的事情、活动，氛围愉悦	色彩缤纷、超现实、奇幻、抽象的，现实主义的、政治主题、革命主题、劳工运动主题，社会正义主题、种族主题，性别主题、女权主义、酷儿主题式	环境、平等、枪支管制，生育权、投票权	美味，价格实惠，创意，Nice 氛围、优质服务、高品质，良好的呈现、美味、便宜，原创、不错的氛围、音响服务-vice
样本草图	美学：{干净的，明亮，设计师}	大胆：{色彩缤纷、超现实、奇幻、Abstract}	重大议题：{气候变化，枪支管制}	朴素概念：{美味、高 quality、价格实惠、创意、氛围愉悦，优质服务，良好的呈现}
	有趣的街区：{氛围愉悦，Central，时尚}	社会正义：{社会正义主题，Race 主题，女权主义，酷儿主题}	议题与观点：{环保，反对枪支，平等权利}	精炼概念：{美味，高质量，便宜，原创}
	现代美：{过时，丑陋，干净的}	政治运动：{政治主题，革命主题，劳工运动主题}	议题 & 详细观点：{生态-友好，反对枪支，支持堕胎权，平等-性，投票权}	经过 *refined* 的概念，扩展：{美味，高品质、便宜、原创、去-人的氛围、优质服务、良好的呈-现}

表5：多个模型草图绘制应用的总结。我们探讨了一项协助用户为旅行规划选择 Airbnb 的任务、一项旨在发掘艺术家与艺术作品之间创意联系的任务、一项旨在帮助选民发现自己与政治候选人之间对齐程度的任务，以及一项允许美食评论员调查他们最看重的因素和可能持有的偏差的任务。

C 定性分析

对于自由文本书面回答，我们旨在总结参与者所呈现的高层主题，因此编码并非最终产物，而是我们的处理过程 [42]。第一作者通过归纳分析总结参与者的模型创作体验。此过程的第一步是多次通读所有书面回答。随后，定性开放编码 [8] 过程以迭代方式进行，分为两个阶段：第一阶段是逐行回答编码，以贴近原始数据（例如，“注意到数据集中对暴力的最简表示”、“计划使用Transformer模型”或“标签之间存在不一致”）。第二阶段则将第一阶段的编码综合为更高层次的主题（例如，“数据代表性”或“标注方法”）。主题生成之后，我们依据这些主题对书面回答进行编码，以描述参与者计划采用的建模方法，以及他们从任务中获得的学习心得和体会。

D 任务前与任务后书面回答问题

D.1 任务前

参与者需要在主要的模型构思任务开始前，针对以下问题提供书面回答。

(1)（教程前）**请描述你的建模方法**。请用几句话描述你打算如何构建一个模型，以识别应从社交媒体平台上删除的帖子。

- 方法：
- 时间估计（您认为完成这一计划方法需要多长时间？）：

（2）（教程后）**初始概念头脑风暴**。请列出 3 个您认为可能与您的任务相关的概念。

D.2 任务后

参与者被要求在完成主要任务后立即对以下问题作出书面回答。

（1）**再次描述您的建模方法**。完成今天的活动后，请用几句话描述您将如何计划构建一个模型，用于检测社交媒体平台上应被删除的帖子。您可以假设自己能够使用类似 ModelSketchBook 的工具，但并非必须使用此类工具。

（2）**学习心得**。根据您今天的探索，您是否对任务、数据或建模目标有了新的认识？您是否发现数据或任务表述中存在任何不足？您是否希望进行任何更改来弥补这些不足？

(3) **选择你最喜欢的草图**。 请直接运行下面的单元格，查看你创建的所有草图。然后，在再下面的单元格中，输入你认为表现最好（或最符合你的建模目标）的草图编号

。最后，运行笔记本的最后一个单元格。

E 任务后调查问题

以下是参与者在完成主要任务后接受调查时需要回答的问题。

(1) 您对使用应用程序接口创建的概念模型有多满意？（单选：非常不满意、不满意、有些不满意、不满意也不满意、有些满意、满意、非常满意）(2) 您对使用应用程序接口创建的草图模型有多满意？（单选：非常不满意、不满意、有些不满意、不满意也不满意、有些满意、满意、非常满意）(3) 在尝试并迭代您的概念和草图的过程中，您觉得自己处理建模任务的方法在多大程度上发生了**变化**？（单选：没有变化、少许变化、有些变化、许多变化）(4) 在尝试并迭代您的概念和草图的过程中，您觉得自己处理建模任务的方法在多大程度上**有所改善**？（没有改善、略有改善、明显改善、大幅改善）(5) 您如何评价自己使用 Scikit-learn、PyTorch 或 TensorFlow 等工具创作机器学习模型时的**舒适程度**？（非常不舒服、不舒服、有些不舒服、不舒服也不舒服、有些舒服、舒服、非常舒服）(6) 您如何评价自己使用 ModelSketchBook API 时的**舒适程度**？（非常不舒服、不舒服、有些不舒服、不舒服也不舒服、有些舒服、舒服、非常舒服）(7) 您如何评价自己使用 Scikit-learn、PyTorch 或 TensorFlow 等工具创作机器学习模型时表达个人想法／意图的能力？（非常缺乏表现力、缺乏表现力、有些缺乏表现力、缺乏表现力也不富有表现力、有些富有表现力、富有表现力、非常富有表现力）(8) 您如何评价自己使用 ModelSketchBook API 时表达个人想法／意图的能力？（非常缺乏表现力、缺乏表现力、有些缺乏表现力、缺乏表现力也不富有表现力、有些富有表现力、富有表现力、非常富有表现力）(9) 您**喜欢**使用 ModelSketchBook API 创作模型的哪些方面？请描述该应用程序接口的优点。(10) 您**不喜欢**使用 ModelSketchBook API 创作模型的哪些方面？请描述该应用程序接口的缺点或您遇到的任何问题。(11)（可选）欢迎写下您可能有的其他评论、反馈或想法。

F 任务后访谈问题

以下是我们在主要任务结束后采访参与者时提出的问题。我们也根据需要提出了后续问题。

(1) **思维过程**。能否带我看一遍你的笔记本，并说明你在设计概念和草图模型时的思维过程？(2) **对比**。与以往创作其他模型的经历相比，这次模型创作过程有何不同？

- 你最喜欢它的哪一点？
- 你最不喜欢它的哪一点？(3) **学习心得**。 你从这个模型草图绘制过程中学到了什么？有什么收获，或对原有观点有何改变？

G 参与者人口统计信息

我们有3名女性、12名男性、1名非二元性别参与者，以及1名不愿透露性别的参与者；其中有14名亚裔参与者、2名白人参与者，以及1名自我认定为白人和西班牙裔的参与者。9名参与者年龄在18至24岁之间，8名在25至34岁之间。在最高学历方面，9名参与者获得了博士学位、硕士学位或其他专业学位，4名获得了四年制学位，3名接受过部分大学教育，1名高中毕业。有9名参与者自我认定为在行业工作的机器学习工程师或数据科学家，另有8名自我认定为在工作或研究中使用机器学习的学生。在行业从业者中，3人是初创公司的员工，6人是大型公司的员工（员工人数为 1000+ ）。参与者报告的机器学习工作经验平均为3.6年。

H 建模差距

表6总结了参与者识别出的建模差距主要类别，并附有参与者的直接引言。

I 参与者的思维过程

参与者认为，在进行模型草图绘制时，他们的思维模式发生了变化。许多参与者觉得，这种方法使他们摆脱了通常不得不处理的繁琐任务。P10将模型草图绘制描述为一种“不同于 [建模] 的思考方式”，相较于他们日常的建模工作——“更多是手动和技术性的，工作量更大，也更复杂”——他们喜欢模型草图绘制“会直接替你完成”。同样，P5表示，模型草图绘制“将机器学习中许多最烦人的部分外包了出去，例如创建用于生成特征的流水线”。

一些参与者认为，这种方法使他们能够以更长远的视角看待建模过程：P1表示，“这一过程有助于正确看待长期目标以及机器学习生命周期的结构。”另一些人将模型草图绘制过程描述为令人耳目一新的创意且艺术化，而他们通常的模型创作体验则更显程式化。正如P12所说，“[模型草图绘制]迫使我以一种平时不必采用的方式进行创造性思考。通常，我的工作相当模板化且重复。在这种情况下，模型的性能更多地受限于我自身的创造力和能力，因此我感觉自己对模型的表现拥有更多艺术指导。”

参与者也很欣赏模型草图绘制所带来的可解释性；与以往的模型创作体验相比，这让他们对模型行为有了更深入的理解。P14 分享道

类别	学习或收获	引言
Task	对于这项任务而言，文本总体上似乎比图像更重要	“图像可以从文本中分离出来，而在这项练习中，文本似乎更有用。”（P7）
	视具体情况而定，可能需要将这些模态结合起来，或专注于文本或图像	“文本与图像如何相互作用，这是一个非常重要的收获。你不能只考虑图像本身；也不能只考虑文本本身，重要的是它们共同产生的效果它们所创作的内容。”（P16），“反驳 [文本的重要性更高] 这一观点的例子，或许是图像血腥图片，因为帖子即使没有文本，也未必就有问题。”（P2）
	迷因很难理解和标注	“迷因很棘手，因为图像可能无害，但文本可能有害。”（P1），“迷因非常含糊，难以标注。其潜台词很难用简单的概念来捕捉。”（P6）
数据集	数据中需要更充分地涵盖各类伤害（恐同、恐跨、裸露）	“需要更多数据 [来采集] 各类有问题的帖子（不只是种族主义或性别歧视） [...] 恐同、恐跨等方面的数据不足。”（P1），“[...] 在这种情况下，裸露不是问题，但在测试时可能会是。”（P16）
	该数据集可能过于不现实且人为设计，难以匹配帖子在线分布	“这些数据似乎与通常在线发布的内容相差甚远，因此这些数据肯定不适合一般用途。”（P8）
标签	类别不平衡会对建模产生很大影响，并使其难以校准模型	“我了解到，类别不平衡对建模影响很大——例如，我自己的标注数据集似乎相对偏向于不移除帖子，因此，针对我认为真正具有冒犯性的内容，尝试 [try] 更具体的针对性术语，收效颇丰，比如鼓吹暴力。”（P5）
	他们自己的标签似乎前后不一致	“我的标签显然也缺乏足够的内部一致性，而且收集更多样本或许有助于平衡噪声。”（P7）
	提高移除分数粒度将很有价值而不仅仅是进行二值化	“一个可以改进的方面是提高任务保真度，实际使用这 5 个标签而不是将它们二值化。”（P3）

表 6：完成模型草图绘制任务后，参与者总结了对其建模方法的启示，并指出了数据集和标注方法中的不足。

“我通常做的很多事情都只是一个黑箱 [.....] 如果它不起作用，就会让人很沮丧”，但“有了另一种自底向上构建的选择，就成了一个相当不错的替代方案。”这种建模方法也让参与者能够从第一性原理出发思考决策。P16 描述道：“从人类层面的推理来说，我很喜欢把问题拆解成更小概念的想法。这似乎是教孩子时一种很直观的方式.....在我看来，这就是在构建一个更容易解释的模型。”

J 参与者体验结果

从对机器学习从业者的益处这一角度回答 RQ2 时，我们观察到参与者认为模型草绘方法既舒服又富有表现力。参与者对这种模型开发方法表现出极大的热情，表示他们“很享受以这种方式思考模型的体验”（P1），并认为该工具“未来对类似任务可能很有帮助”（P11），或者可以“展示某件事是可行的，并拓展想象力”（P12）。

J.1 与现有创作工具的比较

与使用 Scikit-learn、PyTorch 和 TensorFlow 等更常见的模型创作工具的体验相比，参与者使用 ModelSketchBook API 时感到的熟悉程度相近（图 12）。尽管这项 30 分钟的任务是他们首次接触我们的工具，而且所有参与者都拥有使用这些更传统工具进行模型创作的丰富经验。与此同时，我们发现，参与者认为使用 ModelSketchBook API 时的表现力高于使用这些同样常见的模型创作工具时的表现力。这些结果初步表明，我们的模型草绘方法有望成功被机器学习从业者采用，成为模型创作的一种轻量且富有表现力的入口。

J.2 对模型草图绘制体验的反思

在调查中，我们还探究了用户对其概念和草图模型的满意程度，并询问他们认为自己的建模方法随着时间推移发生了多大变化或改善。我们观察到，参与者对自己的概念和草图模型都非常满意。除两位参与者外，所有人对两种模型的满意度都评为“比较满意”“满意”或“非常满意”。除一位参与者外，所有人都表示，在不断迭代概念和草图的过程中，自己的建模方法发生了变化；绝大多数参与者（17 人中有 14 人）认为自己的建模方法“略有改善”“明显改善”或“显著改善”。

在访谈反馈和问卷开放式回答中，参与者对模型草绘方法表示热情（表 7）。他们最常提及的模型草图绘制优点包括：高度可解释、易于使用（尤其是无需编写代码）、能够提供快速反馈回路，以及促使他们采用一种不同的、更高层次的思考方式。与此同时，参与者指出的主要局限包括：笔记本环境常常变得杂乱；以示例为中心（而非以指标为中心）的可视化需要花更多时间来理解；此外，他们也希望在头脑风暴构思概念方面得到更多帮助。多位参与者对这样的机会感到兴奋：让缺乏机器学习知识的非技术用户和领域专家也能使用 ModelSketchBook，并发挥其领域专业知识。他们还建议

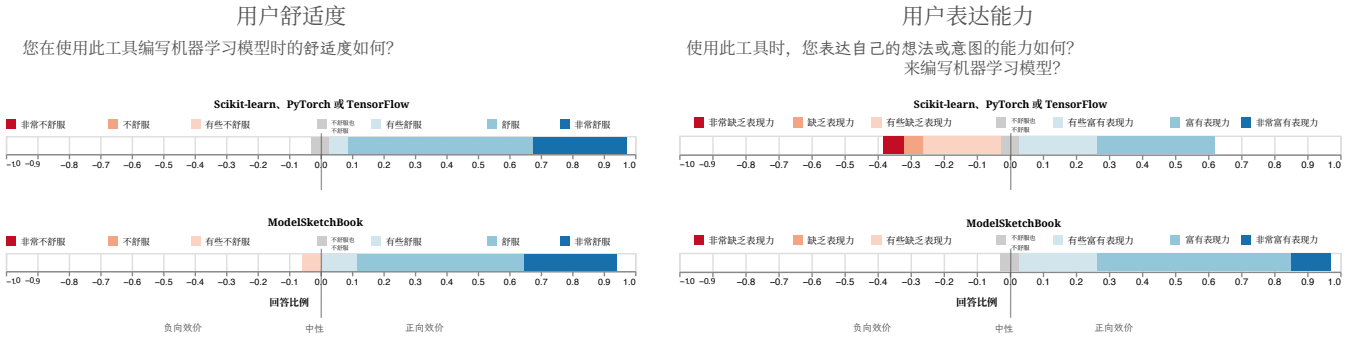

图 12：首次使用 ModelSketchBook API 时，参与者表示，他们的舒适程度与使用传统模型构建工具时相近（左），而使用 ModelSketchBook 时，他们的表达性更强（右）。

我们不应止步于笔记本原型，而应搭建专门的用户界面；他们还建议提供辅助工具，以应对概念中可能出现的偏差和过拟合问题。

K 最先进模型评估

K.1 基准数据集和任务详情

对于影评情感任务，我们从 IMDB电影评论数据集 [40], 这一二元情感分析数据集中收集数据。该数据集中的每个示例都是一条来自互联网电影数据库（IMDB）的评论，并已标注为正向或负向情感。该数据集中的正标签和负标签数量均衡，且经过筛选，只包含引发对立的评论：在 10 分制评分中，负面评论评分为 4 分或以下，正面评论评分为 7 分或以上。我们用于与草图模型进行比较的最先进水平模型是 distilbert-imdb [60], , 该模型在此二元分类任务上的准确率为 0.928。

对于评论有害性任务，我们采用了 Kumar等人 [35] 使用的数据集，该数据集汇集了来自 Reddit、Twitter 和 4chan 的评论，并收集了人们对这些评论毒性的不同看法。该数据集包含 17,280 名美国调查参与者给出的评分。每个项目由一条评论以及某位参与者对该评论给出的毒性评分组成，评分采用 5 分制，从“完全无毒”到“极其有毒”。该数据集包含超过 100,000 个示例，且评分偏向较低毒性。Kumar等人研究的一项重点是，人们对于何种内容构成有毒内容存在不同看法，这导致数据集中的评分者之间存在大量分歧，也使毒性预测任务颇具挑战性。我们将该任务定义为预测每个项目在所有标注者中的聚合评分，因为草图模型旨在进行项目级预测，而非（项目，标注者）级预测。因此，该任务的最先进水平模型是 Gordon等人 [22], 提出的聚合模型变体，其平均绝对误差为 0.90。

最后，对于仇恨表情包任务，我们采用了 Kiela等人 [31], 在第 5.1.2 节中介绍的完整训练集。与研究中使用的数据集样本不同，此处我们没有筛选出标签为“非仇恨性”的示例，而是采用 Meta AI 提供的集中式“仇恨迷因”定义，而非依据参与者各自偏好的定义。该任务的最先进水平模型取自仇恨表情包挑战赛 [69], 的第一名方案，其报告的曲线下面积为 0.845；我们还将其与数据集发布时 Meta AI 报告的曲线下面积为 0.714 的最先进水平结果进行了比较。

K.2 草图模型详情

对于每个任务数据集，我们随机抽取了最多 500 个示例来训练草图模型。相比之下，这些任务的最先进水平模型分别使用 25,000 个示例训练 IMDB 任务、使用 100,000 个示例训练评论毒性任务，以及使用 8,500 个示例训练仇恨表情包任务。随后，对于每项任务，我们团队的一名成员仅根据抽取的训练集，反复创建草图模型，力求基于训练数据打造性能尽可能好的草图模型。

对于影评情感任务，这一过程最终得到一个包含 16 个概念的草图模型：令人愉悦、无聊、平庸、引人入胜、评价过高、令人失望、获奖、浪费时间、仇恨、喜爱、正向情感、负向情感、讽刺、赞扬、称赞和嘲弄。由于该数据集仅包含影评文本，因此所有概念都是文本概念。该草图模型使用线性回归汇总指定的概念。接着，对于评论有害性任务，最终草图模型包含 14 个概念：种族主义、性别歧视、恐同、恐跨、歧视残障人士、宗教歧视、毒性、伤人、威胁、错误信息、网络挑衅、仇外、友善和无害。同样，由于该数据集中的所有示例都是文本评论，因此所有概念都是文本概念。该草图模型也使用线性回归汇总这些概念。最后，对于仇恨表情包任务，最终草图模型包含 7 个概念：种族主义的、仇恨言论、粗俗、无害、宗教、性别歧视的和暴力。所有这些都是基于表情包文本的文本概念；与研究参与者的看法一致，我们发现，对于这项任务，图像概念的效果不如文本概念。该草图模型使用带有两个隐藏层的多层感知机（MLP）分类器来汇总所选概念。

最后，为了评估草图模型性能，我们从测试集中随机抽取了 100 个示例（保持与完整测试集相同的标签分布）。我们将训练好的草图模型应用于该测试集，以生成草图模型预测。随后，我们将这些预测与真实标签进行比较，从而计算出与各任务最先进水平模型所报告的主要性能指标相对应的性能指标。

类别	主题	引言
点赞	易于使用和上手	“……我最喜欢的一点是，我能很快上手，比如我运行了几行代码，马上就看到了一些结果。所以我认为 ModelSketchBook 确实让机器学习变得更易于使用，这一点非常重要。”（P9），“我喜欢不用为任务编写任何代码。”（P5），“它很容易上手，也不容易产生误差。”（P7），“无需编写任何代码，就能获得一个非常稳健的基线。”（P3）
	可解释、可说明	“它的交互性很强，也非常容易理解，而我平时使用的机器学习模型通常并非如此。所以这真的非常有帮助。这就像是一种不同的思考方式。”（P10），“In在决策的可解释性和说明性方面，这在这类任务中也很重要，它[ModelSketchBook] 非常有帮助。”（P4），“我喜欢概念的语义理念，因为它确实让模型的可解释性成为关注的重点，而且根据我的许多经验，有时你会尝试一大堆模型而且从一开始，你并没有真正把它们视为语义概念。”（P16）
	快速反馈回路	“这个建模过程有更多快速反馈回路，也能更快地迭代特征探索，然后很快将其转化为模型，这当然是件好事。”（P5），“我之前在工作中建立的模型，我们做广告定向，所以很难获得这种实时反馈。”（P6）
	不同的高层次思维方式	“我喜欢通过定义这些概念，从人的思维方式出发来处理问题。”（P16）：“……侧重于理念和对任务的思考，而不是超参数或机器学习的技术细节。”（P10）：“这个过程有助于理清长远目标，以及机器学习生命周期和需要构建的流水线的结构：需要什么类型的数据、优先处理哪些数据，以及文本数据在哪些方面能比图像数据提供更多洞见，反之亦然。”（P1）
点踩	Notebook 界面有时显得杂乱	“复制粘贴某些单元格有点烦人。”（P5）
	需要花时间浏览示例来评估性能	“逐个查看每个示例很费劲，因为内容太长了。”（P6），“如果不用每次都显示所有示例，只为了评估性能，那就好了。”（P16）
	需要更多头脑风暴概cepts	“很难制定准确的概念 [……] 这更多是手动选择相关特征时遇到的问题。”（P9），“在如何创造性地提出概念方面，没有太多经验或指导。”（P10），“我不确定如何改进模型。”（P12）
建议	构建专用的用户界面	“我认为，将其从 IPython 中移出并采用更原生的用户界面会更好。”（P3），“我认为拥有一个图形用户界面会很好。”（P16）
	一个面向非技术用户和领域专家的版本	“我认为，对于没有机器学习或深度学习正规背景的模型开发者来说，这个应用程序接口非常出色。这个应用程序接口对产品工程师和产品经理来说是个非常实用的工具。”（P3）。“我觉得，这适用于对于那些对机器学习了解不多，但对……非常熟悉的人来说，这可能会很有参考价值，也很有用。手头任务了如指掌的人，让他们能亲手把东西做出来。”（P12）
	提供用于应对过拟合和偏差的工具	“人们很容易陷入一个误区：只顾着优化在所见示例上的性能，然后……忘了其他数据是什么样的。”（P8）。“人工干预概念设计可以有偏差。” (P4)，“也许我会遗漏边缘情况。” (P12)
	允许对权重进行更多手动控制以及模型类型	“我认为，给某个概念增加更多权重，可以改进许多草图。” (P13)，“缺少了一些特征（非线性模型、更多类型的信号）。” (P12)

表 7：在书面反馈和访谈中，参与者表示，他们喜欢模型草图绘制所带来的过程和认知模式。他们认为，可以通过专门的用户界面和支架来改进该工具，以协助构思概念。