AutoML-GPT：使用 GPT 实现自动机器学习

Shujian Zhang ChengyueGong Lemeng Wu Xingchao Liu
Mingyuan Zhou 德克萨斯大学奥斯汀分校 {szhang19,
mzhou}@utexas.edu

arXiv:2305.02499v1 [cs.CL] 4 May 2023

摘要

人工智能任务涵盖广泛的领域和学科。尽管已经针对特定任务和应用设计了大量人工智能模型，但在寻找合适的模型架构、优化算法和超参数方面，往往仍需要投入大量人工精力。近年来，ChatGPT 等大语言模型（LLM）取得了显著进展，在推理、理解和交互等多个方面展现出卓越的能力。因此，我们提出开发面向任务的提示词，并自动利用大语言模型实现训练流程的自动化。为了实现这一理念，我们提出了 AutoML-GPT，该系统利用 GPT 作为连接各种人工智能模型的桥梁，并通过优化后的超参数动态训练模型。AutoML-GPT 根据模型卡和数据卡中的用户请求动态生成相应的提示词段落。最终，借助这一提示词段落，AutoML-GPT 将自动完成从数据处理、模型架构设计、超参数调优到预测训练日志生成的整个实验过程。通过充分发挥 AutoML-GPT 强大的语言能力以及现有人工智能模型的优势，AutoML-GPT 能够应对各种任务和数据集中的众多复杂人工智能任务。该方法在计算机视觉、自然语言处理以及其他具有挑战性的领域取得了显著成果。大量实验和消融实验表明，我们的方法具有通用性、有效性，并能够为众多人工智能任务带来益处。

1 引言

人工智能（AI）近年来取得了显著进展。在这些进展中，ChatGPT [OpenAI, 2023] 尤其突出，因其具备推理、理解和交互能力[Wu et al., 2023]。根据指令执行新任务的能力，是迈向通用人工智能的关键一步，而大语言模型（大语言模型）的卓越能力催生了许多新兴研究主题，例如上下文学习 [Ram et al., 2023; Xie et al., 2021], 思维链提示 [Pilault et al.,2023; Wei et al.,2022b], 检索与阅读 [Izacard and Grave, 2020;Zhang et al., 2021, 2022], 以及基于GPT的智能系统 [Zheng et al., 2023]。这些领域旨在探索大语言模型的巨大潜力，并为构建复杂的人工智能系统带来无限机遇。

大语言模型（例如 GPT-4 [Brown et al., 2020年; OpenAI, 2023], LLaMA [Touvron et al., 2023], Flan-T5[Chung et al., 2022], 以及 PaLM [Chowdhery et al., 2022], ）展现出对自然语言的深入理解能力，以及生成连贯且符合上下文的回答的能力。这一进展为涉及不同领域数据的复杂任务开辟了新的潜在应用，例如图像和文本处理，以及领域专属知识的融入。在此背景下，大语言模型发挥着至关重要的作用，因为它们理解和生成自然语言的能力，使 AI 能够更好地理解并应对各种挑战。

在本文中，我们旨在开发一种名为 AutoML-GPT 的自动机器学习（AutoML）系统，该系统利用大语言模型，根据用户输入和描述自动在数据集上训练模型。大语言模型被用作自动训练系统，以建立与多功能模型之间的联系并处理输入。我们建议将语言用作大语言模型进行交互的通用接口和提示词。

与用户交互。通过将数据和模型描述纳入提示词，大语言模型可以管理用于数据处理、模型架构设计和超参数调优的人工智能模型。它们可以根据需要调用这些模型来处理人工智能任务，并返回预测训练日志。然而，将多个人工智能模型纳入大语言模型，需要大量高质量的模型描述。为克服这一挑战，我们建议利用模型卡 [Mitchell et al., 2019]，其中提供了定义明确的模型描述；同时利用数据卡 [Gebru et al., 2021]，用于特定的人工智能任务。这种方法能够让我们通过基于语言的接口连接多种模型，从而促进复杂人工智能任务的解决。通过捕捉模型和数据集之间的相似性，还能提升它们之间的可迁移性。

AutoML-GPT 将多样化的机器学习模型、训练流水线和数据集连接起来，以解决众多复杂的人工智能任务。具体而言，对于我们旨在解决的每项人工智能任务，我们利用相应的描述（如模型卡和数据卡），将段落融合为提示词，输入预训练大语言模型（如 ChatGPT），从而建立自动机器学习流水线。随后，在我们的系统中，大语言模型会执行自动训练，并针对用户输入的问题返回预测的训练日志。基于这些训练日志，我们可以进一步与大语言模型交互，以解决图1所示的请求（如超参数调优）。因此，AutoML-GPT 的整个流程可分为四个阶段：1）数据处理；2）模型架构设计；3）根据预测的训练日志进行超参数调优；4）基于实验数据进行人工反馈。

得益于这种设计，图1中的 AutoML-GPT 能够使用外部模型，从而处理知名基准测试中的多项任务，并在仅提供元数据（数据卡）的情况下，将所学知识迁移到未知的私有数据集。此外，该流水线还使 AutoML-GPT 能够持续吸收特定任务专家的能力，从而实现可扩展、可成长的人工智能能力。总而言之，我们的贡献如下：

- 为结合大语言模型和专家模型的优势，我们提出了 AutoML-GPT。它负责数据处理和模型架构设计，并针对每项具体任务自动开展实验。

- 通过将模型卡与模型描述、数据卡与数据描述相结合，我们提供固定格式的提示词段落，并构建训练流程以处理通用人工智能任务。
- 在涵盖语言、视觉和持续学习的多种人工智能任务上进行了大量评估，结果展现了 AutoML-GPT 自动训练的能力。此外，结果还证明了为未见过或新数据集提供超参数调优的有效性。

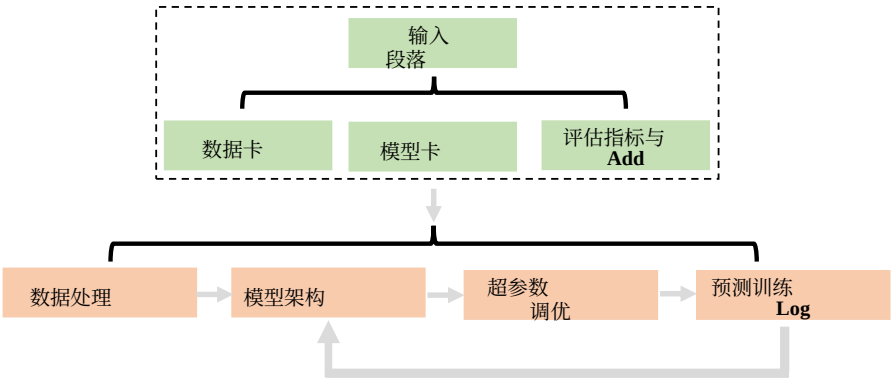


图1：AutoML-GPT 概览。图中标注了部分符号及其对应组件。“评估指标与附加请求”指评估指标和额外请求。

2 AutoML-GPT

AutoML-GPT 是一个协作系统，它依据数据和模型信息来组织提示词输入段落。大语言模型充当控制器，而众多专家模型则作为协作执行者。

AutoML-GPT 的工作流程包含四个阶段：数据处理、模型架构设计、超参数调优和训练日志生成。具体而言，我们为 AutoML-GPT 提出了一套通用方案：1）结合模型卡和数据卡生成固定格式的提示词段落；2）构建训练流程，并根据所选数据集和模型架构处理用户请求；3）生成性能训练日志并调整超参数；4）使用自动建议的超参数调整模型。

2.1 输入分解

在 AutoML-GPT 的第一阶段，大语言模型会接收用户输入。为了提升大语言模型的性能并生成有效的提示词，我们会针对输入提示词采用特定的指令。这些指令包含以下三个部分。

数据卡 为了明确数据集的预期用途，并尽量避免在不适用的场景中使用数据集，我们使用数据卡，为该数据集提供全面的文档说明。如图 2 所示，数据卡的关键组成部分包括数据集名称、输入数据类型（例如图像数据或文本数据）、标签空间（例如类别类型或分辨率）以及默认评估指标。

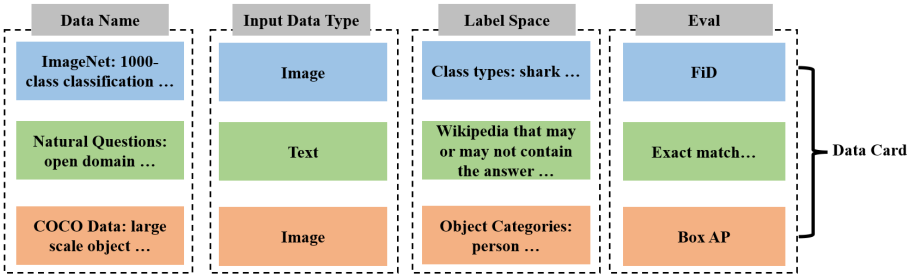
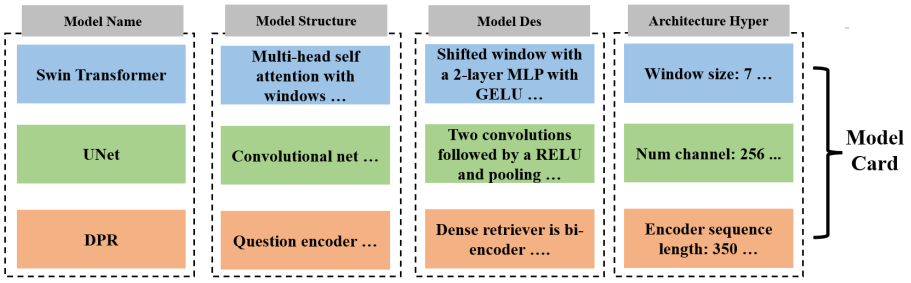


图 2：数据卡包含数据名称、输入数据类型、标签空间和评估指标。在数据卡中，相同颜色表示信息来自同一个数据集。

模型卡 图 3 中的模型卡是对前文所述“数据卡”的补充，作为所提出的范式之一，用于报告训练和测试数据集所使用模型的详细信息。模型卡由模型名称、模型结构（例如，Swin Transformer [Liu 等人提出的模型， 2021] 搭配 UperNet [Xiao 等人提出的 2018]头部）、模型描述和架构超参数组成。通过提供这些信息，模型卡可以向大语言模型说明所使用的机器学习系统，以及用户希望在模型架构上拥有的灵活程度。这将进一步借助大语言模型产生更具包容性的结果。



Figu图 3：模型卡包含模型名称、模型结构、模型描述和架构 re
hyper参数。在模型卡中，同一种颜色表示来自同一模型卡的信息。

评估指标和其他请求 除了模型卡和数据卡之外，用户还可以选择请求更多评估基准测试、评估指标或任何约束条件。除默认的评估指标外，在选择模型架构时，我们还可以根据用户的请求添加特定指标或约束条件。例如，给定“推理时间小于 10 帧每秒”这一约束条件后，我们会根据评估指标和约束条件处理用户请求。得益于这一指令以及用户对这些评估指标和其他请求提供的人工反馈，大语言模型能够更好地遵循指令。AutoML-GPT 将这些任务规范作为高层级指令提供给大语言模型，以便其据此分析用户请求。

2.2 数据处理

数据处理是机器学习中不可或缺的一步，因为数据的质量以及从中提取的有用信息会直接影响模型的学习能力。因此，在将数据输入模型之前对其进行处理至关重要。例如，在计算机视觉中，数据处理是指用于准备原始图像数据，以便进行分析或供机器学习算法使用的一系列技术和方法。这可能包括图像缩放、归一化、数据增强和过滤。同样，在自然语言处理（NLP）项目中，数据处理是指将原始文本数据转换为结构化且干净的格式，使机器学习算法能够轻松理解和处理。通常会使用分词、停用词移除、转为小写，以及移除特殊字符和数字等技术。AutoML-GPT 会根据提供的数据卡和数据描述，结合项目要求和数据特性，提供具体的数据处理技术。

2.3 模型架构

在处理任务列表后，AutoML-GPT 需要将每项任务与相应的模型进行匹配，本质上就是为列表中的每项任务选择合适的模型。为此，我们首先从用户输入中获取模型卡和模型描述。随后，利用上下文内任务-模型分配机制，动态地为各项任务分配模型。这种方法通过整合所提供的模型描述并更好地理解用户请求，实现了增量式模型访问，同时提供了更高的开放性和灵活性。

模型架构是指对机器学习模型的设计、结构和组件所作的详细说明。这些描述通常包括以下要素：输入层和输出层、隐藏层、激活函数、损失函数，以及模型特定的组件（例如注意力机制、卷积层或循环层）。

2.4 使用预测训练日志进行超参数调优

为了找到一组能够使特定数据集上的给定模型获得最佳性能的最优超参数，超参数调优是机器学习中的关键步骤。超参数是在训练过程中不会被学习、而是预先定义的配置设置，用于控制模型学习行为的各个方面。常见超参数的示例包括学习率、批次大小、隐藏层数量以及每层的神经元数量。

为了无需在真实机器上进行训练就能调优超参数，我们会根据提供的数据卡和模型卡，针对给定的超参数设置生成训练日志，并据此预测性能。AutoML-GPT 将自动执行训练并返回训练日志。模型在数据集上的性能训练日志会记录训练过程中收集的各种指标和信息。它有助于了解模型的进展、发现潜在问题，并评估所选架构、超参数和优化技术的有效性。典型的训练日志包含训练轮次编号以及训练和验证指标。通过检查训练日志，我们可以根据用户反馈初步了解模型性能。

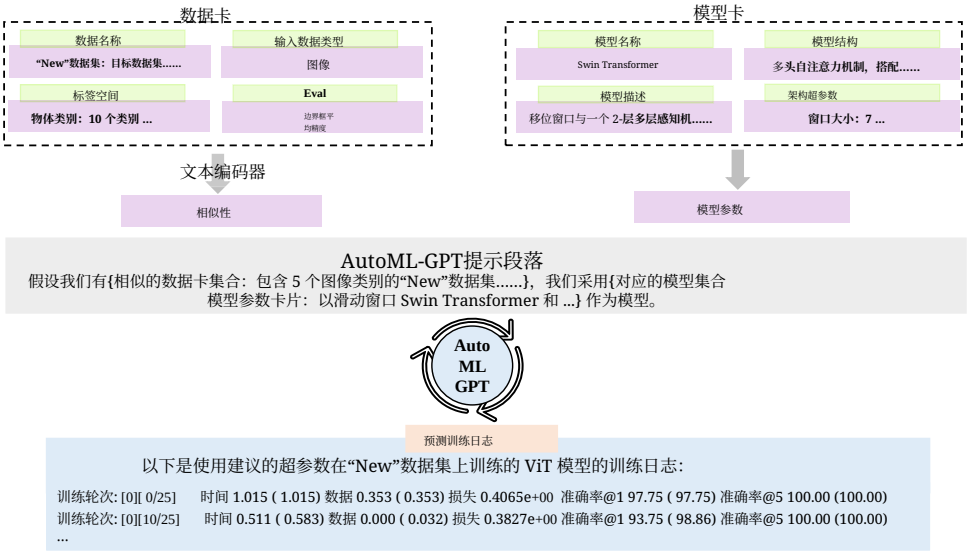


图 4：AutoML-GPT 在未见过的数据集上的概览：顶部模块展示数据卡和模型信息。我们首先记录多个数据集的训练信息。通过文本编码器处理这些数据集的数据卡，以获取相似度分数；随后将这些分数与相应已训练模型的模型参数结合，形成 AutoML-GPT 提示段落。底部模块展示基于为未见过的数据集推荐的超参数设置生成的预测训练日志。

未见过的数据集 针对未见过的私有数据集进行超参数调优可能更具挑战性。给定未见过的数据集的元数据，AutoML-GPT 可以推荐一个可能适用于该数据集的超参数配置。我们依靠数据卡利用必要的文本描述，并识别未见过的数据集与现有数据集之间的相关性。基于这种相关性，我们将现有数据集的超参数设置迁移到新的未见过的数据集。

为了计算相关性，我们使用文本编码器对数据卡进行编码。具体而言，数据卡包含类别类型、分辨率、图像大小及其他相关元数据等信息。我们将数据集规模、任务描述、标签空间以及输入/输出数据类型作为文本编码器（例如，CLIP [Radford et al., 2021]）的输入，并使用编码后潜在表示的相似度分数来描述这个未见过的数据集与现有数据集之间的相关性。

3 实验

我们使用 ChatGPT（OpenAI 的“GPT-4”版本）[1]评估 AutoML-GPT 的性能，并通过它实现该方法。我们开展了多个案例研究，从多个角度展示我们方法的有效性。

3.1 未见过的数据集

在图4中，我们展示了使用 AutoML-GPT 在未见过的数据集上进行训练的结果。为验证其在实际场景中的性能，我们整理了一组已训练数据集和一些尚未训练的数据集的性能及超参数信息。我们将为这些未训练的数据集预测超参数配置。我们的测试环境基于 Vinyals 等人 [2016]所述的分类设置。我们还遵循 MiniImageNet [Vinyals 等人， 2016] 对训练数据集 [Deng 等人， 2009] 进行子采样，并将其划分为 80% 和 20% 两部分。我们利用 80% 的数据构建数据卡和相应的模型卡（其中包含模型的最佳超参数）。我们随机选择十五个类别来创建各种子集数据集（例如，数据集A、B等），通过网格搜索确定超参数，对 ViT base 模型 [Dosovitskiy 等人， 2020] 进行微调，并记录这些子集数据集上的最佳性能。随后，我们利用剩余 20% 数据中的十个图像类别，创建一个名为“New”的数据集。

[1]https://platform.openai.com/

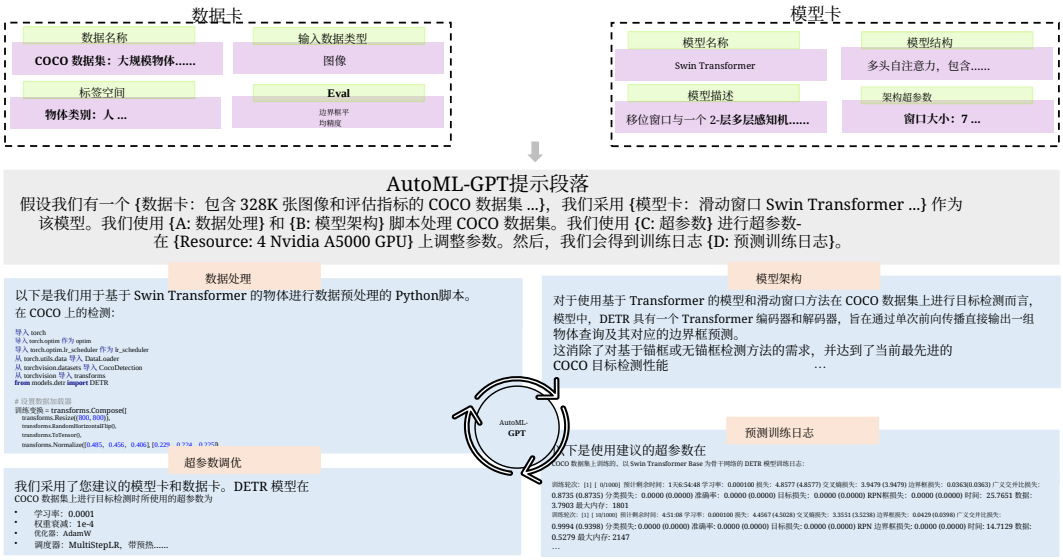


图5：目标检测版 AutoML-GPT 概览：顶部模块展示数据卡和模型卡。中间模块展示AutoML-GPT提示段落，该段落由数据卡和模型卡生成。底部模块概述四个步骤：数据处理、模型架构、超参数调优和预测训练日志。我们利用预测训练日志调整超参数，然后将超参数反馈给用户。

为展示我们的方法在未见过的数据集上的能力，我们使用 AutoML-GPT，根据提供的数据卡和模型卡，为“New”数据集推荐最佳训练配置。在数据卡中，我们记录标签空间、$i.e.$以及每个类别的文本描述。在实际应用中，我们会通过文本编码器（例如CLIP文本编码器）对数据卡中的文本进行编码，并计算相似度，从而得出两张数据卡之间的相似度分数。具体而言，在图4中，我们指出“New”数据集与数据集A的标签空间相似度为 60%，与数据集B的标签空间相似度为 40%。利用这些信息以及数据集A和B的数据卡中的超参数设置，AutoML-GPT 可以为“New”数据集上的训练推荐合适的超参数设置。在我们的实验中，Top 1预测的准确率达到 98%，而采用随机选择的超参数的平均 Top 1 准确率为 80%。此外，即使不提供任何额外数据集，我们也使用 AutoML-GPT 建议的超参数设置初始化模型。采用此配置后，我们取得了 82% 的 Top 1 准确率，优于随机选择的超参数的平均表现，但不如我们推荐的设置。这也表明，ChatGPT 能够为特定任务（$e.g.$图像分类）提供良好的超参数设置。这证明了我们提出的自动训练方法能够有效解决机器学习问题，即使面对未见过或新的数据集也不例外。这些发现凸显了我们自动训练方法通过提供准确的超参数推荐来提升机器学习的潜力。

3.2 目标检测

图5展示了我们在 COCO 数据集上的目标检测结果 [Lin 等人， 2014] 。 ① 顶部区块展示了根据用户输入生成的 COCO 数据集数据卡和 ImageNet 模型卡。中间区块展示了根据输入分解生成的AutoML-GPT提示段落。数据卡和模型卡中的信息会自动整合到我们的提示词格式中。我们报告了数据处理、模型架构设计、超参数调优和训练日志生成的结果。 ② 在数据处理环节，AutoML-GPT 会生成用于处理输入数据集的脚本。我们还在图5中提供了一个 Python 脚本示例。在模型架构设计环节，我们的流程会生成供后续训练使用的模型组合。数据和模型准备就绪后，我们会在超参数调优阶段提供详细配置（$e.g.$，学习率： 10^{-4}，权重衰减： 10^{-4}），并根据预测的训练日志进一步调优。 ③这些结果进一步验证了我们的方法能够作为

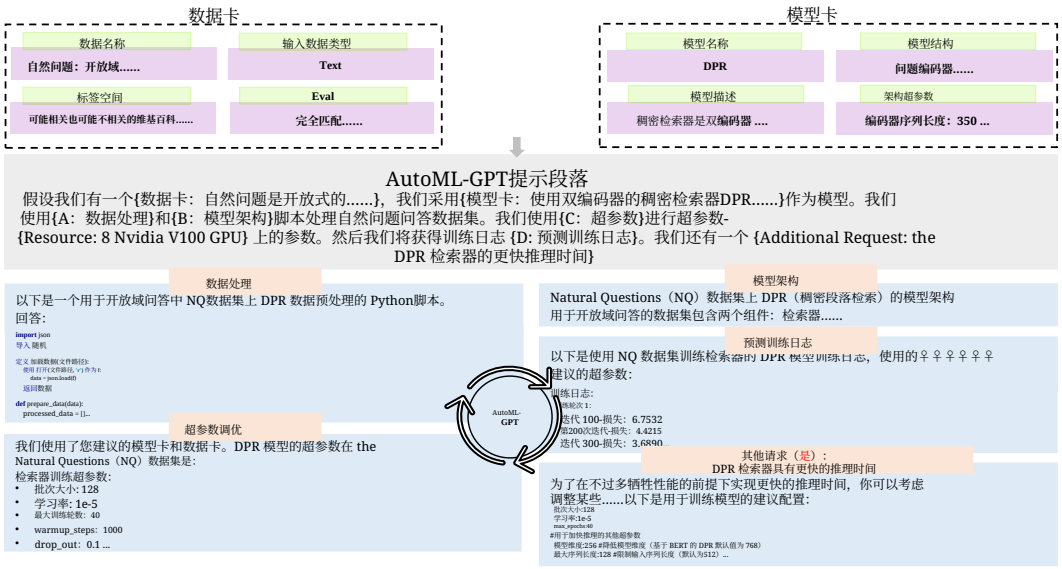


图 6：用于问答的 AutoML-GPT 概览：顶部模块展示数据卡和模型信息，中部模块突出显示由数据卡和模型卡共同生成的 AutoML-GPT提示段落。底部模块详细介绍四个步骤：数据处理、模型架构、超参数调优和预测训练日志。

一种有效的流程，灵活地将大语言模型适配到下游任务。我们的方法利用数据卡和模型卡生成 AutoML-GPT 提示段落，也可视为专注于增强大语言模型提示词组件的工作的补充模块。

3.3 问答

我们在 Natural Questions Open 数据集 [Kwiatkowski 等人， 2019] 上的实验结果如图 6 所示。我们采用稠密段落检索（DPR） [Karpukhin 等人， 2020]。 ① 对于数据卡，用户输入数据名称、输入数据类型、标签空间和评估指标。 ② 对于模型卡，其中包括模型名称、模型结构、模型描述和架构超参数。 ③ 借助生成的 AutoML-GPT提示段落，AutoML-GPT 执行数据处理、创建模型架构、超参数调优，并生成预测训练日志。如“超参数调优”部分所示，AutoML-GPT 生成的超参数与 DPR 提供的超参数高度一致， $e.g.$，学习率为 10^{-5}，最大训练轮次为 40。 ④ 获得预测训练日志后，我们展示了这样一种场景：用户可以根据需求向 AutoML-GPT 请求不同的评估指标或模型架构，如图 6“其他请求：DPR 检索器具有更快的推理时间”所示。从图 6 返回的响应中可以看出，AutoML-GPT 还会提供诸如“不会牺牲太多性能”之类的提示。AutoML-GPT 会根据这些请求和预测日志进一步调整超参数。我们的方法展现了自动开展实验并执行交互式超参数调优的强大能力。这进一步证实了我们的方法适用于各种数据集，并能够在不同输入类型和领域之间实现泛化。

3.4 分类

我们还在 UCI 成人数据集 [Dua and Graff, 2017] 上评估 AutoML-GPT，并使用 XGBoost。与之前一样，我们提供数据卡和模型卡，以生成输入提示词段落。此处采用相同的训练流程，如图 7 所示。我们还遵循 AutoML-GPT 建议的超参数设置，并训练 XGBoost 模型。训练后的最终验证损失为 0.277，准确率为 85.92%。尽管输入和任务各不相同，我们提出的 AutoML-GPT 仍能在分类任务中持续展现出色的性能。这进一步表明，AutoML-GPT 可用于广泛的机器

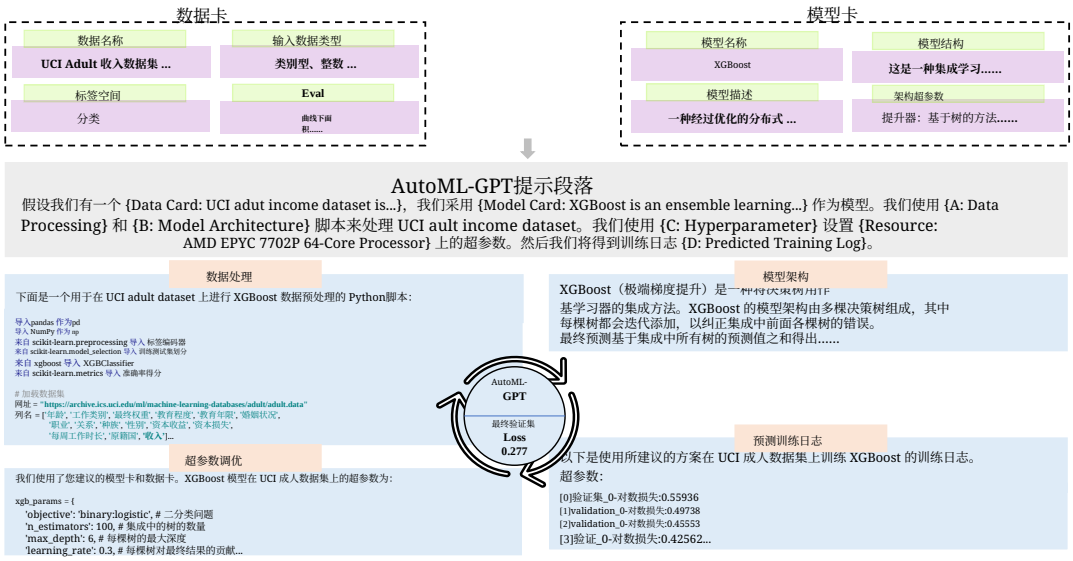


图 7：用于分类的 AutoML-GPT 概览：顶部区块展示数据卡和模型信息，中间区块展示 AutoML-GPT 提示段落，该段落根据数据卡和模型卡生成。底部区块概述了四个步骤：数据处理、模型架构、超参数调优和预测训练日志。此外，我们还展示了遵循 AutoML-GPT 的超参数推荐并训练模型后得到的最终验证结果。

学习问题，涵盖各种任务。

4 相关工作

先进的大语言模型大语言模型通过零样本学习和少样本学习展现出了强大的鲁棒性和泛化能力，其参数规模已超过千亿。大语言模型的典型代表包括拥有 5300 亿参数的 Megatron-turing NLG [Smith 等人， 2022] 拥有 2800 亿参数的 Gopher [Rae 等人，2021] 以及拥有 5400 亿参数的 PaLM [Chowdhery 等人， 2022] 。大语言模型的规模扩展解锁了此前在较小规模模型中未曾观察到的新兴能力 [Wei 等人，2022a]。这些大语言模型已经证明了其在零样本学习方面的卓越性能。在现有的大语言模型中，ChatGPT 具有独特的特点。它能够以类似对话的方式与用户互动，同时保留通过预训练积累的知识和泛化能力。在此基础上，本研究进一步探索 ChatGPT 在对话之外的不同任务中的零样本学习能力。

思维链 思维链（思维链）提示词促使大语言模型在回答 [Wei et al., 2022b]之前生成中间推理步骤。目前，关于思维链提示词的研究主要有两个方向。一条方向探索人工设计的思维链。在人工设计的思维链中，大语言模型会针对推理过程调整人工设计的特征和示例 [Wei et al., 2022b]。Wang et al.[2022] 提出了一种新的解码策略——自洽性，用于替代思维链提示词中使用的朴素贪心解码。最近，Interactive-Chain-Prompting [Pilault et al.，2023] 被提出，用于解决跨语言条件生成中的歧义。另一条方向是在零样本设置下开展研究，其中，STaR [Zelikman et al., 2022] 被引入用于自生成，并帮助模型自我改进；自动推理与工具使用（ART） [Paranjape et al., 2023] 则是一个利用冻结的大语言模型自动生成中间推理步骤作为程序的框架。

基于 GPT 的系统 GPT [Brown 等人， 2020] 已展现出令人瞩目的性能提升。近期的一系列研究主要聚焦于将 GPT 模型集成到人工智能系统中。HuggingGPT [Shen 等人，

2023] 基于 HuggingFace transformers 库构建，并使用 GPT 作为交互代理。VisualGPT [Wu 等人， 2023]整合了不同的视觉基础模型，使用户能够与 ChatGPT 进行交互。OpenAGI [Ge 等人， 2023], 是一个开源 AGI 研究平台，旨在提供复杂的多步骤任务，并配备特定任务的数据集。同样，我们也将 GPT 集成到自动机器学习流水线中。另有一种基于 GPT 的系统，可以整合来自搜索引擎的额外信息，即 $e.g.$、AutoGPT [2]。AutoML-GPT 从自动训练的角度重新审视了 ChatGPT 的影响。我们专注于构建训练流程，并从头到尾建立一个 AutoML 系统。

5 结论

我们的工作展示了基于 GPT 构建 AutoML 系统的优势。所提出的方法能够自动进行机器学习实验。这种自动化学习大幅提高了训练效率并提升了模型的性能。我们展示了该方法在计算机视觉、自然问题回答和分类基准测试中的应用案例。此外，我们还针对未见过的数据集进行了详细的应用案例研究，并展示了用户与 AutoML-GPT 之间的额外交互。总而言之，所提出的 AutoML-GPT 高效且通用，有望为各种任务中的机器学习模型调优打造自然语言接口。未来，我们将：1）为知名基准自动生成模型卡和数据卡，并将其纳入系统；2）借助 ChatGPT，从大型预训练模型中提取任务感知型子网络。

参考文献

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared J Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。2022。PaLM：通过路径扩展语言建模。*arXiv* 预印本 *arXiv:2204.02311*。

Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。2022。扩展经过指令微调的语言模型。*arXiv* 预印本 *arXiv:2210.11416*。

Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。2009。ImageNet：大规模分层图像数据库。载于*2009* 年 *IEEE* 计算机视觉与模式识别会议，第 248–255 页。IEEE。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020年。图像胜过 16×16 个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*。

Dheeru Dua 和 Casey Graff。2017年。UCI 机器学习资料库。

Yingqiang Ge、Wenyue Hua、Jianchao Ji、Juntao Tan、Shuyuan Xu 和 Yongfeng Zhang。2023年。OpenAGI：当大语言模型遇上领域专家。*arXiv* 预印本 *arXiv:2304.04370*。

Timnit Gebru、Jamie Morgenstern、Briana Vecchione、Jennifer Wortman Vaughan、Hanna Wallach、Hal Daumé Iii 和 Kate Crawford。2021年。数据集的数据说明书。*ACM* 通讯，64(12):86–92。

[2]https://github.com/Significant-Gravitas/Auto-GPT

Gautier Izacard 和 Edouard Grave。2020年。利用生成模型进行段落检索，以实现开放域问答。*arXiv* 预印本 *arXiv:2007.01282*。

Vladimir Karpukhin、Barlas O˘guz、Sewon Min、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。面向开放域问答的稠密段落检索。自然语言处理实证方法（*EMNLP*）。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Redfield、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Matthew Kelcey、Jacob Devlin、Kenton Lee、Kristina N. Toutanova、Llion Jones、Ming-Wei Chang、Andrew Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。Natural Questions：问答研究的一个基准。*TACL*。

Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Dollár 和 C Lawrence Zitnick。2014年。Microsoft coco：上下文中的常见物体。载于计算机视觉–*ECCV 2014*：第*13*届欧洲会议，苏黎世，瑞士，*2014*年*9*月*6*日至*12*日，论文集，第*V*卷第*13*册，第740–755页。Springer。

Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。2021年。Swin Transformer：使用移位窗口的分层视觉变换器。载于*I EEE/CVF* 国际计算机视觉会议论文集，第 10012–10022 页。

Margaret Mitchell、Simone Wu、Andrew Zaldivar、Parker Barnes、Lucy Vasserman、Ben Hutchinson、Elena Spitzer、Inioluwa Deborah Raji 和 Timnit Gebru。2019年。用于模型报告的模型卡。在公平性、问责性与透明度会议论文集，第 220–229 页。

OpenAI. 2023年。GPT-4 技术报告。*arXiv*。

Bhargavi Paranjape、ScottLundberg、Sameer Singh、Hannaneh Hajishirzi、Luke Zettlemoyer 和 Marco Tulio Ribeiro。2023年。ART：大语言模型的自动多步推理和工具使用。*arXiv* 预印本 *arXiv:2303.09014*。

Jonathan Pilault、Xavier Garcia、Arthur Bražinskas 和 Orhan Firat。2023年。交互式链式提示：通过交互实现跨语言条件生成的歧义消解。*arXiv* 预印本 *arXiv:2301.10309*。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021年。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR。

Jack W Rae、Sebastian Borgeaud、Trevor Cai、Katie Millican、Jordan Hoffmann、Francis Song、John Aslanides、Sarah Henderson、Roman Ring、Susannah Young 等。2021年。扩展语言模型：Gopher 训练的方法、分析与见解。*arXiv* 预印本 *arXiv:2112.11446*。

Ori Ram、Yoav Levine、Itay Dalmedigos、Dor Muhlgay、Amnon Shashua、Kevin Leyton-Brown 和 Yoav Shoham。2023年。上下文检索增强语言模型。*arXiv* 预印本 *arXiv:2302.00083*。

Yongliang Shen、Kaitao Song、Xu Tan、Dongsheng Li、Weiming Lu 和 Yueting Zhuang。2023年。HuggingGPT：使用 ChatGPT 及其在 Hugging Face 上的好友解决人工智能任务。*arXiv* 预印本 *arXiv:2303.17580*。

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti 等。2022年。使用 DeepSpeed 和 Megatron 训练 Megatron-turing NLG 530B：一个大规模生成式语言模型。*arXiv* 预印本 *arXiv:2201.11990*。

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar 等。2023年。Llama：开放且高效的基础语言模型。*arXiv* 预印本 *arXiv:2302.13971*。Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra 等。2016年。用于单样本学习的匹配网络。神经信息处理系统进展，29。Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi 和 Denny Zhou。2022年。自洽性提升了语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler 等。2022a。大语言模型涌现能力。*arXiv* 预印本 *arXiv:2206.07682*。Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le 和 Denny Zhou。2022b。思维链提示能激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2201.11903*。

Chenfei Wu、Shengming Yin、Weizhen Qi、Xiaodong Wang、Zecheng Tang 和 Nan Duan。2023年。Visual ChatGPT：借助视觉基础模型进行对话、绘图与编辑。*arXiv* 预印本 *arXiv:2303.04671*。

Tete Xiao、Yingcheng Liu、Bolei Zhou、Yuning Jiang 和 Jian Sun。2018年。用于场景理解的统一感知解析。在欧洲计算机视觉会议（*ECCV*）论文集中，第418–434页。Sang Michael Xie、Aditi Raghunathan、Percy Liang 和 Tengyu Ma。2021年。将上下文学习解释为隐式贝叶斯推断。*arXiv* 预印本 *arXiv:2111.02080*。Eric Zelikman、Yuhuai Wu、Jesse Mu 和 Noah Goodman。2022年。STaR：通过推理进行自举推理。神经信息处理系统进展，35：15476–15488。Shujian Zhang、Chengyue Gong 和 Eunsol Choi。2021年。更深入地了解问题会有所帮助：改进问答中的校准。*arXiv* 预印本 *arXiv:2106.01494*。Shujian Zhang、Chengyue Gong 和 Xingchao Liu。2022年。段落掩码：一种适用于检索器-阅读器模型的可学习正则化策略。*arXiv* 预印本 *arXiv:2211.00915*。Mingkai Zheng、Xiu Su、Shan You、Fei Wang、Chen Qian、Chang Xu 和 Samuel Albanie。2023年。GPT-4 能否执行神经架构搜索？*arXiv* 预印本 *arXiv:2304.10970*。