cs336 lec15

liuyang2967299295

July 2025

语言模型的训练，pre-training 过程中，通常只是对语言模型进行了建模，但模型却不能很好的控制它的行为，包括指令遵循与安全与内容审查。

- 当前的困境（Problem Statement）：
  "Pretraining data isn't quite what we want (but it scales).."
  当前的预训练数据虽然可以大规模获取和使用("scales"，即规模化、可扩展)，但它并不能完全满足我们对模型输出行为的精确控制需求。这意味着预训练数据可能包含太多我们不想要的行为，或者无法确保模型总能产生我们期望的输出。它好在“量大”，但不好在“质不符我们心意”。
- 提出的解决方案/核心问题(Proposed Solution/Core Question)：
  "Can we collect data of behaviors we do want and train the LM?"
  我们是否可以通过收集那些我们明确期望模型展现的特定行为数据，并用这些数据来训练（或微调）语言模型?
- 需要思考的三个子问题（Three Key Questions to Address)：
  - 1. What does that data look like?
    理解：这种“我们期望的行为数据”具体是什

么样的？这指的是这种数据应该以何种形式存在？是正面示例（Positive Examples）？反面示例（Negative Examples）？还是某种带有特定标签或指令的数据？如何定义和捕捉这些“期望的行为”？例如，是用户对模型输出的评分，还是人类专家对某个输出的编辑和修正？

– 2. How do we best make use of that data?
理解：我们如何才能最好地利用这些数据来训练模型？收集到数据后，最佳的利用方式是什么？是进行模型微调（Fine-tuning）？还是通过强化学习（Reinforcement Learning from Human Feedback, RLHF）？或者有其他更有效的方法能让模型学习到这些期望的行为，并将其转化为实际的输出控制？

– 3. Do we need scale for this?
理解：为了达到目标，我们是否也需要大规模地收集这种“期望行为数据”？这个问题探讨的是，为了达到对模型输出的有效控制，我们是否也需要像预训练数据那样大规模地收集这种“期望行为数据”？还是说，少量的、高质量的、精心策划的数据就能达到目标？这关系

到数据收集的成本和可行性。

这边给出一个标准做法，SFT 后跟 RL。

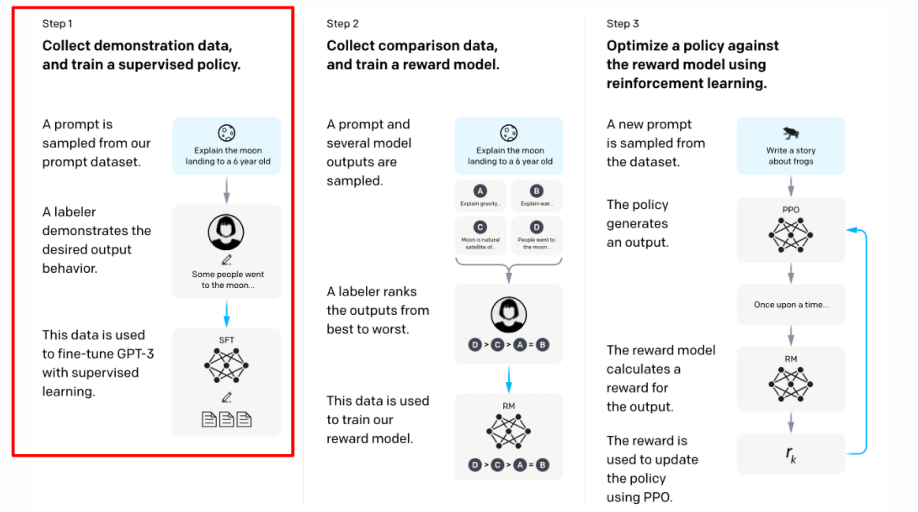

图 1: the ‘supervised finetuning’ part

SFT

SFT (Supervised Fine-Tuning) 是预训练模型之后的一个重要步骤，通过有标签数据对模型进行微调，使其更好地适应特定任务。

SFT 的关键点主要在于训练数据与数据使用策略, 下面将介绍几个常见的监督数据集:

训练数据

- TO-SF (Tasks for Supervised Fine-tuning)

内容：涵盖了常识推理、问题生成、闭卷问答(QA)、对抗性问答、抽取式问答、标题/上下文生成、主题分类、结构化文本转换等广泛的自然语言处理任务。

规模：包含 55 个数据集，14 个类别，193 项任务。
解读：这是一个非常广泛和通用的 NLP 任务集合，旨在为模型提供全面的基础能力。

- Muffin

  内容：包括自然语言推理、代码指令生成、程序合成、对话上下文生成、闭卷问答、对话问答、代码修复等。

  规模：包含 69 个数据集，27 个类别，80 项任务。

  解读：yi 这个集合在通用 NLP 任务的基础上，更侧重于代码理解与生成以及对话能力。

- CoT (Reasoning) (链式思考/推理)

  内容：专注于推理能力，包括算术推理、常识推理、隐含推理、解释生成、句子组合等。

  规模：包含 9 个数据集，1 个类别，9 项任务。

  解读：CoT (Chain-of-Thought) 是当前大型语言模型提升推理能力的关键技术。这个集合虽然规模较小，但其任务类型高度集中于逻辑推理和逐步思考，对于模型生成高质量、有逻辑的回答至关

重要。

- Natural Instructions v2

  内容：涵盖因果效应分类、常识推理、命名实体识别、毒性语言检测、问答、问题生成、程序执行、文本分类等。

  规模：包含 372 个数据集，108 个类别，1554 项任务。

  解读：这是所有微调任务中规模最大、种类最丰富的集合。它强调通过统一的“指令”格式来训练模型执行各种任务，极大地提升了模型的指令遵循能力和泛化性。

数据使用策略

模型训练往往不是一次性完成的，而是分阶段进行的，每个阶段的数据配比有所不同。以 Open Assistant 为例，在“稳定阶段”，模型主要通过通用的大规模数据和基础指令数据进行训练，建立广泛的能力。在“衰减阶段”，则逐渐减少通用数据的比例，增加更多样化、高质量、领域特定的数据，以进一步提升模型的专业能力和对话质量。

- Data Mixture of Stable Stage (稳定阶段的数据混合)

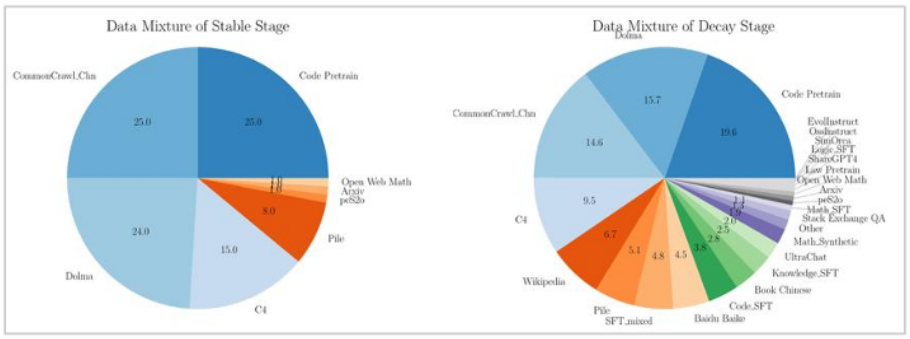

图 2: Open Assistant 数据使用策略图

主要组成:

- CommonCrawl.Chn (25.0%):通用网络爬取数据(中文)。
- Code Pretrain (25.0%): 代码预训练数据。
- Dolly (24.0%): 一个高质量的指令遵循数据集,通常包含用户与模型的对话。
- C4 (15.0%): 大规模通用网络文本数据集。

解读: 这个阶段的数据混合相对平衡,主要由大规模通用网络数据(中英文)、代码数据和高质量的指令微调数据组成。这可能是模型训练的初始或基础阶段,旨在建立模型广泛的语言理解、生成和初步的指令遵循能力。

- Data Mixture of Decay Stage (衰减阶段的数据混合)

  主要组成:

- CommonCrawl.Chn (15.7%), Code Pretrain (19.6%), C4 (9.5%)：这些通用数据比例有所下降。
- 新增或显著增加的细分数据：Wikipedia (6.7%), Pif (5.1%)，Baidu Baike (4.8%)，Book Chinese (4.5%)，以及许多更小、更专业的 SFT 数据集，如 Evollstruct, Oasst, LogiSFT, MSET, Ultra-Chat, Knowledge SFT, Stack Exchange QA, Math Synthetic 等。

解读：这个阶段的数据混合呈现出通用数据比例降低，而高质量、特定领域、多样化小数据集比例显著增加的特点。这个阶段的调整旨在提升模型在特定领域的知识、专业技能（如数学、编程、知识问答）以及在高质量对话场景中的表现。这种策略有助于防止模型在通用数据上过度拟合，并使其在更广泛和专业的对话场景中表现更出色。

数据特性是塑造模型行为和性能的根本

训练数据的内在特性，尤其是其“风格”方面（如输出长度和格式），将深刻地影响大型语言模型的表现、用户的偏好以及其评估结果。

- 数据本身存在显著的风格差异：

  训练数据本身就呈现出多种风格变异，例如在长度、列表使用、引用复杂性、数据规模和安全性等

What about benchmarks?

These factors are (mostly) not that relevant for other benchmark perfs

Table 3: Comparison of different instruction tuning datasets, showing that different instruction-tuning datasets can excel in different aspects, and mixtures perform best on average. Cells are blue if the finetuning boosts the vanilla LLAMA performance, and orange if the finetuning hurts the performance.

	MMLU (factuality)	GSM (reasoning)	BBH (reasoning)	TydiQA (multilinguality)	Codex-Eval (coding)	AlpacaEval (open-ended)	Average
	EM (0-shot)	EM (8-shot, CoT)	EM (3-shot, CoT)	F1 (1-shot, GP)	P@10 (0-shot)	Win % vs Davinci-003	
Vanilla LLaMa 13B	42.3	14.5	39.3	43.2	28.6	-	-
+SuperNI	49.7	4.0	4.5	**50.2**	12.9	4.2	20.9
+CoT	44.2	40.0	41.9	47.8	23.7	6.0	33.9
+Flan V2	**50.6**	20.0	40.8	47.2	16.8	3.2	29.8
+Dolly	45.6	18.0	28.4	46.5	31.0	13.7	30.5
+Open Assistant 1	43.3	15.0	39.6	33.4	31.9	58.1	36.9
+Self-instruct	30.4	11.0	30.7	41.3	12.5	5.0	21.8
+Unnatural Instructions	46.4	8.0	33.7	40.9	23.9	8.4	26.9
+Alpaca	45.0	9.5	36.6	31.1	29.9	21.9	29.0
+Code-Alpaca	42.5	13.5	35.6	38.9	34.2	15.8	30.1
+GPT4-Alpaca	46.9	16.5	38.8	23.5	**36.6**	63.1	37.6
+Baize	43.7	10.0	38.7	33.6	28.7	21.9	29.4
+ShareGPT	49.3	27.0	40.4	30.5	34.1	**70.5**	42.0
+Human data mix.	50.2	38.5	39.6	47.0	25.0	35.0	39.2
+Human+GPT data mix.	49.3	**40.5**	**43.3**	45.6	35.9	56.5	**45.2**

图 3: 多种指令微调数据集在多个基准测试上的表现

方面。不同训练数据集的统计数据（特别是平均响应长度），本身就存在巨大的风格多样性。这意味着模型在学习过程中，会吸收并反映其训练数据固有的长度、结构和表达习惯等“风格”特征。

- 数据风格影响评估和偏好：

  无论是人类还是基于 AI 的评估者，在判断模型输出质量时，都会受到输出“风格”的影响，例如内容的长度和是否采用列表格式。这表明，数据中包含的风格偏好，会直接塑造模型最终的输出风格，进而影响其被用户接受和评价的程度。

综上所述，在这里提出一个洞察：大型语言模型不仅仅是学习了数据中的知识和信息，更深层次地，它们

也继承了训练数据的 “风格”。因此，数据中不同风格和内容特性的存在，是决定模型输出质量、行为模式以及其在评估中表现的关键因素。

幻觉与安全