scientific reports

Check for updates

OPEN

使用轻量级 LLMs 进行设备上查询意图预测，以支持无处不在的对话

Mateusz Dubiel、Yasmine Barghouti、Kristina Kudryavtseva 和 Luis A. Leiva

会话代理 (CA) 已经开始为用户提供交互式帮助。然而，当前CA的对话建模技术主要基于硬编码规则和严格的交互流程，这对其灵活性和可扩展性产生了负面影响。大型语言模型 (LLMs) 可以作为替代方案，但不幸的是，它们并不总是为最终用户提供良好水平的隐私保护，因为它们大多数都在云服务上运行。为了解决这些问题，我们利用迁移学习的潜力，研究如何最好地微调轻量级预训练的LLMs来预测用户查询的意图。重要的是，我们的LLMs允许在设备上部署，使其适合个性化、无处不在和隐私保护的场景。我们的实验表明，考虑到这些限制，RoBERTa 和 XLNet 提供了最佳权衡。我们还表明，经过微调后，这些模型的性能与 ChatGPT 相当。我们还讨论了这项研究对相关利益相关者（包括研究人员和从业者）的影响。总而言之，本文深入探讨了LLM设备上 CA 的适用性，并强调了LLM性能和内存占用之间的中间立场，同时还考虑了隐私影响。

关键词 对话代理、设计、信息检索、图形用户界面

当使用基于云的通信平台时，用户通常会失去对其隐私的控制，因为他们的数据由第三方服务器处理（并最终存储在第三方服务器上），这些服务器也可能用于服务提供商的进一步培训。此外，正如之前的研究表明，用户的隐私意图往往与他们的行为不同步，这可能导致用户无意中泄露敏感信息。当涉及到与旨在模仿类人交互的系统（例如会话代理（CA））进行交互时，这个问题是相关的，特别是在移动设备上，这些设备可以被视为用户很少分开的“亲密”对象。

如今，CA 变得越来越普遍。它们有多种形状和形式，例如智能手机上的数字助理（例如 Apple Siri、Google Assistant、三星 Bixby）、独立设备（例如 Amazon Echo Show、Google Nest 和腾讯听听）或汽车系统（例如，BMW 智能个人助理、Cerence 汽车平台），仅举几例。 CA 应用程序的热门领域包括健康和福祉、辅导和生产力。

如前所述，在云服务上运行的 CA 并不总是提供良好的隐私保护水平，因为用户无法保证他们的语音或文本命令将在那里得到安全处理。虽然已经证明传统 CA 可以完全离线运行，即使在 RaspberryPi 等低资源设备上也是如此，但这种方法的扩展性不佳。具体来说，开发 CA 的传统方法涉及使用预定义的槽填充机制和严格的交互流程，从而阻碍了新任务或领域的灵活性和可扩展性。例如，用户的话语“显示我的储蓄账户余额”表示带有“账户类型”槽的“显示余额”意图。总体而言，训练槽填充系统具有挑战性，因为它需要对每个槽和每个意图的用户话语的多种变体进行大量的手动编码。事实上，这种方法现在被认为已被弃用，仅适用于非常简单的情况，例如用户必须在某些给定选项之间进行选择的情况；参见呼叫中心的客户服务。

机器学习 (ML) 是一种替代方法，允许从数据中学习 CA 的行为，而无需对其进行显式编程，这使其更具可扩展性和通用性。更具体地说，迁移学习最近已成为自然语言处理 (NLP) 中事实上的 ML 方法，其中预训练的大型语言模型 (LLMs) 通过微调超参数来适应新任务在一个

1 卢森堡大学，4365 Esch-sur-Alzette，卢森堡。这些作者做出了同等贡献：Mateusz Dubiel 和 Luis A. Leiva 电子邮件：luis.leiva@uni.lu

小但具有代表性的数据集。然而，由于高计算要求和相关的高货币成本，许多研究人员无法对大多数现代 LLMs（例如 PaLM、LLaMA 或 GPT 系列）进行微调。因此，许多研究人员和从业者的唯一选择往往是依赖基于云的服务，为这些LLMs提供接口，从而损害用户的隐私，特别是对于使用敏感信息进行操作的 CA或旨在鼓励信息披露。

我们应该注意到，在本文中，我们所说的“模型”指的是“计算模型”，即经过训练（通过示例）来映射输入和输出的数据驱动结构。因此，计算模型既是结构又是数据。如上所述，如果没有数据，就不可能进行模型预训练。此外，有充足的证据支持高质量数据可以打造更好模型的说法，而不是相反。我们在本文末尾的“影响”部分详细阐述了这一观察结果。

为了弥合用户隐私和基于 LLM 的 CA 可扩展性之间的差距，我们研究了轻量级 LLMs 上的迁移学习，该迁移学习可以部署用于设备上的推理任务，这是一个基本的先决条件适用于移动和无处不在的系统；见图 1。具体来说，我们研究预测四个查询意图代理，如“方法”部分所述。意图代理对于理解交互式对话的上下文至关重要，因为它们决定了 CA 在正确解释用户输入并成功解决问题时的效率。我们的调查利用了“与 GUI 的对话”数据集，因为它为移动系统提供了一个有趣的测试平台，如“材料”部分中所述。

Rico、Enrico、VINS 或 WebUI 等 GUI 数据集可以在应用程序设计和开发的早期阶段发挥作用，为各种应用程序功能提供灵感和见解。虽然此类数据集包含有关 GUI 属性和相关技术规范的丰富信息，但查询它们可能需要使用开发人员的专业知识或复杂的基于 JSON 的 API，这使得没有编程经验的用户无法访问它们。为了解决这个问题，Todi 等人。提议使用对话模式来支持用户使用自然语言导航复杂的 GUI 数据集。在本文中，我们通过一系列适用于设备上 NLP 任务的轻量级 LLMs 进一步探索了这一概念。应该注意的是，虽然即时工程可以通过开发和优化指导模型的指令来更有效地使用 LLMs，但我们在 中探索的轻量级 LLMs 并不支持它。这篇论文。然而，出于比较目的，我们还评估了一个更大的、最先进的 LLM (ChatGPT) 的性能，它是通过即时工程进行微调的。

虽然LLMs变得越来越普遍，但它们很容易受到数据泄露的影响，从而损害了最终用户的隐私。为了探索依赖外部云服务进行部署的常规 LLMs 的替代方案，我们在此研究轻量级 LLMs 在涉及在商用移动设备上运行时预测查询意图的任务上的性能例如可以“随时随地”使用的智能手机和平板电脑。继斯塔尔等人之后。我们将移动设备定义为“一种便携式无线计算设备，无需额外设备即可携带，并且足够小，可以握在手中使用”。我们制定以下研究问题：

- RQ1 涉及模型在意图预测任务上的性能：
  - RQ1a：哪些预训练模型在针对不同任务中的查询意图进行微调后实现了最佳性能?
  - RQ1b：是否有一个模型在所有任务中都表现最好?
- RQ2关注模型性能和微调时间之间的关系：

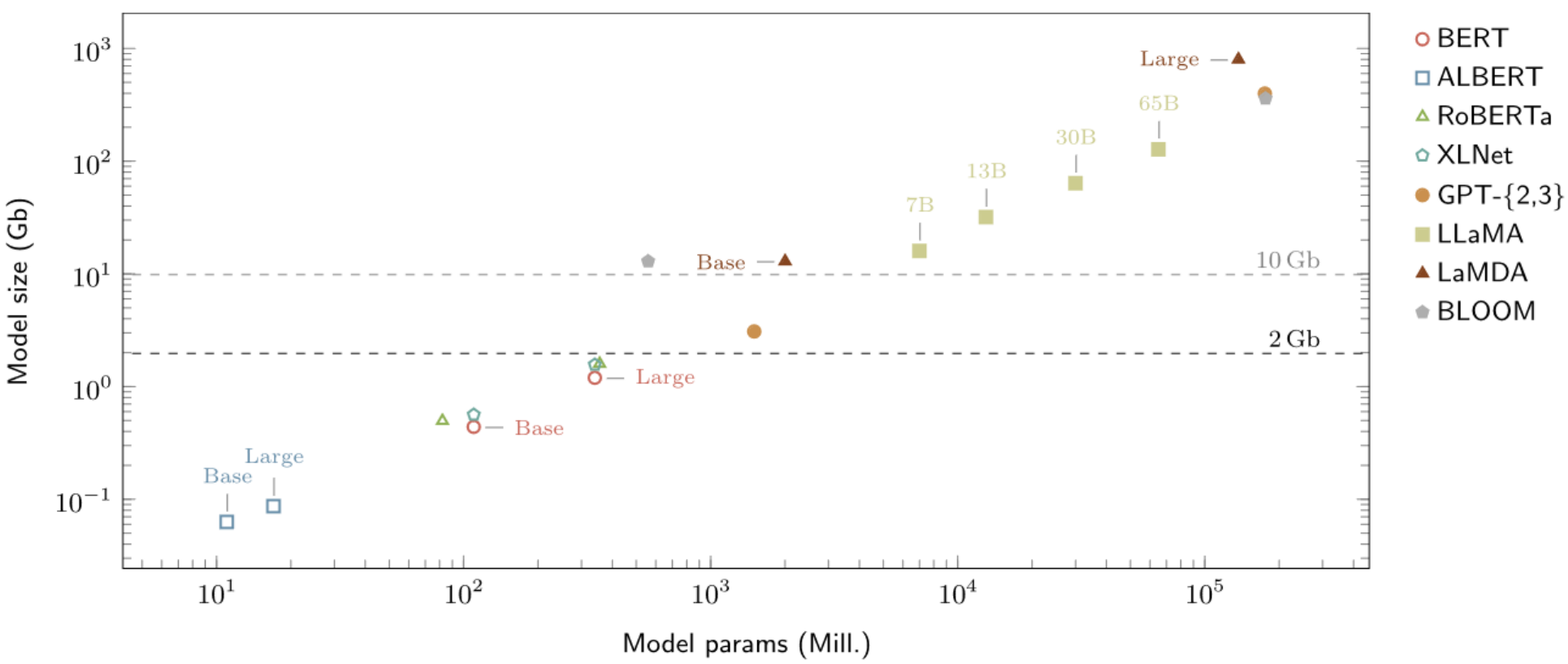

图 1. 一些流行的 LLMs 的概述。轴采用对数刻度以方便可视化（模型大小和模型参数之间几乎呈线性关系）。每个模型都有不同的可用变体，例如基本模型和大型模型（请参阅绘图注释）。如下一节所述，我们将 2 Gb 设置为 LLM 大小的上限，以便可以将其部署在商用移动设备和无处不在的设备上（有关更多详细信息，请参阅“模型”部分）。因此，在本文中，只有低于该上限的模型才被认为是轻量级的。

- RQ2a：每个预训练模型的最小微调周期数是多少?
- RQ2b：每个模型实现最佳性能的最佳微调 epoch 数量是多少?

通过解决上述研究问题，我们的工作做出了以下贡献：

- 我们深入了解轻量级 LLMs 对于设备上 NLP 任务的充分性，以及它们在 GUI 对话上下文中是否适合特定类型的用户查询。为了完整性，我们还对其他数据集进行了额外的实验。
- 我们阐明了性能与隐私权衡，并演示了在移动和无处不在的设备上部署基于 LLMs 的 CA 的可行性。我们表明，轻量级 LLMs 需要比之前假设的更多的微调周期才能达到其峰值性能。
- 我们讨论我们的研究对不同类型的利益相关者的影响，包括研究人员、开发人员、设计师和最终用户。虽然 ChatGPT 擅长零样本分类任务，但轻量级 LLMs 在微调后也能实现类似的性能（有时甚至更好）。

总体而言，这项工作对需要有效平衡性能和内存占用的移动和无处不在的系统做出了实证贡献，同时还考虑了对最终用户的隐私影响。更具体地说，我们使特定预训练模型的选择更加明智，并有助于在应用轻量级 LLMs 开发 CA 时避免试错选择方法。本文的主要前提是LLMs应该在商用硬件上进行微调，而不需要访问高性能计算设施或云服务提供商。

背景及相关工作

我们讨论了 CA 的相关性，以克服捕获用户信息需求的挑战（也称为语义差距），并解释迁移学习如何变革这一挑战。我们还讨论了 LLMs 当前的隐私和可持续性问题，这些问题可能对其广泛使用产生重要影响。

用于 GUI 交互的会话代理

“CA”术语是一个总括术语，包括两种主要类型的自动对话系统，即（1）旨在实现特定目标（例如预订航班）的任务导向型代理和（2）非任务导向型代理。在本文中，我们关注前一种意义，考虑一个代理，其目标是通过自然语言支持用户完成 GUI 数据集探索任务。

我们可以找到最近应用于 GUI 设计和布局任务的 CA 示例，例如：绘制草图、在应用程序中创建 UI 屏幕的任务快捷方式以及从自然语言短语创建低保真 UI 模型。然而，与我们的调查最相关的是托迪等人的工作。他们提出了一个 CA 原型来探索“与 GUI 的对话”数据集中的大量视觉设计。该原型以文本、数字、GUI 或设计的一部分的形式回答了用户的问题。例如，用户可以发出诸如“向我显示搜索栏设计的示例”或“应用程序上次更新时间是什么时候？”之类的查询。查找可以帮助他们满足搜索需求或提供有用参考点的信息。

最近，通过对话解决用户信息需求的话题越来越受到人们的关注，从而导致了一些基于CA的交互系统的开发。例如，贾汉巴赫什等人。 构建了人机交互的人工智能问答系统，协助用户处理业务文档。该系统非常符合实际用户的需求，因为他们的问题是在用户自然地处理文档（即执行日常工作任务）时就地收集的。在另一项研究中，王等人。研究了使用 LLM 与移动界面实现多功能对话交互的可行性。虽然他们设计了提示技术来使 LLM 适应移动 UI，但他们几乎没有探索单 UI 交互的信息查询。在我们的工作中，我们还探索导航查询并将研究扩展到 GUI 之外的 (i) 完整的移动应用程序和 (ii) 数据集。

使用 LLMs 进行情感分析

与我们工作相关的另一个研究领域是情感分析，这是一种 NLP 技术，其目标是检查话语或文本片段的情绪基调。Varia等人提出了一个统一的框架来解决基于方面的情感分析（ABSA）。 ABSA 是一项情感分析任务，涉及用户生成文本中的四个元素：方面术语、方面类别、观点术语和情感极性。瓦里亚等人。以涵盖所有子任务以及整个四重预测任务的多任务学习方式，通过指导提示对 T5 模型进行了微调。他们表明，所提出的多任务提示方法在几次学习环境中产生了性能提升。

在一项类似的研究中，Simmering 和 Huoviala 评估了 GPT-3.5 在 ABSA 任务的零样本和微调设置中的性能。他们发现，经过微调的 GPT-3.5 在 SemEval-2014 任务 4 的方面术语提取和情感极性分类方面均达到了 83.8% 的最先进的 F1 分数，在最先进的基础上进行了改进InstructABSA 模型提高了 5.7%。然而，性能的代价是需要微调 1000 倍的模型参数，以及相关的成本，并且推理时的延迟也会增加。 Simmering 和 Huoviala 的结果表明，虽然详细的提示可以提高零样本和少样本设置中的性能，但对于微调模型来说它们并不是必需的。

张等人。将 LLMs 与在特定领域数据集上训练的小型 LM 的功能进行了比较，例如对话分类和主观文本的多方面分析等任务。总的来说，张等人。评估了 26 个数据集上 13 项任务的性能，发现 LLMs 表现出了令人满意的性能

在较简单的任务中，小型语言模型在较复杂的任务中表现优于小型语言模型，因为小型语言模型需要更深入的理解或结构化的情感信息。尽管如此，LLMs 在少量学习环境中显着优于较小的模型，这表明它们在数据管理和标签有限时的潜力。

语义差距

语义差距（以计算机可可靠理解的方式表达信息需求的困难）是每个信息检索系统的基本挑战。 CA 越来越多地被用来弥补这一差距，允许用户用自然语言表达他们的查询。

在 GUI 相关 CA 的背景下，Todi 等人。引发了超过 1000 个查询意图，这些查询意图被手动标记为不同的类别，并用于开发 CA 原型。虽然他们介绍了如何设计智能系统以直观地与 GUI 数据集交互，但他们的 CA 原型基于流行的 Rasa 框架 [https://rasa.com]，该框架依赖于预定义的手写规则和用户故事。虽然基于规则的方法具有高度可解释性并且能够适应新的领域和语言，但它并不能完全捕捉自然语言的可变性，并且依赖于规则的质量和覆盖范围，这显然是不可扩展的。为了解决这个限制，在我们的工作中，我们采用了预先训练的 LLMs ，它提供了很大的灵活性，可以适应新任务，只需很少的编程工作，并且可以部署在商用移动设备上。

迁移学习

迁移学习是一种机器学习方法，其中预训练的模型可以用作新任务或领域模型的起点。例如，在 ImageNet 等通用图像数据集上训练的模型可以适应理解更具体的图像，如 X 射线图像。类似地，在语言消歧任务上训练的模型可以重新用于另一个任务，例如查询消歧。迁移学习的主要优点之一是，与仅从头开始使用少量数据进行训练相比，可以获得更好的性能。这是可能的，这要归功于模型超参数对新数据的适应（又称微调），这允许快速且更充分的优化。 NLP 中微调的直觉是，在预训练阶段，模型已经学习了语言的丰富表示，这使其能够更轻松地学习（或“微调”）下游语言理解的要求诸如句子分类之类的任务。有趣的是，之前的研究发现，经过良好微调的小语言模型可以胜过大规模语言模型。

在不同的未标记文本语料库上对 LLMs 进行预训练，在将 ML 用于 NLP 任务方面取得了多项突破。此类模型的一些最著名的示例包括 BERT、RoBERTa、XLNet、PaLM、LLaMA 和 GPT 系列，包括其最近且非常流行的变体 ChatGPT 和开源替代方案 BLOOM。这些LLMs成功的主要组成部分是变压器架构。在本文中，正如之前所暗示的，我们研究了一系列轻量级的LLMs，它们可以部署在商用移动设备上，以便离线运行推理任务。图 1 和表 2 提供了这些 LLMs 的概述，并与上述流行的 LLMs 进行比较。

LLMs 的隐私和可持续发展问题

随着 LLMs 变得越来越普遍，它们对用户隐私的影响也变得更加明显。先前的研究表明，LLMs 可能容易受到训练数据泄漏的影响，其中可以从模型中提取敏感信息。由于训练过程中处理的参数数量和数据集大小很大，大型模型特别容易无意中记住部分训练数据，而这些数据在使用过程中可能会被反省。反过来，基于这些模型构建的 CA 也可能容易遭受此类隐私泄露。然而，正如最近的一项研究所示，使用较小的模型有助于缓解 LLM 记忆问题。

此外，当今 CA 提供的类人交互为用户引导、欺骗和操纵提供了可能性。例如，当用户将个性化代理与人类混淆时，他们可能会披露更多信息和/或过度依赖个性化代理。有趣的是，如果 CA 更具社交互动性，并且更有可能向此类代理进行亲密的、隐私敏感的披露，那么人们对 CA 的信息共享做法（例如，与第三方共享用户数据）的负面看法往往会减少。最近一项关于智能手机使用情况的调查结果表明，用户对隐私的态度因性格特征而异，有些群体对风险持谨慎态度，而另一些群体则忽视潜在威胁，这可能会增加他们无意中泄露敏感数据的可能性。此外，用户向 CA 披露私人信息的倾向，加上缺乏有关信息收集、存储和披露实践的知识，似乎与他们所宣称的对其个人数据的透明度和控制的需求相矛盾。

基于大型神经网络的训练 CA 与高能耗相关，这反过来又会对环境产生长期影响，正如之前的工作所强调的那样。罗尔等人。提到本地设备上部署经过微调的 LLMs 可以提供一种增强隐私并减少环境足迹的方法。有趣的是，哈金斯等人。证明只需 25 个训练示例即可通过微调的 BERT 模型实现高意图识别准确度，展示了在个人笔记本电脑上训练小语言模型的可行性。在本文中，我们从尺寸、性能和整体微调时间方面系统地分析了8个轻量级LLMs。我们还讨论了它们在移动和无处不在的设备上部署的适用性；即 LLMs 可以加载到设备上并运行，无需与外部服务器通信。

团体	(id.) 示例查询	范围	目的	格式	信息。壮举。
设计师	79.有视频支持吗?	UI	通知	二进制	功能性
	862.我怎样才能建立个人资料?	UI	教育	文本	设置
	957.显示安全功能	数据集	筛选	图像	隐私
开发商	227.显示应用评分	App	通知	数字	元数据
	854.横幅在哪里可见?	UI	Find	图像	元素
	1154.有没有类似的应用	数据集	筛选	二进制	元数据
最终用户	16.如何保存搜索?	UI	教育	图像	功能性
	792.我可以查看我的信用余额吗?	UI	通知	二进制	元素
	887.显示使用我的位置的所有应用程序!	数据集	筛选	图像	传感器

表 1. 针对不同用户组的查询示例（逐字记录，无特定顺序，随机选择）。我们为每个查询提供一个 id，它引用“Conversations with GUIs”数据集中的行号。

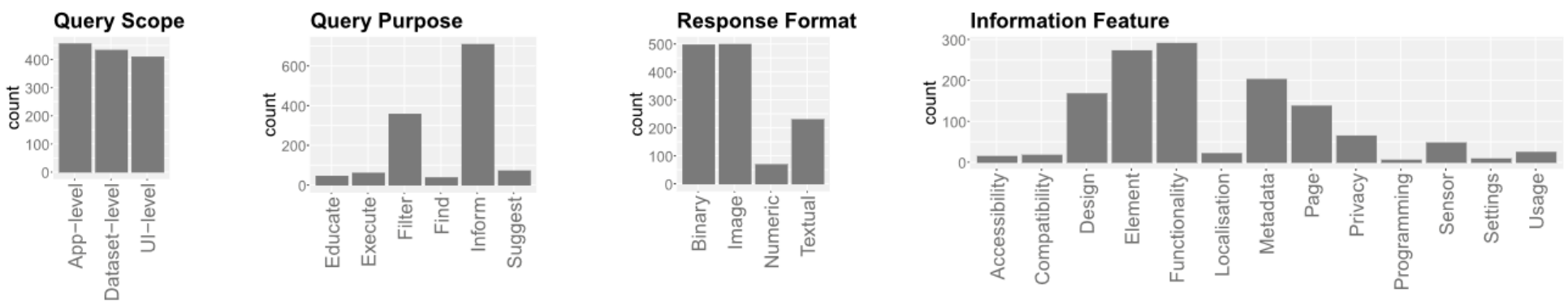

图 2.我们研究中考虑的目标变量的类别分布。

模型	层数	注意。头	参数	Size
BERT 基础	12	12	110M	440MB
BERT 大号	24	16	340M	1.2GB
阿尔伯特基地	12	12	11M	63MB
阿尔伯特·大号	24	16	17M	87MB
罗伯塔基地	12	12	82M	499MB
罗伯塔·拉格	24	16	355M	1.6GB
XLNet基地	12	12	110M	565MB
XLNet 大号	24	16	340M	1.57GB

表 2.所研究的预训练轻量级 LLMs 的描述。模型大小与层数、注意力头和可训练参数成正比。

材料

我们使用“Conversations with GUIs”数据集，其中包含 1317 个带标签的用户查询，作为我们的 LLMs 的培训材料。该数据集引出了来自三个不同用户组（最终用户、设计人员和开发人员）的四个目标变量（也称为意图：查询分数、查询目的、响应格式和信息特征）的示例查询，并与不同的 GUI 屏幕截图一起提供。 。

我们选择这个数据集的动机有三个。首先，它包含 LLMs 的域外数据，因此模型微调有望使它们充分执行。其次，它包括可被视为隐私敏感的个性化用户数据。第三，与医疗记录数据集相反，它是可公开访问的。总体而言，该数据集为我们探索 CA 性能和隐私考虑之间的权衡提供了一个有趣的基础，这对于移动和无处不在的系统至关重要。

表 1 中列出了每个用户组的示例查询。例如，查询“显示应用程序评级”(id.227) 是应用程序级意图的一个示例，其目标是获取有关应用程序元数据的数字信息，而查询“是否有类似的应用程序”（原文如此，id.1154）指的是数据集，其目的是过滤信息。请注意，查询类型的难度不同，其中一些查询的标签不明确，导致我们在“错误分类示例”部分中讨论不同类型的分类错误。另请注意，如前所述，即使该数据集并不意味着考虑隐私敏感数据，但许多查询也可以被视为此类数据（例如，参见 id.957、id.792 和 id.887）。

方法

接下来，我们根据“与 GUI 的对话”数据集提供的真实标签以及执行任务所选择的模型来定义四个意图预测任务。我们将查询意图预测任务构建为将用户话语（或查询）分类为四个不同的类别，我们将其称为目标变量（或意图）：

1. 查询范围（3 类） 查询是指单个 GUI、应用程序还是整个数据集。
2. 查询目的（6 类） 查询背后的可操作目标；例如根据某些标准进行过滤、获取更多信息、请求建议等。
3. 响应格式（4 类） 检索到的信息的预期传递格式：图像、文本、数字或二进制。
4. 信息特征（13类）查询所指的特定特征；例如，与应用程序的可访问性或隐私、其设计等相关。

最后，图 2 提供了每个目标变量的类分布。可以看出，我们在本文中解决了四个多类分类问题。此外，我们还可以看到许多类别是不平衡的。因此，正如稍后所解释的，我们在衡量意图分类性能时将考虑这一观察结果。

型号

我们利用 Ernie 存储库中的 LLMs 来进行我们的研究：https://github.com/labteral/ernie。选择 Ernie 的主要原因是它是公开可用的，并且包含适合设备上部署的最先进的轻量级 LLMs。根据最近的一项调查，低端（< 150 美元）到中端（< 550 美元）移动设备的 RAM 容量在 3 到 8 Gb 之间。考虑到大部分 RAM 将被后台服务和其他正在运行的应用程序占用，我们将 LLM 大小的上限设置为 2 Gb，以便可以部署在商用移动设备上。

BERT（来自 Transformers 的双向编码器表示）是一种基于 Transformer 的模型，也是同类中第一个突破性的 LLM。它将从左到右和从右到左的训练与掩蔽学习策略结合起来，其中训练序列中的每个单词都被模型必须预测的特殊标记替换。 BERT 在 BookCorpus（包含超过 11k 未出版的书籍）和英语维基百科上进行了预训练。

RoBERTa（稳健优化的 BERT 方法）建立在 BERT 的语言屏蔽策略的基础上，其中模型学习预测在其他未注释的语言示例中有意隐藏的文本部分。 RoBERTa 接受了五个数据集的联合训练：(1) BookCorpus、(2) 英语维基百科、(3) CC-News 包含 6300 万篇英文新闻文章、(4) OpenWebText 和 (5) Stories 包含以下内容的子集 CommonCrawl 语料库经过过滤，以匹配 Winograd Schemas 的故事风格。

ALBERT（A Lite BERT）是一种基于 BERT 的 Transformer 架构，但它包含的超参数要少得多（10M vs 110M）。为了实现这一目标，ALBERT 在不同层之间共享相同的权重：它有一个编码器层，对输入应用 12 次。由于 ALBERT 的超参数比 BERT 少大约 10 倍，因此它对计算资源的压力显着减轻。 ALBERT 使用与 BERT 相同的数据进行预训练。

XLNet 是一种自回归预训练 LLM，它使用双向上下文并最大化文本句子在所有排列顺序上的预期可能性，在 20 种不同的任务上优于 BERT。它融合了 Transformer-XL 架构的思想，克服了 BERT 和衍生模型的固定长度上下文限制，成为 NLP 应用的强大工具。 XLNet 在与 BERT 加上 CommonCrawl、Giga5（16 Gb 文本）和 ClueWeb 2012-B 相同的数据集上进行预训练。

有两点值得一提。首先，所有型号都可以根据其大小进行区分（例如，基本型号与大型型号），但它们都符合我们制定的 2 GB 限制。其次，所有模型都区分大小写，这意味着它们可以消除普通名词和专有名词之间的歧义；例如，苹果（水果）与苹果（品牌名称）。对于任何现代 CA 来说，这是一个可以在实践中使用的便利功能。

除了这些模型之外，我们还考虑了 ChatGPT，这是由 OpenAI 在未公开的大量数据上进行训练的最先进的专有 LLM；参见https://help.openai.com/en/articles/6783457。我们的目的是更好地了解轻量级 LLMs 与目前最流行的 LLM 相比如何。我们使用 ChatGPT 的 gpt-3.5-turbo-1106 版本，该版本可通过基于 JSON 的付费 API 进行自定义微调。

微调程序

我们将数据集中的所有编码查询随机分为三个部分：70% 用于训练，10% 用于验证，20% 用于测试。我们使用分层抽样来确保每个分区中类的分布相同。训练和验证分区用于模型微调，而测试分区用于模型性能评估，因为该分区模拟看不见的数据。

我们对 epoch 的增量数量（从 1 到 20）进行微调，在训练期间使用 16 个查询的批量大小，在验证时使用 32 个查询的批量大小。我们采用具有学习率的 Adam 优化器 $\eta = 2$ and 指数衰减 $\beta = 0.9$。为了稳定训练，我们将 Clipnorm 值设置为 1。总共，我们在“Conversations with GUIs”数据集上进行了 720 次微调实验，对应于 9 个模型的组合 $\times \times$ 4个查询意图预测任务。

:(0123456789) 卷:.(1234567890)

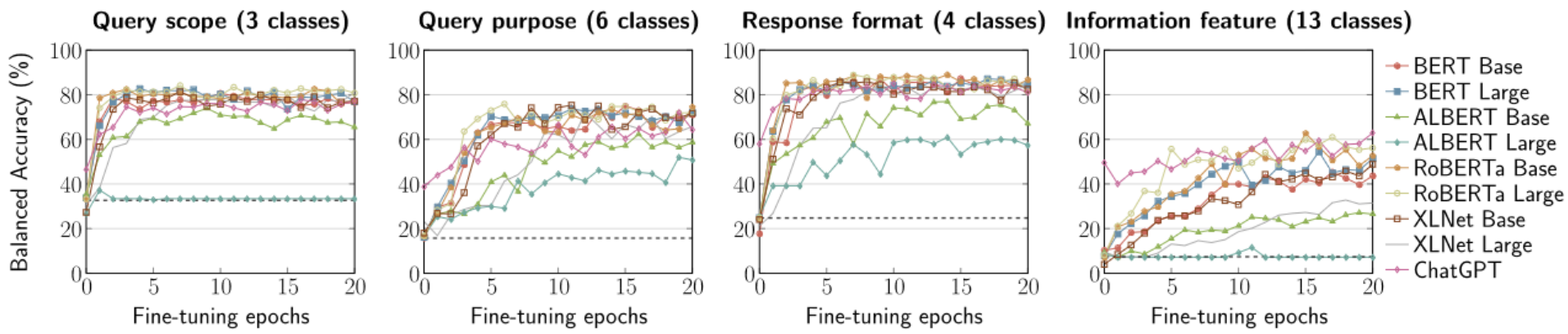

图 3. 平衡的精度结果。

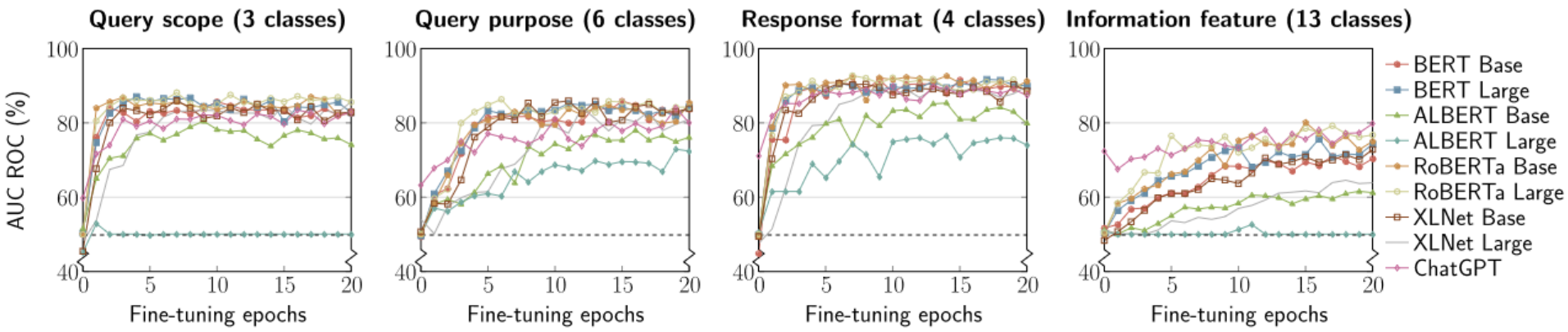

图 4.AUC ROC 结果。

所有实验，包括我们在补充材料中报告的其他实验，都是在单个 Tesla V100（SXM2，32 GB）GPU 卡中执行的。请注意，经过微调后，模型就可以部署在商用移动设备上。为了简化复制和进一步的后续工作，我们将在发布时分享我们的代码和模型检查点。有关 ChatGPT 微调过程以及在其他数据集上进行的实验的详细信息，请参阅补充材料。

结果

在下图中，我们根据平衡精度和 ROC 曲线下面积报告性能结果，如下定义。水平虚线表示随机分类器的分类性能（计算为 100/c，其中 c 是每种情况下要预测的类别数）。随机分类器提供了理论上的下限，即在模型微调后，LLM 的性能不应比随机分类器差。正如我们在下图中看到的，在所有情况下，ChatGPT 都实现了最佳的零样本性能，但在微调后，它的表现优于其他模型。

平衡的精度

分类精度（定义为所有预测中正确预测的数量）是分类问题中的标准评估指标，但它对不平衡数据非常敏感，即当目标类别之一比其他类别出现的频率更高时；参见图 2。因此，为了解释这一点，我们报告平衡准确性，即敏感性（真阳性率）和特异性（真阴性率）的算术平均值。图 3 总结了结果。

在查询范围预测方面，ALBERT 模型明显优于所有其他模型。总体而言，RoBERTa Large 是表现最好的，在 7 个 epoch 后达到了 84% 的平衡准确度。值得注意的是，BERT Large 在短短 4 个 epoch 内仅取得了稍差的性能 (83%)。至于查询目的，我们可以看到所有模型的性能都比查询范围的情况稍差。具体来说，RoBERTa Large 在 6 个 epoch 时达到了 76% 平衡精度的最佳结果，然而，在几个 epoch 后其性能下降。此任务中所有模型的性能下降可能是因为在这种情况下目标变量的数量是其两倍（6 类），因此可能比查询范围情况（3 类）有更多的歧义空间。

关于响应格式预测，该任务涉及 4 个类别，并产生与查询范围预测类似的模型性能。 RoBERTa Base 是性能最好的模型，在 7 个 epoch 后达到了 89.9% 的平衡准确度，紧随其后的是 BERT Large，在 18 个 epoch 后达到了相同的结果。总体而言，此任务中所有模型的行为与查询范围预测实验的行为非常相似，但 ALBERT Large 除外，在这种情况下表现明显更好。

最后，当涉及到信息特征预测时，由于目标变量数量巨大（13 个类别），该任务似乎最具挑战性。可以看出，所有模型都需要更多的训练周期才能达到最佳性能。该任务性能最佳的模型是 RoBERTa Large，在 15 个 epoch 内达到了 62.8% 的平衡准确度，其次是 ChatGPT，它在 20 个 epoch 后实现了相同的性能。值得指出的是，大多数模型的性能在超过 20 个 epoch 后仍能持续改进。

曲线下面积

ROC 曲线下面积 (AUC ROC) 是评估任何分类器判别能力的流行指标。 ROC 曲线提供了有关模型在一系列分类阈值范围内的假阳性率与真阳性率的信息，AUC ROC 是该曲线下的面积。由于 AUC ROC 是针对二元分类问题定义的，并且我们所有的实验都有两个以上的类，因此我们以一对一的方式计算它，以考虑多类分类。图 4 总结了结果。

可以看出，平衡精度实验中观察到的结果有类似的趋势（参见图3）。在查询范围预测方面，RoBERTa 模型表现最好，模型在 3 个 epoch 内达到 87% AUC ROC，而 ALBERT 模型表现最差。对于查询目的，RoBERTa Large 表现出最好的性能，在 6 个 epoch 内达到了 86% AUC ROC。 ALBERT 模型再次表现最差。这可能是因为与其他模型相比，其超参数数量要少得多，这使得 ALBERT 不适合对 GUI 相关任务进行微调。

在响应格式预测方面，RoBERTa Large 是性能最好的模型，在 7 个时期内达到 93% AUC ROC。与其他目标变量的性能相比，ALBERT Large 的性能明显更好（12 个时期的 76%），而其他目标变量没有实现任何改进。

最后，在信息特征预测方面，RoBERTa Base 收敛速度最快，在 15 个 epoch 时达到了 80% 的最佳 AUC ROC。 ChatGPT 在 20 个 epoch 时达到了 79.8% 的峰值性能。另一方面，与平衡精度相反，XLNet Base 和 XLNet Large 模型之间存在较大差异，这次 XLNet Base 往往比其 Large 变体表现更好。

调查结果摘要

我们已经看到所有研究LLMs的类别数量和分类性能之间存在清晰且有趣的关系。首先，对于少数类别，趋势类似于对数曲线，模型在几个时期后就饱和。然后，随着类别数量的增加，曲线变得更平坦，模型需要更长的时间才能达到最佳性能。虽然 ChatGPT 最初的表现优于所有其他模型（零样本分类），但它表现出相同的微调趋势。有趣的是，它在微调后优于其他模型，这与Zhang等人之前进行的实验一致。 。

我们的分析强调了LLM对查询意图预测任务进行微调的重要性，并强调了为当前任务选择适当模型的必要性。我们的分析还有助于根据模型复杂性和效率之间的权衡确定预测查询意图的最佳模型选择。重要的是，我们在商用 GPU 卡中对所有轻量级 LLMs 进行了微调，因此研究人员和从业者可以轻松重现我们的发现。

我们观察到的总体趋势是，随着时间的推移，模型收敛到最佳点或“最佳点”。我们应该注意到，我们的性能指标中观察到的小波动是由于历元内训练损失的波动造成的。它们主要是由于（1）梯度下降的随机性，（2）我们无法在单个批次中容纳所有查询，以及（3）模型的大尺寸与数据集的小尺寸成比例。

为了简洁起见，表 3 总结了所考虑的查询意图预测任务的前 3 个执行模型。所有模型均根据平衡精度的最佳值进行排名；即，当每个模型在最少的时期内达到最大平衡精度时。该表还报告了微调每个模型直至达到最佳状态的时间（以分钟为单位）。我们可以看到，RoBERTa Large 是唯一一个在所有任务中系统排名前三的模型。请注意，我们不是报告所有模型的摘要，而是通过分析性能最佳的前 3 个模型，为有兴趣构建聊天机器人原型的研究人员和开发人员提供更简洁、更有针对性的见解。事实上，

Task	前 3 名型号	零射击加速器(%)	巴尔。附件。 (%)	纪元	时间（分钟）
查询范围（3类）	罗伯塔·拉格	33.3	84.2	7	6.59
	BERT 大号	33.3	82.8	4	3.75
	罗伯塔基地	33.3	82.6	17	7.11
查询目的（6类）	罗伯塔·拉格	16.6	75.8	6	5.43
	XLNet基地	17.9	75.3	10	4.91
	BERT 基础	15.8	75.0	15	6.85
响应格式（4类）	罗伯塔基地	24.2	89.0	14	5.88
	罗伯塔·拉格	25.0	88.9	7	6.51
	伯特·基地	17.8	87.4	15	6.03
信息特征（13类）	罗伯塔基地	7.1	62.8	15	6.41
	聊天GPT	49.4	62.8	20	58.5
	罗伯塔·拉格	8.1	54.4	16	14.67

表 3. 根据实现的平衡精度（越高越好）和 epoch 数量（越低越好），微调后每个任务的最佳性能模型摘要。“Zero-shot Acc.”列表示微调之前的分类精度。计算微调时间（越低越好），直到在各自的行中报告最佳时期。 ChatGPT 仅进入信息功能任务的前 3 名。

在研究的初始阶段，尝试允许更快迭代周期的模型可能更实际。例如，RoBERTa Base 提供与 RoBERTa Large 类似的性能，但优化微调所需的时间却只有 RoBERTa Large 的一半。

为了进一步结合我们的发现，我们想强调关于模型复杂性和训练数据可用性之间的权衡的长期讨论。正如之前的工作所示，无论其复杂性如何，拥有更高质量的数据而不是更多的数据将提高模型性能。这种关系似乎也反映在我们的实验结果中，其中较大的模型（例如 BERT Large）或在更多数据源上训练的模型（例如 XLNet）并不总是能带来更好的性能。

讨论

我们首先回答有关模型性能 (RQ1) 以及训练周期数与性能之间的关系 (RQ2) 的主要研究问题。我们还提供了一些错误分类的示例，以更好地结合我们的发现。然后，我们讨论我们的研究对相关利益相关者的影响，考虑我们研究的局限性，并提出未来工作的几种可能的途径。

型号性能

在微调之前，所有轻量级 LLMs 都能够捕获一般语言特征和模式，但它们对于手头的任何任务都没有表现出足够的性能。然而，经过微调后，模型从训练数据中学习并表现出更好的性能，使它们能够更好地处理每个考虑的预测任务。需要注意的是，只有 ChatGPT 在零样本分类上表现良好，最初优于所有轻量级LLMs，但后来经过微调后被其他模型超越。 ChatGPT 出色的零样本性能归因于这样一个事实：它比我们研究过的任何轻量级LLMs都要复杂得多，并且对（书面）世界有更多的了解。

RQ1a：哪些预训练模型在根据范围、目的、响应格式和信息特征进行微调以预测查询意图后实现了最高的平衡精度?

我们观察到 RoBERTa 模型在所有四个预测任务中表现最好。具体来说，RoBERTa Large 在查询范围（84.2%）和查询目的（75.8%）方面表现最好，而 RoBERTa Base 在响应格式（89%，比 RoBERTA Large 好 0.1 个百分点）和信息特征（62.8%）方面表现最好。但必须指出的是，这两个模型在收敛之前的微调时间上存在显着差异，如下一节所述。

RQ1b：是否有一个模型在上述所有四项任务中都表现最佳?

根据我们之前的讨论，我们认为，虽然除了擅长零样本分类任务的 ChatGPT 之外，没有单一的赢家通吃模型，但 RoBERTa Large 是适合所有任务的最明智的模型。严格来说，虽然 RoBERTa Base 在预测响应格式方面取得了最佳性能，但与 RoBERTa Large 相比，其差距可以忽略不计：89% 与 88.9%，并且差异不具有统计显着性。因此，为了获得最佳整体性能（即最高精度），我们建议使用 RoBERTa Large 来开发用于 GUI 辅助的 CA。

尽管如此，还应该指出的是，尽管 RoBERTa Base 尺寸很小（几乎是 RoBERTa Large 的三分之一），但 RoBERTa Base 在包含 13 个不同类别的信息特征预测这一最具挑战性的任务中表现得异常出色（性能详细信息请参见表 3）。因此，应该考虑用于具有大量类别的消歧任务，特别是考虑到其收敛时间短（RoBERTa Large 为 6 分钟 vs 16 分钟）。

训练周期与表现之间的关系

正如前面所暗示的，每个模型都有其首选的“微调最佳点”。接下来，我们讨论每个模型在历元方面观察到的变异性，以实现其最佳性能。

RQ2a：每个预训练模型的最小微调周期数是多少?

在所有考虑的预测任务中，至少需要六到七个时期才能实现有竞争力的性能，除了信息特征预测之外，通常需要至少十五个时期。此外，除了信息特征之外，超过 15 个 epoch 后，一些模型开始过度拟合。这是因为该任务的高度复杂性，与其他任务相比，需要更长的训练时间。总的来说，建议不要在像我们分析的数据集上对这些模型进行超过 15 个时期的微调。

RQ2b：每个模型实现最佳性能的最佳微调 epoch 数量是多少?

我们观察到，对于所有任务，这个数字始终在 7-15 范围内，但具有挑战性的信息特征预测任务除外，其中所有模型都需要更多时间来收敛。我们观察到，除了 ALBERT 系列和 XLNet Large 之外的所有模型只需要 3-5 个 epoch 就开始接近其最佳性能。这一观察结果与之前报告的 BERT 模型类似范围的工作一致。在信息特征预测方面，对于 BERT 系列模型，需要 10-15 个 epoch 才能达到最佳性能，而对于 RoBERTa 模型，这个范围在 7 到 12 个 epoch 之间。这里一个值得注意的例外是 XLNet 模型，其性能遵循的趋势可能会在 20 个 epoch 后达到峰值。总的来说，我们观察到每个意图的类别数量越多意味着学习曲线越渐进。

错误分类的例子

表 4 包含之前报告的前 3 个最佳性能模型错误分类的查询示例。虽然模型总体表现良好，但一些用户查询由于其模糊性而被证明是困难的。例如，“如何创建购物篮？”（id.985）被 RoBERTa Large 预测为“建议”而不是“教育”目的，考虑到有限的背景，理论上可以分为这两个类别。同样，在“我们收集什么数据”（id.581）中，RoBERTa 模型预测信息特征是“元数据”而不是“隐私”类。同样，该查询可能对模型具有挑战性，因为它的

Task	前 3 名型号	(id.) 示例查询	查询意图	
			预测	真实情况
查询目的	罗伯塔·拉格		教育	教育
	BERT 大号	985.如何创建购物篮?	教育	教育
	罗伯塔基地		建议	建议
查询范围	罗伯塔·拉格		UI层	应用程序级
	XLNet基地	916. 我可以放大应用程序窗口吗?	UI层	应用程序级
	BERT 基础		UI层	应用程序级
响应格式	罗伯塔基地		图像	图像
	聊天GPT	813.隐私在哪里?	二进制	图像
	罗伯塔·拉格		文本	图像
信息特征	罗伯塔基地		元数据	隐私
	罗伯塔·拉格	581.我们收集什么数据	隐私	隐私
	BERT 大号		隐私	隐私

表 4. 表现最好的 3 个模型所犯的分类错误示例（以粗体突出显示）。
跨模型测试相同的查询。

简短且缺乏更广泛的上下文信息。另一个模棱两可的例子是“隐私在哪里？”，这对 ChatGPT 和 RoBERTa Large 来说是有问题的。（id.813），它被识别为“二进制”或“文本”响应而不是“图像”响应格式的请求。预测真实格式再次变得困难，因为两者都是解决此查询的同样明智的候选者。

还应该注意的是，数据集包含一些重复或接近重复但具有不同真实标签的查询。例如，“有登录页面吗？”（id.21 和 id.65）在信息特征方面被标记为 id.21 的“页面”和 id.65 的“功能”。这些情况虽然并不常见，但可能会在模型中引入一些噪声，从而使预测任务更具挑战性。

影响

总体而言，用户目前面临两种选择。他们可以使用 ChatGPT 而不进行微调来实现有竞争力的分类性能（特别是对于具有大量类别的意图），但代价是损害他们的隐私和一些金钱成本（每 1K 代币 0.008 美元，每个意图类别大约 1 美元）“Conversations with GUIs”数据集），或在自己的场所微调轻量级LLMs以获得更好的性能。

接下来，我们将讨论我们的研究结果对相关利益相关者的影响，包括开发人员、设计师、最终用户以及移动和无处不在的多媒体社区。值得注意的是，这些建议主要基于我们对“Conversations with GUIs”数据集的发现，该数据集比其他 NLP 数据集更具挑战性。我们参考补充材料进行其他实验，这些实验强调了我们在本文中研究的大多数模型的优异结果。

对于开发商

如果不进行微调，除了 ChatGPT 之外的所有模型在大多数情况下都表现得像随机分类器（见图 3 和图 4 中的虚线）。因此很明显，如果没有适当的微调，轻量级 LLMs 还没有准备好在 CA 上下文中支持用户的需求。这可以通过以下事实来解释：所有研究的轻量级 LLMs 都是在通用数据上进行预训练的，而“与 GUI 的对话”数据集特定于用户界面，因此可以被认为是“out-域外数据。有趣的是，仅仅一个时期之后，所有轻量级 LLMs 的分类性能结果就表现出了提升。此外，我们观察到较小的（基础）模型不一定比较大的模型需要更少的微调时期。

对于设计师和最终用户

CA 有潜力通过提供额外的通信渠道来简化与 GUI 的交互。例如，用户可以发出对话查询（通过语音文本）来快速访问有关应用程序隐私设置（例如 GPS 跟踪）的信息，否则这些信息将隐藏在冗长的技术规范文档中。了解哪种模型可以提供最准确的内存占用权衡，可以帮助用户决定是否值得在与模型交互上花费额外的时间来提高性能。应该注意的是，硬件限制可能会使无法访问高性能计算的用户无法对非常大的模型进行微调。这一点也适用于那些关心经济利用可用资产的公司的设计师。实际上，在全球范围内，我们的工作可以有助于更合理、更环保地使用计算资源。

适用于移动和普适计算

无处不在的应用程序预计将在动态环境中运行，在这种环境中，移动设备可以无缝运行，而不依赖于数据连接。我们的工作在这方面奠定了基石，允许感兴趣的研究人员在基于 CA 的应用程序的商用硬件上部署高效的轻量级 LLMs。

这些应用程序的一些示例包括开发多方 CA 或维护电子阅读器中的阅读流程。因此，我们的工作应该被视为移动和普适计算社区的支持技术。

正如 Mhlanga 所强调的那样，保护数据隐私不仅是一项道德义务，体现了对用户权利的尊重，而且也应该成为他们雇用的公司所有者和开发人员的首要任务。虽然《通用数据保护条例》(GDPR) 立法要求公司和组织保护最终用户的个人数据，但实际上不太可能实现 100% 合规。在我们的调查中，我们设想最终用户可以在自己的设备上本地运行基于轻量级 LLMs 的 CA，以避免向基于云的服务发送查询，从而保护他们的隐私。根据技术水平和可用资源（我们在实验中使用的 GPU 卡成本约为 3000 美元），轻量级 LLMs 可以由最终用户自己直接训练或作为一个单独提供-关闭购买公司提供的聊天机器人插件。

局限性和未来的工作

需要注意的是，虽然我们采用 2 Gb RAM 作为模型部署的上限，但这个大小可能会超过一些旧移动设备的容量，因此建议使用远低于该阈值的模型，以确保更广泛的范围的兼容性。总体而言，在我们的研究中，较大的模型产生了最佳性能，XLNet Base（用于查询目的）和 RoBERTa Base（用于查询范围）与性能最佳的模型紧密匹配，提供了可行的替代方案，同时大幅减少了所需的 RAM（减少了约 60%）适用于内存容量较低的旧移动设备。

我们在这项工作中没有探索的一个方面是中低端设备上的运行时性能分析，因为我们没有部署我们的模型。这个实施方面应该在未来的工作中进行探索。此外，建议考虑在线学习场景，其中最终用户在使用设备与模型交互时提供新的（看不见的）查询。这可以按照我们提出的相同微调方法来实现，但使用批量大小 1，一次摄取一个新查询。

未来的工作可以考虑更先进的微调技术，例如增量调整和低秩适应，以便微调LLMs，这对于许多研究人员来说成本高昂（就计算资源而言），例如如图 1 所示。然而，应该指出的是，与传统的微调相比，这些技术需要更多的数据才能收敛。最后，未来的工作还应该探索特定型号的中低端移动设备的运行时间和电池消耗，以提供实用的见解，从而提供有关在商品设备上部署轻量级 LLMs 的实用见解。

展望未来，我们希望向现有产品提出三种可能的隐私保护轻量级 LLMs 应用，这些产品可以在未来开发以支持不同的 GUI 用户组。首先，CA 可以支持开发人员从命令行界面按需使用它。或者，这样的 CA 也可以嵌入到集成开发环境中，例如 Visual Studio Code [https://code.visua lstudio.com/]。其次，设计人员可以受益于集成到 Figma [https://www.figma.com/] 或 Sketch [https://www.sketch.com/] 等界面设计工具中的 CA，以帮助他们协作创建新界面。在这种情况下，CA 可以以合乎道德的方式聚合匿名用户查询（例如，使用 NLP 方法删除品牌名称或实体），以便选择通过通知第三方服务来改进 CA 功能的用户。第三，由于最终用户最关心隐私功能，我们建议可以将 CA 集成到 Google Play [https://play.google.com/store/games] 或 iOS App Store [https://www.google.com/store/games] 中。 apple.com/app-store/]以便用户可以查询特定应用程序的隐私和其他元数据相关功能。

在结束本节时，我们要承认，尽管它具有增强隐私的潜力，但在我们自己的场所进行微调LLMs可能会引发一些道德问题。由于没有监督模型如何“在野外”部署，因此它们可能会被应用于恶意活动，例如窃取用户凭据（参见 FraudGPT [https://thehackernews.com/2023/07/new- ai-tool-fraudgpt-emerges-tailored.html], WormGPT。 [https://www.infosecurity-magazine.com/news/wormgpt-fake-emails-bec-attacks/]等）。尽管如此，我们相信，综合考虑，在内部微调轻量级 LLMs 给用户带来的好处大于风险。

结论

我们研究了如何最好地微调不同的轻量级预训练 LLMs 以进行设备上查询意图预测，以在与 CA 进行 GUI 相关交互期间为用户提供支持。我们的结果表明，在将隐私交给某些第三方云服务以换取性能提升（特别是在零样本分类场景中）和采用扩展性不佳的传统 CA 开发之间存在中间立场。虽然 RoBERTa Large 被证明是总体上表现最好的，但在所有探索的模型中，RoBERTa Base 和 XLNet Base 在性能（意图预测准确性和 AUC ROC）和内存占用之间提供了最佳权衡，因此它们可能同样适合用于设备上 CA 部署。总而言之，我们的研究结果为使用或使用 GUI 的不同利益相关者以及对开发需要平衡性能和内存占用同时考虑隐私影响的移动和无处不在的系统感兴趣的不同利益相关者提供了宝贵的见解。我们的模型检查点和软件可在 [URL TBA] 上公开获取。

数据可用性

当前研究期间使用和/或分析的数据集可根据合理要求从相应作者处获得。

收稿日期：2023 年 10 月 28 日；接受日期：2024 年 5 月 28 日
Published online: 03 June 2024

参考

1. Norberg, P. A.、Horne, D. R. 和 Horne, D. A. 隐私悖论：个人信息披露意图与行为。 J。消耗。亲。 41, 100–126 (2007)。
2. Nissenbaum, H. 背景下的隐私：技术、政策和社会生活的完整性。在上下文中的隐私（2009）。
3. Adam, M. & Klumpe, J. 通过聊天引导 - 消息交互性和平台自我披露对用户披露倾向的影响。参见 Proc，ECIS (2019)。
4. Bickmore, T. & Cassell, J. 关系代理：建立用户信任的模型和实现。见 Proc，CHI (2001)。
5. Panova, T. & Carbonell, X. 智能手机成瘾真的是一种成瘾吗？ J.行为。瘾君子。 7, 252–259 (2018)。
6. Kocielnik, R.、Xiao, L.、Avrahami, D. 和 Hsieh, G. 反思伴侣：一种让用户反思身体活动的对话系统。在 PACM 互动中。暴民。可穿戴无处不在的技术。 2（2018）。
7. 科切尔尼克，R.等人。我可以和您谈谈您的社交需求吗？了解健康领域对话用户界面的偏好。在Proc，CUI（2021）中。
8. Czerwinski, M.、Hernandez, J. 和 McDuff, D. 构建一个能够感知的人工智能：具有情商的人工智能系统可以学得更快、更有帮助。IEEE 光谱。 58, 32–38 (2021)。
9. Yuan, T.、Moore, D. 和 Grierson, A。用于教育辩论的人机对话系统：计算辩证法。国际。 J.阿蒂夫。英特尔。教育。 18, 3–26 (2008)。
10. Graesser, A. C.、VanLehn, K.、Rosé, C. P.、Jordan, P. W. 和 Harter, D. 具有对话功能的智能辅导系统。人工智能杂志。 22、39（2001）。
11. Darves, C. 和 Oviat, S. 从眉毛到信任：评估具身对话代理。与数字鱼交谈：为教育软件设计有效的对话界面，卷。 7（Ruttkay, Z. 和 Pelachaud, C. 编辑）（Springer，Dordrecht，2004 年）。
12. Brandtzaeg, P. B. 和 Følstad, A. 为什么人们使用聊天机器人。摘自 INSCI (2017)。
13. Grover, T.、Rowan, K.、Suh, J.、McDuff, D. 和 Czerwinski, M. 智能代理原型的设计和评估，以帮助提高工作中的注意力和生产力。在 Proc，IUI (2020)。
14. Bermuth, D.、Poeppel, A. 和 Reif, W. Jaco：离线运行的隐私感知语音助手。载于 HRI 程序 (2022)。
15. Pieraccini，R.等人。基于语义统计表示的语音理解系统。在过程中。 ICASSP 卷。 1（1992）。
16. Bobrow，D.G. 等人。 GUS，一个框架驱动的对话系统。阿蒂夫。英特尔。 8, 155 (1977)。
17. Jurafsky, D. 和 Martin, J. H. 语音和语言处理。在聊天机器人和对话系统（2023）中。
18. 李，T.J.-J。 & Riva, O. KITE：从移动应用程序构建对话机器人。在 Proc，MobileHCI (2018)。
19. Bhardwaj，V.等人。对话式人工智能——最先进的评论。在对话式人工智能中，（Rajavat, A. 等编辑）（Wiley，2024）。
20. 费舍尔，S.等人。 GRILLBot 实践：为适应性强的会话任务助手部署大型语言模型的经验教训和权衡 arXiv:2402.07647 (2024)。
21. 刘，X.等人。 P-tuning v2：快速调整可以与跨尺度和任务的普遍微调相媲美 arXiv:2110.07602 (2021)。
22. 李，X.等人。 FLM-101B：开放的 LLM 以及如何使用 10 万美元的预算对其进行训练 arXiv:2309.03852 (2023)。
23. Kirk, H. R.、Vidgen, B.、Röttger, P. 和 Hale, S. A. 界限内的个性化：用于将大型语言模型与个性化反馈相结合的风险分类和政策框架 arXiv:2303.05453 (2023)。
24. Kronemann, B.、Kizgin, H.、Rana, N. 和 Dwivedi, Y.K. 人工智能如何鼓励消费者分享他们的秘密？拟人化、个性化和隐私问题的作用以及未来研究的途径。跨度。 J·马克。 ESIC 27, 3–19 (2023)。
25. Lee, A. N.、Hunter, C. J. 和 Ruiz, N. Platypus：对 LLMs arxiv:2308.07317 (2023) 进行快速、廉价且强大的改进。
26. Gascó, G.、Rocha, M.-A.、Sanchis-Trilles, G.、Andrés-Ferrer, J. 和 Casacuberta, F。更多的数据总是能产生更好的翻译吗？载于 Proc，EACL (2012)。
27. Todi, K.、Leiva, L. A.、Buschek, D.、Tian, P. 和 Oulasvirta, A. 与 GUI 的对话。载于 Proc，DIS (2021)。
28. 德卡，B.等人。 Rico：用于构建数据驱动设计应用程序的移动应用程序数据集。参见 UIST (2017) 的 Proc。
29. Leiva, L. A.、Hota, A. 和 Oulasvirta, A. Enrico：移动 UI 设计主题建模的数据集。在 Proc 中，MobileHCI (2020)。
30. 布尼安，S.等人。 VINS：移动用户界面设计的视觉搜索。参见 Proc，CHI (2021)。
31. 吴，J.等人。 WebUI：用于通过 Web 语义增强视觉 UI 理解的数据集。参见 Proc，CHI (2023)。
32. 库马尔，R.等人。 Webzeitgeist：设计挖掘网络。见 Proc，CHI (2013)。
33. 怀特，J.等人。使用 ChatGPT arXiv:2302.11382 (2023) 增强提示工程的提示模式目录。
34. Carlini, N.、Liu, C.、Erlingsson, Ú.、Kos, J. 和 Song, D. 秘密共享者：评估和测试神经网络中的无意记忆。在 Proc 中，USENIX (2019)。
35. Stal, J. 和 Paliwoda-Pękosz, G. 在组织中使用移动技术提供知识的 SWOT 分析。在 Proc，ICTM (2018)。
36. Huang, F.、Schoop, E.、Ha, D. 和 Canny, J. F. Scones：走向对话式草图创作。载于 UIST 的 Proc (2020)。
37. Arsan, D.、Zaidi, A.、Sagar, A. 和 Kumar, R. 基于应用程序的虚拟助手任务快捷方式。载于 UIST 的 Proc (2021)。
38. Huang, F., Li, G., Zhou, X., Canny, J. F. & Li, Y. 使用深度学习模型根据高级文本描述创建用户界面模型 arXiv:2110.07775 (2021)。
39. 特·霍夫，M.等人。与文档对话：以文档为中心的援助的探索。参见 Proc，SIGIR (2020)。
40. Jahanbakhsh, F.、Nouri, E.、Sim, R.、White, R. W. 和 Fourney, A。理解处理业务文档时出现的问题 arXiv：2203.15073 (2022)。
41. Feng, S.、Jiang, M.、Zhou, T.、Zhen, Y. 和 Chen, C. Auto-Icon+：用于 UI 开发中图标设计的自动化端到端代码生成工具。 ACM 翻译。相互影响。英特尔。系统。 12(4), 1–26 (2022)。
42. Wang, B., Li, G. 和 Li, Y. 使用大型语言模型实现与移动 UI 的对话交互。参见 Proc，CHI (2023)。
43. 瓦里亚，S.等人。基于少量方面的情感分析的指令调整 arXiv:2210.06629 (2022)。
44. Simmering, P. F. & Huoviala, P. 用于基于方面的情感分析的大型语言模型 arXiv:2310.18025 (2023)。
45. 斯卡里亚，K.等人。 InstructABSA：基于方面的情感分析的指令学习 arXiv：2302.08624（2023）。
46. 张文伟、邓云、刘波、潘 S. J. 和 Bing L. 大语言模型时代的情感分析：现实检验 arXiv:2305。15005 (2023).
47. Smeulders, A. W.、Worring, M.、Santini, S.、Gupta, A. 和 Jain, R. 早年末基于内容的图像检索。 IEEE 传输。模式肛门。马赫。英特尔。 22, 1349 (2000)。
48. Torrey, L. 和 Shavlik, J. 迁移学习。机器学习应用和趋势研究手册：算法、方法和技术（编辑：Olivas, E.、Guerrero, J.、Martinez-Sober, M.、Magdalena-Benedito, J. 和 Serrano López, A.）（2010）。
49. 邓，J.等人。 ImageNet：大规模分层图像数据库。参见 Proc，CVPR (2009)。
50. Yadav, A.、Patel, A. 和 Shah, M. 关于解决自然语言处理中歧义的全面综述。人工智能公开赛 2, 85–92 (2021)。
51. Jurafsky, D. 和 Martin, J. H. 语音和语言处理，第一章。微调和屏蔽语言模型（2023）。
52. He, P.、Liu, X.、Gao, J. 和 Chen, W. DeBERTa：具有解纠缠注意力的解码增强 bert arXiv:2006.03654 (2020)。
53. 图夫龙，H.等人。 LLaMA：开放高效的基础语言模型 arXiv:2302.13971 (2023)。
54. Tenney, I.、Das, D. 和 Pavlick, E. BERT 重新发现了经典的 NLP 流程。在 Proc，ACL (2019) 中。
55. Devlin, J.、Chang, M.-W.、Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向 Transformer 的预训练 arXiv：1810.04805 (2018)。

:(0123456789) 卷:.(1234567890)

56. 刘，Y.等人。 RoBERTa：一种稳健优化的 BERT 预训练方法 arXiv：1907.11692（2019）。
57. 杨，Z.等人。 XLNet：用于语言理解的广义自回归预训练。参见 Proc，NeurIPS (2019)。
58. 乔杜里，A.等人。 PaLM：使用路径扩展语言建模 arXiv:2204.02311 (2022)。
59. Radford, A.、Narasimhan, K.、Salimans, T. 和 Sutskever, I。通过生成预训练提高语言理解 (2018)。
60. 雷德福，A.等人。语言模型是无监督的多任务学习者。 OpenAI 博客 1(8), 9 (2019)。
61. 布朗，T.B.等人。语言模型是小样本学习者。过程。 NeurIPS 33, 1877–1901 (2020)。
62. 斯蒂农，N.等人。学习根据人类反馈进行总结。参见 Proc，NeurIPS (2020)。
63. 大科学。 BLOOM：176B 参数的开放访问多语言语言模型。 arXiv：2211.05100 (2023)。
64. 瓦斯瓦尼，A.等人。您所需要的就是关注。参见 Proc，NeurIPS (2017)。
65. 张，C.等人。神经语言模型中的反事实记忆 arXiv:2112.12938 (2021)。
66. 迪南，E.等人。预测 e2e 对话式 AI 中的安全问题：框架和工具 arXiv:2107.03451 (2021)。
67. 卡利尼，N.等人。量化神经语言模型的记忆 arXiv:2202.07646 (2022)。
68. 魏丁格，L.等人。语言模型带来的风险分类。载于 Proc，FAccT (2022)。
69. Tourangeau, R.、Couper, M. P. 和 Steiger, D. M. 人性化的自我管理调查：网络和 IVR 调查中的社会存在实验。计算。哼。行为。 19, 1–24 (2003)。
70. Sannon, S.、Stoll, B.、DiFranzo, D.、Jung, M. F. 和 Bazarova, N. N.“我刚刚分享了您的回复”：将通信隐私管理理论扩展到与会话代理的交互。在 PACM 嗡嗡声中。计算。相互影响。 4（2020）。
71. Fleischhauer, D.、Engelstätter, B. 和 Tafreschi, O。智能手机用户的隐私悖论。在 Proc，MUM (2022)。
72. Lutz, C. & Tamò, A. RoboCode-Ethicists：隐私友好型机器人，工程师的道德责任？摘自 Proc，网络科学 (2015)。

73. Keysermann，M.U. 等人。我可以相信你吗？：与人工同伴共享信息。载于 AAMAS（2012 年）。
74. Strubell, E.、Ganesh, A. 和 McCallum, A. NLP 深度学习的能源和政策考虑因素 (2019)。
75. Bender, E. M.、Gebru, T.、McMillan-Major, A. 和 Shmitchell, S. 关于随机鹦鹉的危险：语言模型会太大吗？载于 Proc，FAccT (2021)。
76. 罗尔，S.等人。开放域对话代理：当前进展、未解决的问题和未来方向 arXiv:2006.12442 (2020)。
77. 哈金斯，M.等人。意图识别实用指南：在现实 HRI 应用中评估最少训练数据的 BERT。载于 HRI 程序 (2021)。

78. Yates, D. 和 Islam, M. Z. 智能手机上的数据挖掘：介绍和调查。 ACM 计算。幸存者。 55, 1–38 (2022)。
79. 朱，Y.等人。协调书籍和电影：通过观看电影和阅读书籍来实现故事式的视觉解释。在过程中。 ICCV 19-27 (2015)。
80. 麦肯齐，J.等人。 CC-News-En：大型英文新闻语料库。见 CIKM (2020)。
81. Trinh, T. H. & Le, Q. V. 常识推理的简单方法 arXiv:1806.02847 (2018)。
82. 共同爬行基金会。 CommonCrawl 数据集，https://commoncrawl.org (2019)。
83. Levesque, H. J.、Davis, E. 和 Morgenstern, L. 维诺格拉德模式挑战。见 Proc，KR (2012)。
84. Lan，Z.等人。 ALBERT：一个精简版 BERT，用于语言表示的自我监督学习。在 Proc，ICML (2019) 中。
85. 戴，Z.等人。 Transformer-XL：超越固定长度上下文的细心语言模型。在 Proc，ACL (2019) 中。
86. Parker, R.、Graff, D.、Kong, J.、Chen, K. 和 Maeda, K. English Gigaword，第 5 版（2011 年）。语言数据联盟，LDC2011T07。

87. Callan, J. Lemur 项目及其 ClueWeb12 数据集。 SIGIR 开源信息检索研讨会 (2012) 特邀演讲。

88. Kingma, D. P. & Ba, J. Adam：随机优化方法 arXiv:1412.6980 (2014)。
89. Powers, D. M. W. 评估：从精确度、召回率和 F 测量到 ROC、信息性、标记性和相关性。国际。 J。马赫。学习。技术。 2 arXiv：2010.16061（2011）。
90. Merchant, A.、Rahimtoroghi, E.、Pavlick, E. 和 Tenney, I。微调期间 BERT 嵌入会发生什么？。在过程中。 BlackboxNLP 肛门研讨会。解释。 NLP 神经网络（2020）。
91. 华红，李X，窦德，徐成志。 & Luo, J. 用于改进 BERT 微调的噪声稳定性正则化 arXiv: 2107. 04835 (2021)。
92. Zargham, N.、Bonfert, M.、Porzel, R.、Doring, T. 和 Malaka, R. 多代理语音助手：用户体验调查。见 Proc，MUM (2021)。

93. Draxler, F.、Rakytianska, V. 和 Schmidt, A. 通过交互式语法增强功能来维持电子阅读器的阅读流程以进行语言学习。在 Proc，MUM (2022)。
94. Mhlanga, D. 在教育中开放人工智能，以负责任且合乎道德的方式使用 ChatGPT 以实现终身学习。金融科技和人工智能促进可持续发展：智能技术在实现发展目标中的作用（Springer，2023）。
95. 丁，N.等人。 Delta 调优：预训练语言模型参数有效方法的综合研究 arXiv:2203.06904 (2022)。

96. 胡，E.J.等人。 LoRA：大型语言模型的低秩适应。在 Proc，ICLR (2021) 中。
97. Pu, G.、Jain, A.、Yin, J. 和 Kaplan, R. LLMs PEFT 技术优缺点的实证分析。 ICLR 理解基础模型研讨会（2023 年）。

致谢

本文提出的实验是使用卢森堡大学的 HPC 设施进行的。这项工作得到了欧盟 Horizon 2020 FET 计划（赠款 CHIST-ERA20-BCI-001）和欧洲创新理事会探路者计划（SYMBIOTIK 项目，赠款 101071147）的支持。

作者贡献

M.D.：概念化、写作——初稿、审查和编辑。 Y.B.：方法论、软件、验证。 K.K.：软件、验证。 L.A.L.：概念化、方法论、软件、可视化、写作评论和编辑。

利益竞争

作者声明没有竞争利益。

附加信息

补充信息 在线版本包含 https://doi 上提供的补充材料。 org/10.1038/s41598-024-63380-6。

信件和材料请求应发送至 L.A.L.