# 澳鹏appen\_全球AI训练数据服务领军者 | 数据标注与采集

2023年，大模型成为了重要话题，每个行业都在探索大模型的应用落地，以及其能够如何帮助到企业自身。尽管微软、OpenAI、百度等公司已经在创建并迭代大模型并探索更多的应用，对于大部分企业来说，都没有足够的成本来创建独特的基础模型（Foundation Model）：数以百亿计的数据以及超级算力资源使得基础模型成为一些头部企业的“特权”。 然而，无法自己创建基础模型，并不代表着大模型无法为大部分公司所用：在大量基础模型的开源分享之后，企业可以使用微调（Fine tuning）的方法，训练出适合自己行业和独特用例的大模型以及应用。 本文即将讨论大模型微调的定义，重要性，常见方法，流程，以及澳鹏如何帮助您利用大模型进行应用落地。 什么是大模型微调？ --------- 大模型微调（Fine-tuning）是指在已经预训练好的大型语言模型基础上，使用特定的数据集进行进一步的训练，以使模型适应特定任务或领域。 其根本原理在于，机器学习模型只能够代表它所接收到的数据集的逻辑和理解，而对于其没有获得的数据样本，其并不能很好地识别/理解，且对于大模型而言，也无法很好地回答特定场景下的问题。 例如，一个通用大模型涵盖了许多语言信息，并能够进行流畅的对话。但是如果需要医药方面能够很好地回答患者问题的应用，就需要为这个通用大模型提供很多新的数据以供学习和理解。例如，布洛芬到底能否和感冒药同时吃？为了确定模型可以回答正确，我们就需要对基础模型进行微调。 为什么大模型需要微调？ ----------- 预训练模型（Pre-trained Model），或者说基础模型（Foundation Model），已经可以完成很多任务，比如回答问题、总结数据、编写代码等。但是，并没有一个模型可以解决所有的问题，尤其是行业内的专业问答、关于某个组织自身的信息等，是通用大模型所无法触及的。在这种情况下，就需要使用特定的数据集，对合适的基础模型进行微调，以完成特定的任务、回答特定的问题等。在这种情况下，微调就成了重要的手段。 大模型微调的两个主要方法 ------------ 我们已经讨论了微调的定义和重要性，下面我们介绍一下两个主要的微调方法。根据微调对整个预训练模型的调整程度，微调可以分为全微调和重用两个方法： 1. 全微调（Full Fine-tuning）：全微调是指对整个预训练模型进行微调，包括所有的模型参数。在这种方法中，预训练模型的所有层和参数都会被更新和优化，以适应目标任务的需求。这种微调方法通常适用于任务和预训练模型之间存在较大差异的情况，或者任务需要模型具有高度灵活性和自适应能力的情况。Full Fine-tuning需要较大的计算资源和时间，但可以获得更好的性能。 2. 部分微调（Repurposing）：部分微调是指在微调过程中只更新模型的顶层或少数几层，而保持预训练模型的底层参数不变。这种方法的目的是在保留预训练模型的通用知识的同时，通过微调顶层来适应特定任务。Repurposing通常适用于目标任务与预训练模型之间有一定相似性的情况，或者任务数据集较小的情况。由于只更新少数层，Repurposing相对于Full Fine-tuning需要较少的计算资源和时间，但在某些情况下性能可能会有所降低。 选择Full Fine-tuning还是Repurposing取决于任务的特点和可用的资源。如果任务与预训练模型之间存在较大差异，或者需要模型具有高度自适应能力，那么Full Fine-tuning可能更适合。如果任务与预训练模型相似性较高，或者资源有限，那么Repurposing可能更合适。在实际应用中，根据任务需求和实验结果，可以选择适当的微调方法来获得最佳的性能。 大模型微调的两个主要类型 ------------ 同时，根据微调使用的数据集的类型，大模型微调还可以分为监督微调和无监督微调两种： 1. 监督微调（Supervised Fine-tuning）：监督微调是指在进行微调时使用有标签的训练数据集。这些标签提供了模型在微调过程中的目标输出。在监督微调中，通常使用带有标签的任务特定数据集，例如分类任务的数据集，其中每个样本都有一个与之关联的标签。通过使用这些标签来指导模型的微调，可以使模型更好地适应特定任务。 2. 无监督微调（Unsupervised Fine-tuning）：无监督微调是指在进行微调时使用无标签的训练数据集。这意味着在微调过程中，模型只能利用输入数据本身的信息，而没有明确的目标输出。这些方法通过学习数据的内在结构或生成数据来进行微调，以提取有用的特征或改进模型的表示能力。 监督微调通常在有标签的任务特定数据集上进行，因此可以直接优化模型的性能。无监督微调则更侧重于利用无标签数据的特征学习和表示学习，以提取更有用的特征表示或改进模型的泛化能力。这两种微调方法可以单独使用，也可以结合使用，具体取决于任务和可用数据的性质和数量。 大模型微调的主要步骤 ---------- 大模型微调如上文所述有很多方法，并且对于每种方法都会有不同的微调流程、方式、准备工作和周期。然而大部分的大模型微调，都有以下几个主要步骤，并需要做相关的准备： 1. 准备数据集：收集和准备与目标任务相关的训练数据集。确保数据集质量和标注准确性，并进行必要的数据清洗和预处理。 2. 选择预训练模型/基础模型：根据目标任务的性质和数据集的特点，选择适合的预训练模型。 3. 设定微调策略：根据任务需求和可用资源，选择适当的微调策略。考虑是进行全微调还是部分微调，以及微调的层级和范围。 4. 设置超参数：确定微调过程中的超参数，如学习率、批量大小、训练轮数等。这些超参数的选择对微调的性能和收敛速度有重要影响。 5. 初始化模型参数：根据预训练模型的权重，初始化微调模型的参数。对于全微调，所有模型参数都会被随机初始化；对于部分微调，只有顶层或少数层的参数会被随机初始化。 6. 进行微调训练：使用准备好的数据集和微调策略，对模型进行训练。在训练过程中，根据设定的超参数和优化算法，逐渐调整模型参数以最小化损失函数。 7. 模型评估和调优：在训练过程中，使用验证集对模型进行定期评估，并根据评估结果调整超参数或微调策略。这有助于提高模型的性能和泛化能力。 8. 测试模型性能：在微调完成后，使用测试集对最终的微调模型进行评估，以获得最终的性能指标。这有助于评估模型在实际应用中的表现。 9. 模型部署和应用：将微调完成的模型部署到实际应用中，并进行进一步的优化和调整，以满足实际需求。 这些步骤提供了一个一般性的大模型微调流程，但具体的步骤和细节可能会因任务和需求的不同而有所变化。根据具体情况，可以进行适当的调整和优化。 然而，虽然微调相对于训练基础模型，已经是相当省时省力的方法，但是微调本身还是需要足量的经验和技术，算力，以及管理和开发成本。为此，澳鹏已经推出一系列定制化服务和产品，助您轻松拥抱大模型。 澳鹏帮助您轻松拥抱大模型 ------------ 澳鹏为所有希望进军大语言模型应用的企业，提供一系列定制化服务及产品： 1. 数据清洗、数据集、采标定制：澳鹏作为人工智能数据行业超过26年的全球领军人，在235+种语言方言方面有深入的研究和大量的数据经验，可以为您提供您需要的使用场景中所需的[多语言数据](https://www.appendata.com/datasets)、定制化[采集](https://www.appendata.com/data-collection)标注、以及多层次[详细标注](https://www.appendata.com/data-annotation)，为您的LLM训练提供强大的数据后盾。 2. 微调/RLHF：拥有全球超过100万的众包及强大的合作标注团队、经验丰富的管理团队，我们可以为您的模型微调提供巨量的[RLHF支持](https://www.appendata.com/intelligent-llm-development-platform)，最大程度减少幻觉（hallucination）的干扰。 3. [LLM智能开发平台](https://www.appendata.com/intelligent-llm-development-platform)：由于大语言模型的应用开发，除了训练和微调之外，还需要多方面的开发流程，以提高开发效率、减少开发阻碍。澳鹏自主开发的[LLM智能开发平台](https://www.appendata.com/intelligent-llm-development-platform)，为您提供多层次、多方面的开发者工具，助您快速训练、部署LLM程序。 4. LLM应用定制服务：同时，对于没有开发能力的企业，我们强大的数据团队、算法团队，提供全面的[定制服务](https://www.appendata.com/contact?from=%2Fblogs%2Ffine-tuning&)。根据您的用例和需求，选择合适的基础模型，并使用最合适的数据进行微调，最后为您部署出您想要的LLM应用。 如想进一步了解澳鹏能够为您的LLM应用提供哪些支持，或有相关需求，可以[联系我们](https://www.appendata.com/contact?from=%2Fblogs%2Ffine-tuning&)，我们的专家团队会为您提供可行建议，或给出服务报价。 澳鹏支持全栈式大模型数据服务，包括数据集，模型评估，模型调优；同时，澳鹏智能大模型开发平台与全套标注工具支持您快速部署大模型应用。