Contents

下载本文 pdf：https://github.com/daiwk/collections/blob/master/pdfs/llm_aigc.pdf

各种学习相关代码

https://github.com/daiwk/llms

1 概述

1.1 LLM 简史

- 2017 年的Learning to generate reviews and discovering sentiment尝试用 rnn 来实现智能系统
- 2018 年的 gpt1：Improving language understanding by generative pre-training，生成式预训练（Generative pre-training, gpt），用 transformer 的 decoder，参数量 117m（0.1b），无监督预训练和有监督微调。确定对自然语言文本建模的基本原则为**预测下一个单词**。
- 2019 年的 gpt2：Language models are unsupervised multitask learners模型结构小改，增加数据，参数量变大为 15 亿（1.5b），无监督语言建模，**无需使用标记数据进行显式微调**。
  - 参考The natural language decathlon: Multitask learning as question answering中**多任务求解的概率形式**：$p(output|input, task)$。
  - 提出“由于特定任务的有监督目标与无监督目标（语言建模）相同，只是在序列的子集上进行评估，因此，无监督目标的全局最小值也是有监督目标的全局最小值”，即每个 NLP 任务可以看作**世界文本子集的单词预测问题**，如果模型有足够能力来复原世界文本，无监督语言建模可以解决各种问题。
  - 仅无监督与监督微调的 SOTA 相比效果还是不太行。虽然 GPT2 模型规模相对较小，但如对话等任务在其基础上做微调还是能拿到很好的效果的，例如DIALOGPT : Large-scale generative pre-training for conversational response generation、End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2
- 2020 年的 gpt3：Language models are few-shot learners，175b（1750 亿）参数，当参数量到达千亿时出现了『涌现』现象，发现可以 in-context learning（这点在 **3.3 亿的 BERT 和 15 亿的 gpt2 中看不到**）。**预训练和 ICL 有相同的语言建**

模范式：预训练预测给定上下文条件下的后续文本序列，ICL 预测正确的任务解决方案，其可被格式化为给定任务描述和示范下的文本序列。

- GPT-3 的两种改进方法：
  - 使用代码数据训练：GPT-3 主要问题是缺乏对复杂任务的推理能力，2021 年 openai 提出了 Codex（Evaluating Large Language Models Trained on Code），在 github 代码上微调的 GPT。A neural network solves and generates mathematics problems by program synthesis: Calculus, differential equations, linear alge- bra, and more发现 Codex 能解决非常困难的编程问题，还能在数学问题上有显著提升。Text and code embeddings by contrastive pre-training提出了训练文本和代码 emb 的对比学习，在线性探测分类、文本搜索、代码搜索等任务上有所提升。GPT-3.5 就是在基于代码的 GPT（code-davinci-002）的基础上开发的。
  - 与人类对齐：2017 年 openai 就在learning from human preference的博客中提出了应用强化学习来学习由人类标的偏好比较，此后 2021 年 7 月 openai 发表了 PPO。2020 年 GPT-2 用 RL 进行微调，Deep reinforcement learning from human preferences，Learning to summarize from human feedback也做了相似工作。2022 年提出了 RLHF 的 InstructGPT(Training language models to follow instructions with human feedback)，其中的 **SFT 就对应于常说的指令微调**。在 openai 的博客Our approach to alignment research中提出了训练 AI 系统的 3 个有前途的方向：**使用人类反馈、协助人类评估、做对齐研究**。
- 2022 年的 ChatGPT：用类似 InstructGPT 的方式进行训练，专门**对对话能力进行优化**，将人类生成的对话（**扮演用户和 AI 两个角色**）与 InstructGPT 数据集结合起来**以对话形式生成**。
- 2023 年的 GPT-4：将文本输入扩展到**多模态信号**。此外，
  - 提升安全性：在 RLHF 训练中加入**额外的安全奖励信号**，采用多种干预策略如 Anthropic 提出的Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned提到的红队评估（read teaming）机制以减轻幻觉、隐私和过度依赖问题。
  - 改进的优化方法：使用**可预测扩展**（predictable scaling）的机制，使用模型训练期间的一小部分计算量**以预测最终性能**。
  - 迭代部署的工程方案：Lessons learned on language model safety and misuse，遵循 5 阶段的开发和部署生命周期来开发模型和产品。

1.2 一些综述

- Foundation Models for Natural Language Processing -Pre-trained Language Models Integrating Media
- 大规模语言模型：从理论到实践，Pre-trained Models for Natural Language Processing: A Survey邱锡鹏等
- 人大的大模型综述：https://github.com/RUCAIBox/LLMSurvey，自己存了一份 pdf，(!!! **本章大部分内容按这个来组织!!!**)
- Talking about large language models
- Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing，引用数 2k+
- A comprehensive survey on pretrained foundation models: A history from BERT to chatgpt，唐杰等
- Pre-Trained Models: Past, Present and Future
- A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT
- Pretrained Language Models for Text Generation: A Survey
- A survey for in-context learning
- Towards reasoning in large language models: A survey
- Reasoning with language model prompting: A survey
- Dense Text Retrieval based on Pretrained Language Models: A Survey
- Fine-tune 之后的 NLP 新范式：Prompt 越来越火，CMU 华人博士后出了篇综述文章

1.3 扩展法则

1.3.1 openai 的扩展法则

2020 年,openai 的Scaling laws for neural language models通过拟合模型在不同数据大小（2000w 到 230 亿个 token）、不同的模型大小（7.68 亿到 15 亿个**非嵌入参数**）的性能，提出了在**计算预算**c 的条件下，L 是用 nats 表示的交叉熵损失，模型性能与**模型规模**N、**数据集规模**D 以及**训练计算量**C 间存在如下幂律关系：

$$L(N) = (\frac{N_c}{N})^{\alpha_N}, \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13}$$

$$L(D) = (\frac{D_c}{D})^{\alpha_D}, \alpha_D \sim 0.05, N_c \sim 5.4 \times 10^{13}$$

$$L(C) = (\frac{C_c}{C})^{\alpha_C}, \alpha_C \sim 0.05, C_c \sim 3.1 \times 10^{8}$$

其中，N_c 表示非嵌入参数数量，D_c 表示训练 token 数量,C_c 表示 FP-days。

1.3.2 Chinchilla 扩展法则

DeepMind 在Training compute-optimal large language models中提出了 Chichilla 扩展法则来指导 LLM 最优计算量的训练。通过变化更大范围的模型大小（7000w 到 160 亿参数）和数据大小（50 亿到 5000 亿个 token）进行实验，拟合了如下的扩散法则：

$$L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$$

其中 $E = 1.69, A = 406.4, B = 410.7, \alpha = 0.34, \beta = 0.28$，通过在约束条件 $C \approx 6ND$ 下优化损失 $L(N, D)$，将计算预算最优地分配给模型大小和数据大小的方法：

$$N_{opt}(C) = G\left(\frac{C}{6}\right)^a, \quad D_{opt}(C) = G^{-1}\left(\frac{C}{6}\right)^b$$

其中 $a = \frac{\alpha}{\alpha+\beta}$，$b = \frac{\beta}{\alpha+\beta}$，$G$ 是由 A, B, α, β 计算出的扩展系数。

随着计算预算的增加，

- openai 的扩展法则更偏向于将更大预算分给**模型大小**，因为其对比各模型时使用了固定的训练数据量和学习率等超参，低估了数据量的作用。
- Chinchilla 扩展法则认为**模型大小和数据大小要同比例增加**，即 a 和 b 取值差不多。因为其在无视模型大小的前提下，发现设置与数据量差不多 match 的学习率能获得更好的 loss。

然而，有一些能力（如涌现）无法根据扩展法则进行预测，只有当模型达到一定规模时才会出现。

1.4 涌现能力

涌现能力：在小型模型中不存在而在大型模型中产生的能力，当规模达到一定程度时，性能显著提升，超出随机水平（参考 Emergent Abilities of Large Language Models）。与物理学中的**相变**现象类似（物质从一种相（状态）转变为另一种相的过程，通常伴随着能量的吸收或释放，并且涉及不同的物理性质，例如固体、液体和气体之间的转变）。

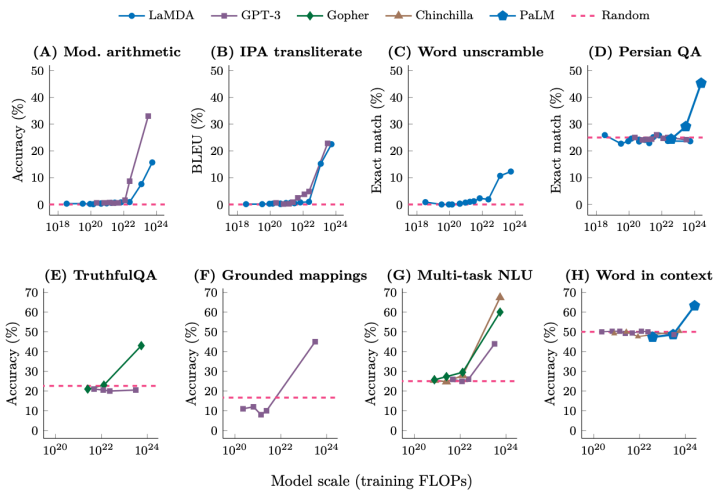


Figure 2: Eight examples of emergence in the few-shot prompting setting. Each point is a separate model. The ability to perform a task via few-shot prompting is emergent when a language model achieves random performance until a certain scale, after which performance significantly increases to well-above random. Note that models that used more training compute also typically have more parameters—hence, we show an analogous figure with number of model parameters instead of training FLOPs as the x-axis in Figure 11. A–D: BIG-Bench (2022), 2-shot. E: Lin et al. (2021) and Rae et al. (2021). F: Patel & Pavlick (2022). G: Hendrycks et al. (2021a), Rae et al. (2021), and Hoffmann et al. (2022). H: Brown et al. (2020), Hoffmann et al. (2022), and Chowdhery et al. (2022) on the WiC benchmark (Pilehvar & Camacho-Collados, 2019).

LLM 的 3 种典型涌现能力及其对应代表模型：

1.4.1 上下文学习

GPT-3（Language models are few-shot learners）提出，只要提供一个自然语言指令和/或几个任务演示，语言模型就能通过完成输入文本的词序列的方式来为测试实例生成预期输出，不用额外的梯度更新。

- ICL 能力小模型不具备：1750 亿的 GPT-3 有 ICL 能力，但 GPT-1 和 GPT-2 无此能力。
- ICL 能力取决于具体下游任务：130 亿的 GPT-3 能在算术任务上有 ICL，但 1750 亿的 GPT-3 在波斯语 QA 上无能为力。

1.4.2 指令遵循

使用**自然语言描述的混合多任务数据集进行微调（指令微调）**，LLM 在**未见过的以指令形式描述的任务**上表现出色，具有更好的泛化能力。例如Multitask prompted training enables zero-shot task generalization、Training language models to follow instructions with human feedback、Finetuned language models are zero-shot learners。

在Finetuned language models are zero-shot learners的实验中，当模型大小达到 680 亿时，经过指定微调的 LaMDA-PT 开始在未见过的任务上显著优于未微调的模型，而 80 亿或更小的模型则没有这个现象。

在Scaling instruction-finetuned language models的实验中，PaLM 至少在 620 亿参数上才能在 4 个评估基准的各种任务上表现良好。

1.4.3 逐步推理

对于涉及多个推理步骤的复杂任务（如数学），可以使用**思维链（Chain-of-Thought, CoT）**提示策略（Chain of thought prompting elicits reasoning in large language models），让 LLM 通过**利用中间推理步骤的提示机制**来解决这类任务。

Chain of thought prompting elicits reasoning in large language models发现，CoT 在模型大于 600 亿的 PaLM 和 LaMBDA 变体中能够提升在算术推理基准任务的效果，而当模型大于 1000 亿时，相比标准提示的优势更明显。

How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources

1.5 LLM 关键点

如何让 LLM 能够**通用**且**有能力**?

1.5.1 扩展

更大的模型、数据规模和更多的训练计算，但计算预算是有限的，可以用扩展法更高效地分配计算资源，如 Chinchilla 在**相同计算预算下增加训练 token 数**，优于更大模型规模的 Gopher，同时需要数据清理。

1.5.2 训练

- 分布式的训练框架：包括 DeepSpeed（Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters）和 Megatron-LM（Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism和Efficient large-scale language model training on GPU clusters using megatron-lm）
- 优化技巧：有助于提升训练稳定性和模型性能，如**重新开始以克服训练损失激增**（Palm: Scaling language modeling with pathways）和**混合精度训练**（BLOOM: A 176b-parameter open-access multilingual language model）。

1.5.3 能力引导

当 LLM 执行某些特定任务时，可能不会显式地展示出其通用求解器的能力，**设计合适的任务指令或具体的 ICL 策略**可以**激发**这种能力，例如

- 通过**包含中间推理步骤的 CoT 提示**
- 使用**自然语言表达的任务描述**，对 LLM 进行**指令微调**

1.5.4 对齐微调

由于预训练语料库包括高质量和低质量的数据，LLM 可能生成有毒、偏见甚至有害的内容，要让 LLM 和人类价值观保持一致，如**有用性、诚实性和无害性**。RLHF 相关工作如Training language models to follow instructions with human feedback和Deep reinforcement learning from human preferences能够产生高质量、无害的回答（例如拒绝回答侮辱性问题）。

1.5.5 工具操作

LLM 本质是基于海量文本语料库进行文本生成训练的，对于不适合以文本形式表达的任务表现不佳（如数字计算），且其能力受限于预训练数据，无法获取最新信息。可以利用外部工具：

- Toolformer: Language models can teach themselves to use tools能利用计算器进行准确计算
- Webgpt: Browser-assisted question-answering with human feed- back能利用搜索引擎检索未知信息

2 RLHF & instructGPT

OpenAI 魔改大模型，参数减少 100 倍！13 亿参数 InstructGPT 碾压 GPT-3

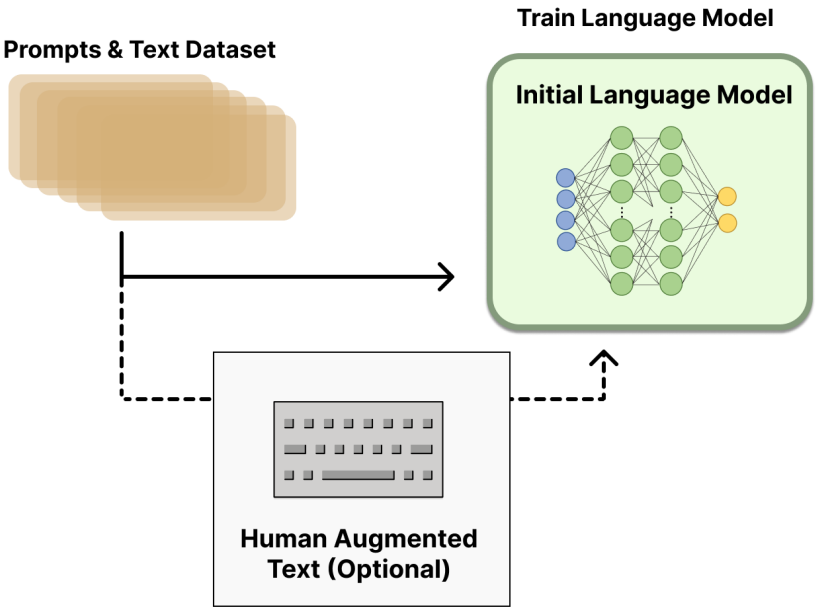
https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

Training language models to follow instructions with human feedback

https://huggingface.co/blog/zh/rlhf

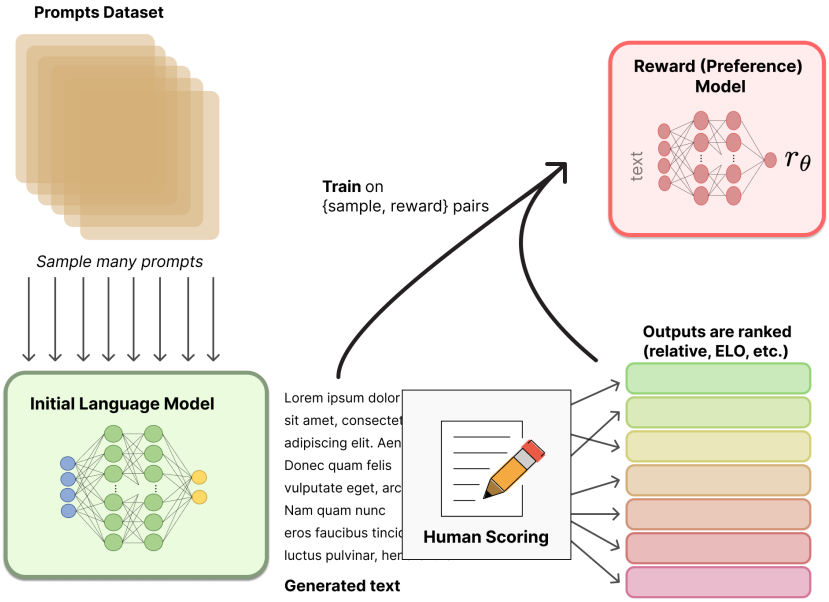
- 预训练一个语言模型 (LM)；
- 聚合问答数据并训练一个奖励模型 (Reward Model，RM)，也叫偏好模型；
- 用强化学习 (RL) 方式微调 LM。

2.1 sft



- openai：instructGPT 使用小版本的 GPT-3，并对“更可取”（preferable）的人工生成文本微调
- Anthropic：1000w-520 亿参数的 transformer，并按“有用、诚实和无害”的标准在上下文线索上蒸馏原始 LM
- DeepMind：2800 亿的模型 Gopher

2.2 rm

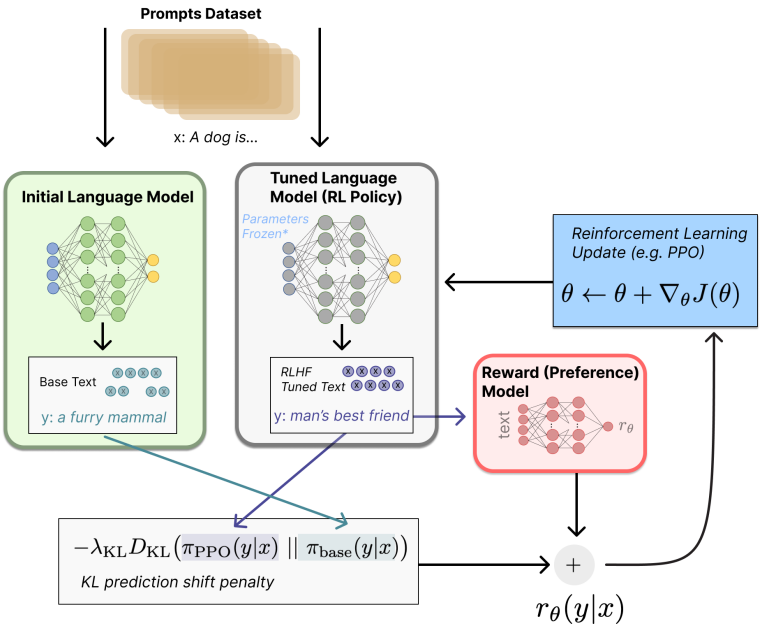


接收一系列文本并返回一个标量奖励，数值上对应人的偏好。我们可以用端到端的方式用 LM 建模，或者用模块化的系统建模 (比如对输出进行排名，再将排名转换为奖励) 。

- **模型选择**：RM 可以是另一个经过微调的 LM，也可以是根据偏好数据从头开始训练的 LM。Anthropic 提出了一种特殊的预训练方式，即用偏好模型预训练 (Preference Model Pretraining，PMP) 来替换一般预训练后的微调过程。因为前者被认为对样本数据的利用率更高。
- **训练文本**：RM 的提示 - 生成对文本是从预定义数据集中采样生成的，并用初始的 LM 给这些提示生成文本。Anthropic 的数据主要是通过 Amazon Mechanical Turk 上的聊天工具生成的，并在 Hub 上可用，而 OpenAI 使用了用户提交给 GPT API 的 prompt。
- **训练奖励数值**：人工对 LM 生成的回答进行排名。起初我们可能会认为应该直接对文本标注分数来训练 RM，但是由于标注者的价值观不同导致这些分数未经过校准并且充满噪音，通过排名可以比较多个模型的输出并构建更好的规范数据集，这些不同的排名结果将被归一化为用于训练的标量奖励值。

目前成功的 RLHF 系统使用了和生成模型具有 不同 大小的 LM，OpenAI 使用了 175B 的 LM 和 6B 的 RM，Anthropic 使用的 LM 和 RM 从 10B 到 52B 大小不等，DeepMind 使用了 70B 的 Chinchilla 模型分别作为 LM 和 RM

2.3 rl



直接微调整个 10B~100B+ 参数的成本过高，参考低秩自适应LoRA和 DeepMind 的Sparrow LM。目前多个组织找到的可行方案是使用策略梯度强化学习 (Policy Gradient RL) 算法、近端策略优化 (Proximal Policy Optimization，PPO) **微调初始 LM 的部分或全部参数。**

- 策略 (policy)：一个接受提示并返回一系列文本 (或文本的概率分布) 的 LM
- 行动空间（action space）：LM 的词表对应的所有词元 (一般在 50k 数量级)
- 观察空间 (observation space)：是可能的输入词元序列，也比较大 (词汇量 ^ 输入标记的数量)
- 奖励函数：偏好模型和策略转变约束 (Policy shift constraint) 的结合。

ppo 确定的奖励函数如下：

- 提示 x 输入初始 LM 和当前微调的 LM，分别得到输出文本 y_1 和 y_2
- 将来自当前策略的文本传给 RM 得到标量奖励 r_θ
- 将两个模型的生成文本进行比较计算差异的惩罚项，一般是输出词分布间的 KL 散度的缩放，即 $r = r_\theta - \lambda r_{KL}$，

惩罚项的好处：+ 用于惩罚策略在每个训练 batch 中生成大幅偏离初始模型，以确保模型输出合理连贯的文本。+ 如果没有这一项，可能导致模型在优化中生成乱码文本，以愚弄奖励模型提供高奖励值。

根据 PPO，按当前 batch 的奖励进行优化。PPO 是置信域优化（TRO，Trust Region Optimization）算法，用梯度约束确保更新步骤不会破坏学习过程的稳定性。

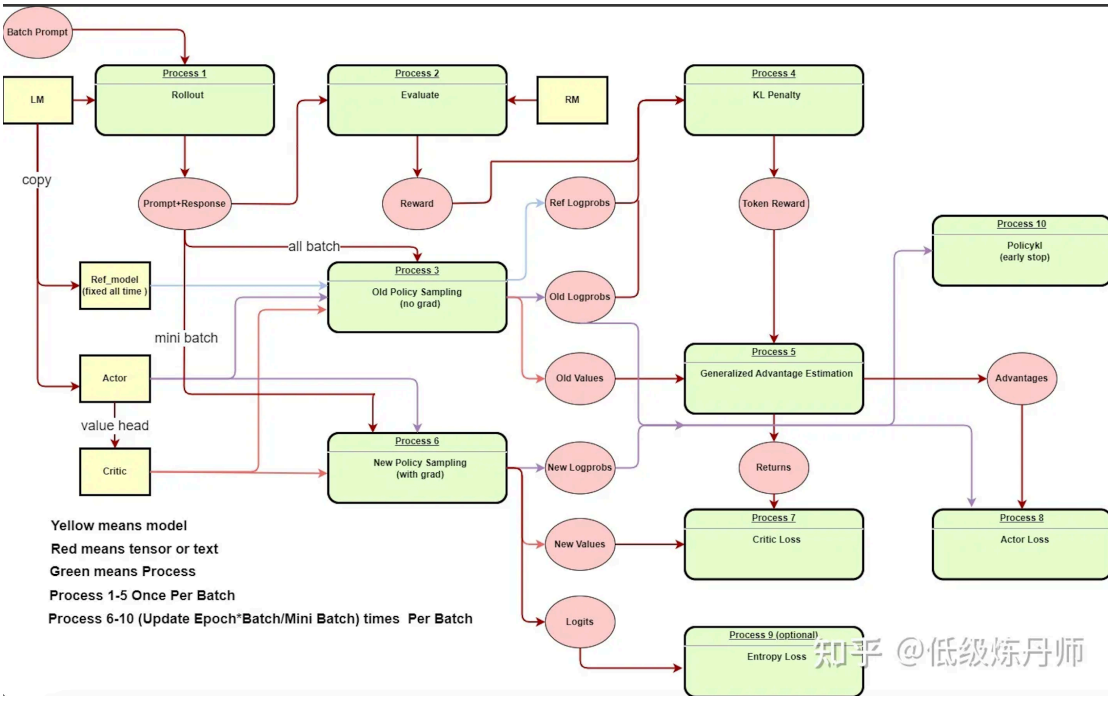
DeepMind 对 Gopher 用了类似的奖励设置，但用的是 A2C 来优化梯度。

2.3.1 rl 流程概述

https://zhuanlan.zhihu.com/p/635757674

Fine-Tuning Language Models from Human Preferences

Secrets of RLHF in Large Language Models Part I: PPO



- Rollout and Evaluation：从 prompt 库里抽样，使用语言模型生成 response，然后使用奖励模型（Reward Model, RM）给出奖励得分。这个得分反映了生成的 response 的质量，比如它是否符合人类的偏好，是否符合任务的要求等。
- Make experience：收集了一系列的“经验”，即模型的行为和对应的奖励。这些经验包括了模型生成的 response 以及对应的奖励得分。这些经验将被用于下一步的优化过程。
- Optimization：使用收集到的经验来更新模型的参数。具体来说，我们使用 PPO 算法来调整模型的参数，使得模型生成的 response 的奖励得分能够增加。PPO 算法的一个关键特性是它尝试保持模型的行为不会发生太大的改变，这有助于保证模型的稳定性。

官方代码 example

```
from tqdm import tqdm

for epoch, batch in tqdm(enumerate(ppo_trainer.dataloader)):
    query_tensors = batch["input_ids"]

    #### Get response from SFTModel
    response_tensors = ppo_trainer.generate(query_tensors, **generation_kwargs)
    batch["response"] = [tokenizer.decode(r.squeeze()) for r in response_tensors]

    #### Compute reward score
    texts = [q + r for q, r in zip(batch["query"], batch["response"])]
    pipe_outputs = reward_model(texts)
    rewards = [torch.tensor(output[1]["score"]) for output in pipe_outputs]

    #### Run PPO step
```

```
    stats = ppo_trainer.step(query_tensors, response_tensors, rewards)
    ppo_trainer.log_stats(stats, batch, rewards)

#### Save model
ppo_trainer.save_model("my_ppo_model")
```

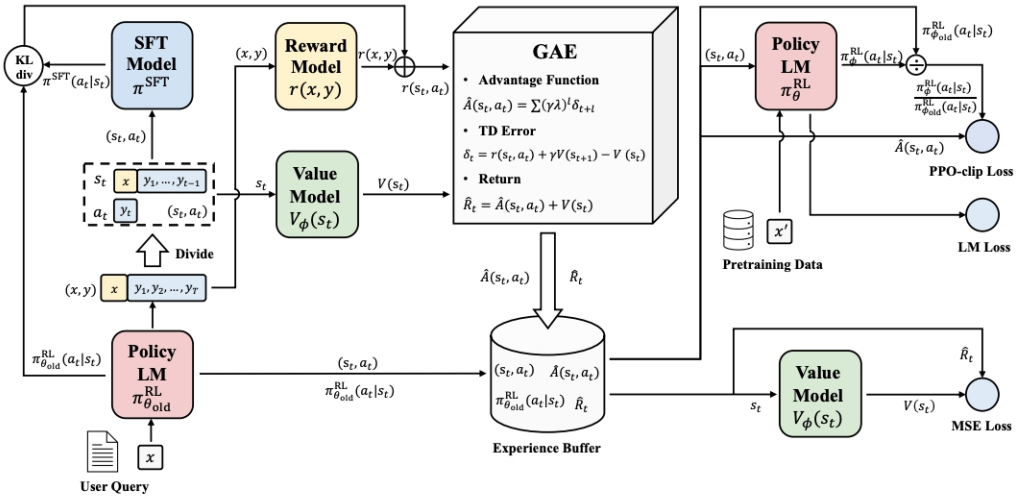


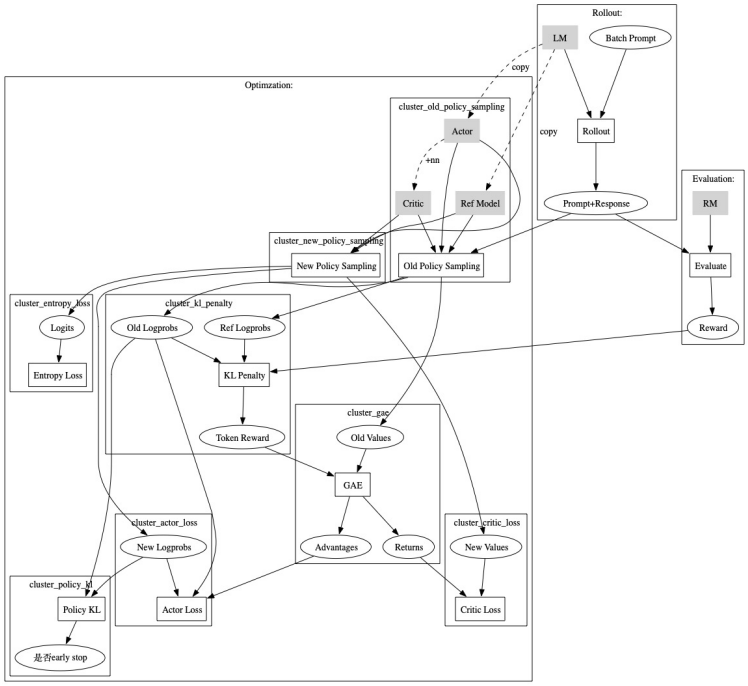
Figure 1: PPO workflow, depicting the sequential steps in the algorithm's execution. The process begins with sampling from the environment, followed by the application of GAE for improved advantage approximation. The diagram then illustrates the computation of various loss functions employed in PPO, signifying the iterative nature of the learning process and the policy updates derived from these losses.

- Rollout：根据策略（LM）生成轨迹（文本）。
  - 输入：Batch Prompt、LM
  - 输出：Prompt+Response
- Evaluate：对生成的轨迹进行评估（RM）。
  - 输入：Prompt+Response、RM
  - 输出：Reward
- Old Policy Sampling：计算并存储旧策略的概率、价值等值，
  - 输入：Ref_model、Actor、Critic、Prompt+Response
  - 输出：Ref Logprobs、Old Logprobs、Old Values
- KL Penalty：计算**当前策略**和**原始 LM** 之间的 KL 散度，用作对策略改变过快的惩罚项。
  - 输入：Ref Logprobs、Old Logprobs、Reward
  - 输出：Token Reward
- Generalized Advantage Estimation (GAE)：G。基于 old value(shape 是 (`batch_size`, `response_length`)) 和 reward 估计优势函数 A，它结合了所有可能的 n-step 进行 advantage 估计
  - 输入：Token Reward、Old Values
  - 输出：Advantages、Returns
- New Policy Sampling：
  - 输入 ref_model、actor、critic，从新的策略中采样概率等信息，
  - 输出 new logprobs、new values 和 logits，供 actor loss、critic loss 以及 entropy loss 用。
- Critic Loss：Critic 的目标是估计状态的价值函数，Critic loss 就是价值函数预测值和实际回报之间的差距。
  - 输入：New Values、Returns
  - 输出：critic 梯度更新
- Actor Loss：Actor 的目标是优化策略，Actor loss 就是基于优势函数的策略梯度。
  - 输入：Old Logprobs，New Logprobs、Advantages
  - 输出：actor 梯度更新

- Entropy Loss：为了增加探索性，通常会添加一个基于策略熵的正则项，它鼓励策略保持多样性。
  - 输入：Logits
  - 输出：entropy loss
- Policykl：这是对策略迭代过程的一个度量，它度量**新策略**和**旧策略**之间的差距。
  - 输入：Old Logprobs、New Logprobs
  - 输出：是否 early stop

在 PPO 中，策略优化的过程涉及到两个策略：一个是”旧的”策略，这是我们在开始每次优化迭代时使用的策略，另一个是”新的”策略，这是我们在优化过程中**不断更新**的策略。

自己整理重画的



2.3.2 几个重要的 loss

2.3.2.1 actor & actor loss Actor 是**策略**，它决定文本会被怎么样生成，是从**策略网络**拷贝来的模拟整个智能体在环境中行动的网络。

优势函数表示在给定的状态下采取某个行动比遵循当前策略的期望回报要好多少。

Actor Loss 如下，用重要性采样比较在**旧策略**和**新策略**下行动的概率（Old Logprobs，New Logprobs），然后将这个比值（也就是 Importance Sampling 的权重）与**优势函数 Advantages** 相乘，得到了对 Actor Loss 的一个估计。

$$L = \pi_{new}/\pi_{old} * A$$

```
# 计算新旧策略下概率的比值
ratio = torch.exp(logprobs - old_logprobs)

# 计算未截断的策略梯度损失
pg_losses = -advantages * ratio

# 计算截断的策略梯度损失
pg_losses2 = -advantages * torch.clamp(ratio, 1.0 - self.config.cliprange,
     1.0 + self.config.cliprange)
```

```
# 选择两者中较大的作为最终的策略梯度损失
pg_loss = masked_mean(torch.max(pg_losses, pg_losses2), mask)

# 计算因为截断导致策略梯度损失改变的比例
pg_clipfrac = masked_mean(torch.gt(pg_losses2, pg_losses).double(), mask)
```

2.3.2.2 critic & critic loss

critic 是专门用来预测 actor 轨迹**每一步价值**的网络，actor 上加几个线性层能够给每个 token 预测一个值。任务是估计状态的价值函数，也就是预测从当前状态开始，通过遵循某个策略，期望能得到的总回报。

Critic Loss 是最小化它的预测价值与实际回报之间的差距，常用 mse

通过最小化 Critic Loss，Critic 的预测能力会逐渐提升。因为 Critic 的预测结果会被用来**估计每个行动的优势（Advantage）**，这个优势值又会被用来计算策略的更新（Actor Loss）。

```
# 将价值函数的预测值裁剪到一个范围内
vpredclipped = clip_by_value(
            vpreds, values - self.config.cliprange_value, values + self.config.cliprange_value
        )

# 计算裁剪前和裁剪后的价值函数损失
vf_losses1 = (vpreds - returns) ** 2
vf_losses2 = (vpredclipped - returns) ** 2

# 最终的价值函数损失是裁剪前和裁剪后损失的最大值的平均值的一半
vf_loss = 0.5 * masked_mean(torch.max(vf_losses1, vf_losses2), mask)

# 计算裁剪操作实际发生的频率
vf_clipfrac = masked_mean(torch.gt(vf_losses2, vf_losses1).double(), mask)
```

2.3.2.3 KL Penalty

用于保证经过强化学习后的模型（新策略 actor）不会过于偏离原始预训练模型（ref model）。

```
# 初始化两个列表来分别存储奖励和非得分奖励
rewards, non_score_rewards = [], []

# 使用 zip 函数并行遍历输入的得分、对数概率、参考模型的对数概率以及 mask
for score, logprob, ref_logprob, mask in zip(scores, logprobs,
        ref_logprobs, masks):
    # 计算 KL 散度，即模型的对数概率与参考模型的对数概率之间的差值
    kl = logprob - ref_logprob

    # 计算非得分奖励，即 KL 散度乘以 KL 控制器值的负值
    non_score_reward = -self.kl_ctl.value * kl
    non_score_rewards.append(non_score_reward)

    # 复制非得分奖励为新的奖励
    reward = non_score_reward.clone()

    # 找到 mask 中最后一个非零元素的索引，这表示输入序列的实际长度
    last_non_masked_index = mask.nonzero()[-1]

    # 对于最后一个非 mask 部分的 token，其奖励是偏好模型的得分加上 KL 散度
    reward[last_non_masked_index] += score

    # 将计算的奖励添加到奖励列表中
```

```
    rewards.append(reward)

# 返回包含所有奖励的张量以及包含所有非得分奖励的张量
return torch.stack(rewards), torch.stack(non_score_rewards)
```

2.3.2.4 GAE GAE 是一种多步优势估计方法。它通过引入一个权衡参数 λ，在**单步 TD 误差**和**多步 TD 误差**之间进行权衡，从而**减小估计的方差**，提高学习的稳定性。其中 σ_{t+l} 是时间步 $t+l$ 的 TD 误差。

$$A_t = \sum_{l=0}^{k-1} (\lambda\eta)^l \sigma_{t+l}$$

$$\sigma_{t+l} = r_{t+l+1} + \eta V(s_{t+l+1}) - V(s_{t+l})$$

```
# 从后往前遍历整个生成的序列
for t in reversed(range(gen_len)):
    # 计算下一个状态的价值，如果当前状态已经是最后一个状态，则下一个状态的价值为 0
    nextvalues = values[:, t + 1] if t < gen_len - 1 else 0.0

    # 计算 delta，它是奖励加上衰减后的下一个状态的价值，然后减去当前状态的价值
    delta = rewards[:, t] + self.config.gamma * nextvalues - values[:, t]

    # 使用 delta 更新 lastgaelam，这是 GAE 公式的一部分
    lastgaelam = delta + self.config.gamma * self.config.lam * lastgaelam

    # 将计算的优势值添加到优势值列表中
    advantages_reversed.append(lastgaelam)

# 将优势值列表反向并转换为张量
advantages = torch.stack(advantages_reversed[::-1]).transpose(0, 1)

# 计算回报值，它是优势值加上状态值
returns = advantages + values
```

2.3.2.5 entropy loss 一个策略的熵越大，意味着这个策略选择各个动作的概率更加“平均”。在 actor 的 loss 里加熵，使得策略的熵尽可能大，从而有更多机会探索可能带来更好奖励的文本轨迹。

```
entropy = -torch.sum(logits* torch.log(logits + 1e-9), dim=-1).mean()
```

新实现：

```
    pd = torch.nn.functional.softmax(logits, dim=-1)
    entropy = torch.logsumexp(logits, axis=-1) - torch.sum(pd * logits, axis=-1)
```

2.3.2.6 Policy kl 在 PPO 中，KL 散度被用作一种约束，以确保在优化过程中新策略不会偏离旧策略太远。这是为了防止过度优化，因为过度优化可能会导致策略性能的大幅下降。

我们希望在优化目标函数的同时，满足以下的 KL 散度约束：

$$KL[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)] \leq \delta$$

在代码中，每个 mini batch 都会进行 early stop 的判定，如果计算出的 KL 散度大于 δ，那么就会停止这一轮的优化，以保证新策略不会偏离旧策略太远。

```
# 计算旧策略和新策略之间的 KL 散度
policykl = masked_mean(old_logprobs - logprobs, mask)
# old_logprobs 是旧策略下行为的概率的对数，logprobs 是新策略下的对数概率
# masked_mean 函数计算差异（old_logprobs - logprobs）的平均值，
# 但只考虑 mask 中对应元素为 True 的元素

# 检查计算出的 KL 散度（policykl）是否大于目标 KL 散度（self.config.target_kl）的 1.5 倍
if policykl > 1.5 * self.config.target_kl:
    self.optimizer.zero_grad()
    # 如果实际的 KL 散度超过了目标的 1.5 倍，那么策略改变过多，这步的梯度也不更新了。
    early_stop = True
    # 并设置 early_stop 标志为 True，表示应提前停止优化，以防止策略从旧策略进一步偏离
```

2.3.3 两个采样

2.3.3.1 Old Policy Sampling（无 bp） 是 **make experience** 的过程，计算并**存储**旧策略的概率、价值等值，来为后面更新的过程服务。

- Old Logprobs：从“旧的”策略 [即在这个 batch 数据中初始的 LM（initial actor）] 中计算每个 token 在旧的策略下的概率 Old Logprobs。
- Old Values：旧策略中每个**时间步**（每个 token 的预测结果）的价值，这个值由 critic 网络进行预测，critic 网络就是需要这个值的原因是 advantage 的计算依赖于 Old Values。
- Ref Logprobs：最最原始的 LM 对于每个时间步的概率预测，一般就是**固定不变的 gpt3**，计算这个值的目的是限制 actor 的更新，防止其偏离原始 gpt3 太远，他的实现在下一个步骤中。

```
all_logprobs, _, values, masks = self.batched_forward_pass(self.model, queries,
    responses, model_inputs)
ref_logprobs, _, _, _ = self.batched_forward_pass(self.ref_model, queries,
    responses, model_inputs)
```

2.3.3.2 New Policy Sampling（有 bp） 在**新的策略**（更新后的 actor）下对轨迹（文本）计算概率的过程，计算 Actor Loss，即策略梯度的损失。

Old Logprobs 是一次性一个 batch 的数据计算的，这是因为在一个 batch 中旧策略都是不变的；而 New Logprobs 是一个 mini batch 计算一次，这是因为新策略每个 mini batch 变一次。

2.3.4 开源 rlhf 库

2.3.4.1 openai 的 lm-human-preferences(gpt2 的 finetune) https://github.com/openai/lm-human-preferences

2.3.4.2 huggingface 的 TRL https://github.com/huggingface/trl

2.3.4.3 CarperAI 的 trlx https://github.com/CarperAI/trlx

2.3.4.4 allenai 的 RL4LMs https://github.com/allenai/RL4LMs

3 llama

LLaMA: Open and Efficient Foundation Language Models

参考代码：https://github.com/huggingface/transformers/blob/main/src/transformers/models/llama/modeling_llama.py

之前的工作考虑的是在训练预算有限的前提下，如何提升模型性能（2022 年 deepmind 的Training Compute-Optimal Large Language Models的 Chinchilla）,llama 考虑在预测时的预算。例如 chinchilla 是一个 10b 的模型在 200b 的 token 上训练，但其

实一个 7b 的模型当用了 1T 的 token 后，性能仍在提升。LLama-13b 比 gpt3 在大多数 benchmark 上好，但 size 只有 1/10，在一个 GPU 上就能跑。

llama 只用公开数据训练，而 Chinchilla、PaLM、GPT-3 都有自己的未公开数据集。其他的 OPT、GPT-NeoX、BLOOM、GLM 虽然也只用公开数据集，但打不过 PaLM-62B 或者 Chinchilla

3.1 预训练数据

- English CommonCrawl(67%)：使用 CCNet pipeline，去重、用 fasttext 把非英文的页面删了，用 n-gram 把低质内容删了。此外，还训了一个线性模型，对页面进行分类：作为维基百科的引用 vs 随机采样的页面，最后把不属于引用这个类别的页面删了
- C4(15%)：与 CCNet 类似，主要区别在质量过滤是基于启发式的规则，如标点符号的存在，或者词数和句子数
- github(4.5%)：使用 Google BigQuery 里的公开 github 数据集，只用 Apache、BSD 和 MIT 证书的。低质判断是启发式规则，如字母数字占比、行的长度等，用正则删掉 head 等样式，最终以文件粒度进行去重。
- wikipedia(4.5%)：2022 年 6-8 月的数据，包括 20 种语言
- Gutenberg and Books3(4.5%)：两个书籍数据集，对有 90% 以上内容重复的书籍做去重。
- Arxiv(2.5%)：拿原始的 tex 文件，删掉 first section 之前的东西，还有一些注释、宏
- Stack Exchange(2%)：高质量的问答网站，按答案的分数排序

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

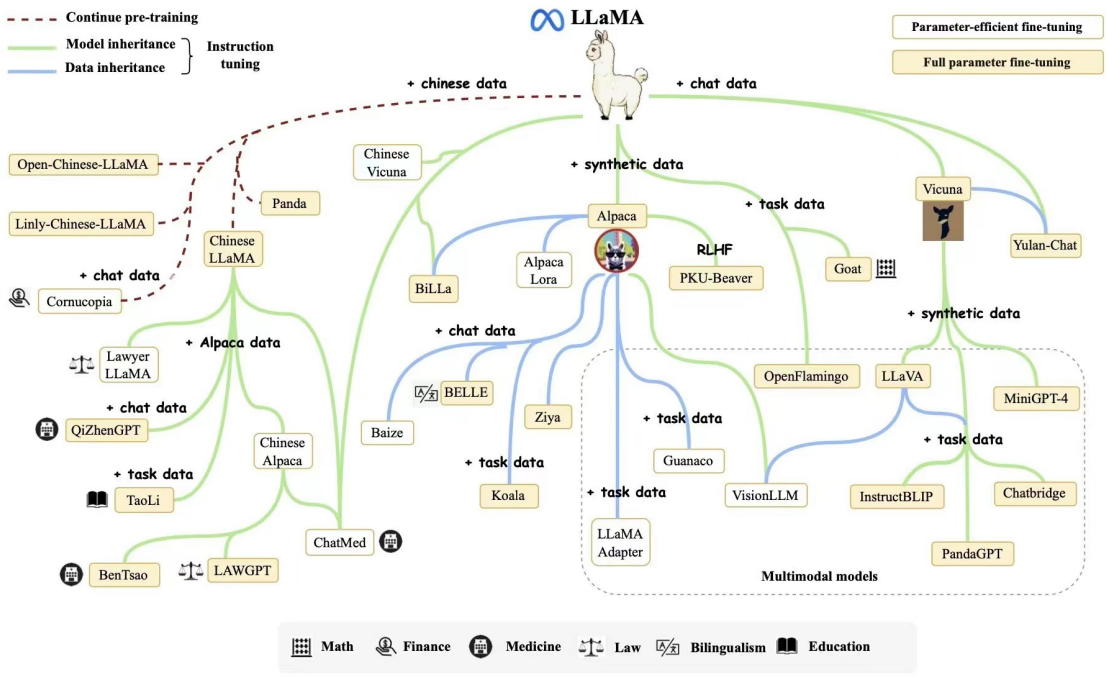
tokenizer：BPE，使用 sentencepiece 的实现。将所有 numbers 切成单个数字，回退到字节去处理未知的 utf8 字符（fallback to bytes to decompose unknown UTF-8 characters）

总共有 1.4T 的 token，对大部分训练数据，每个 token 在训练时只用了一次，除了维基和 book 大概用了两次。

附：gpt4 说：当我们说”一个 token 只训练一次”，我们其实是在说在一个 epoch（一个完整遍历训练集的过程）中，我们只遍历一次完整的数据集。如果一个特定的 token 在数据集中出现多次，那么在一个 epoch 中，这个 token 就会被用来训练模型多次。

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

Table 2: **Model sizes, architectures, and optimization hyper-parameters.**



3.2 网络结构

- pre-normalization(gpt3)：提升训练**稳定性**，对每个子层的输入做 norm，而非输出。此外，使用的是 RMSNorm 函数 (Root mean square layer normalization)
- SwiGLU 激活函数 (PaLM)：Glu variants improve trans- former，把 PaLM 里的 $4d$ 改了 $2/34d$
- Rotary embeddings(GPTNeo)：删掉原来的绝对位置编码，加上 rotary positional embedding(RoPE)，网络的每一层都加，参考Roformer: En- hanced transformer with rotary position embedding

优化器：AdamW，cosine 学习率 schedule，最终学习率是最大学习率的 10%。0.1 的 weight decay 和 1.0 的 gradient cliping，使用 2000steps 的 warmup

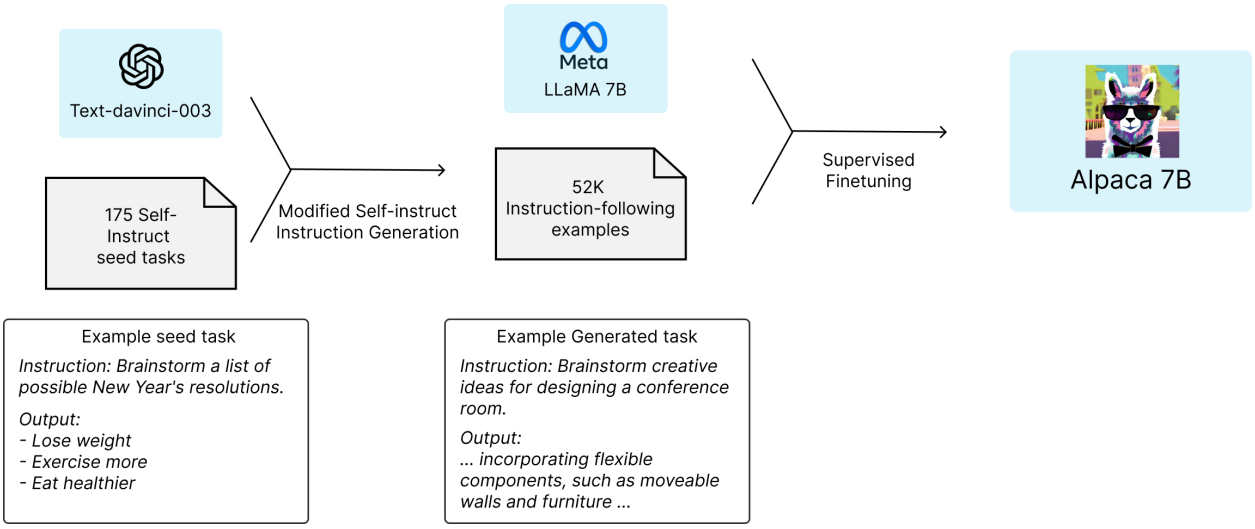
3.3 训练加速

- 对 causal multi-head attention 加速：实现在http://github.com/facebookresearch/xformers中，降低内存使用和运行时间，参考self-attention does not need $o(n^2)$ memory，以及Flashattention: Fast and memory-efficient exact attention with io-awareness。思想是
  - 不存储 attention weights
  - 不计算被 mask 的 key/query 得分
- 减少 xxx：

3.4 衍生：Alpaca

Alpaca: A Strong, Replicable Instruction-Following Model

在 LLaMA 模型的基础上的一个著名的项目是 Stanford 的羊驼（Alpaca）模型，有 70 亿（7b）参数，**没有使用 RLHF**，而是使用**监督学习**的方法。其数据集是通过查询基于 GPT-3 的 text-davinci-003 模型的结果，得到的 52k 的指令-输出对（instruction-output pairs）。因此，Alpaca 本质上使用的是一种弱监督（weakly supervised）或以知识蒸馏（knowledge-distillation-flavored）为主的微调。可以理解为是『用 LLM 来训练 LLM』，或者称之为『用 AI 来训练 AI』。



4 llama2

Llama 2: Open Foundation and Fine-Tuned Chat Models

https://zhuanlan.zhihu.com/p/636784644

5 Anthropic 的一些工作

Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Studying Large Language Model Generalization with Influence Functions

Measuring Faithfulness in Chain-of-Thought Reasoning

6 ChatGLM

ACL22 GLM: General Language Model Pretraining with Autoregressive Blank Infilling

iclr23 GLM-130B: An Open Bilingual Pre-trained Model

7 PALM-E

【IEEE Fellow 何晓东 & 邓力】多模态智能论文综述：表示学习，信息融合与应用，259 篇文献带你了解 AI 热点技

Multimodal Intelligence: Representation Learning, Information Fusion, and Applications

BERT 在多模态领域中的应用

CV 领域：VisualBert, Unicoder-VL, VL-Bert, ViLBERT, LXMERT。

CLIP

PaLM-E: An Embodied Multimodal Language Model

8 pathways

8.1 Google 的大规模稀疏模型设计

DESIGNING EFFECTIVE SPARSE EXPERT MODELS

代码：https://github.com/tensorflow/mesh/blob/master/mesh_tensorflow/transformer/moe.py

9 megatron-lm

https://zhuanlan.zhihu.com/p/646406772

10 deepspeed

https://zhuanlan.zhihu.com/p/343570325

11 ray-llm

https://github.com/ray-project/ray/releases/tag/ray-2.4.0

12 medusa-llm

decoder 的并行化：https://zhuanlan.zhihu.com/p/368592551

https://sites.google.com/view/medusa-llm

用了 tree-attention

13 大模型的一些现象

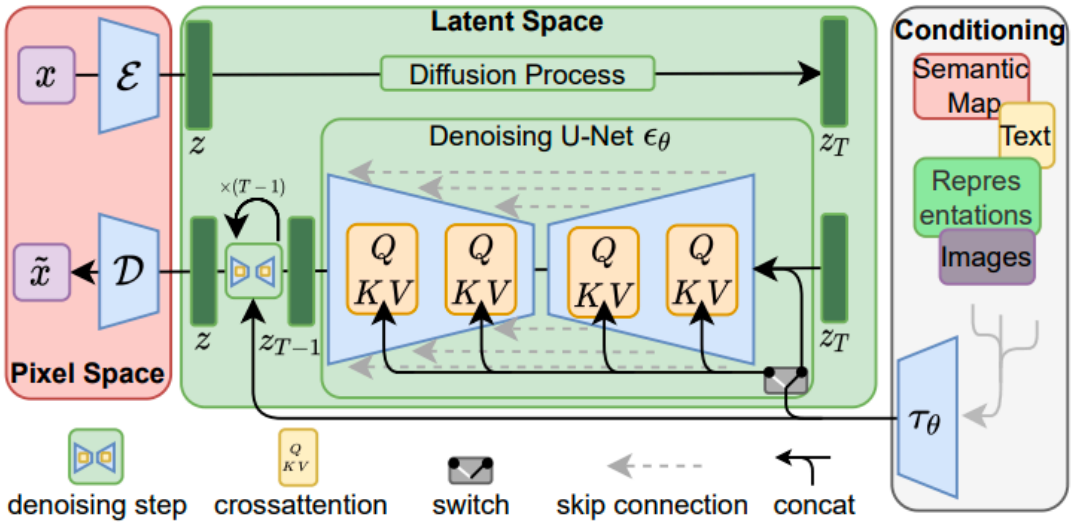
13.1 重复生成

https://www.zhihu.com/question/616130636

https://mp.weixin.qq.com/s/cSwWapqFhxu9zafzPUeVEw

14 stable diffusion

High-Resolution Image Synthesis with Latent Diffusion Models



输入图像，经过编码器得到 z，z 通过前向扩散不断加噪声得到 z_T（正向扩散）

输入条件，经过条件编码器 (原文是 BERT，到了 DALL-E2 就改成 CLIP 了) 得到 τ_θ

z_T 在 τ_θ 的指导下不断去噪（反向扩散），得到新的 z，再通过解码器得到最终生成的图像

15 LLM+ 推荐

15.1 综述

https://github.com/nancheng58/Awesome-LLM4RS-Papers

15.2 P5

Recommendation as Language Processing (RLP):A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)

15.3 llm vs ID

推荐系统范式之争，LLM vs. ID?

16 其他

16.1 公开资源

16.1.1 模型

是否有 ckpt	模型	发布时间	大小	预训练数据规模	硬件	训练时间
有	T5	2019.10	11B	1 万亿 tokens	1024 TPU v3	-
无	GPT-3	2020.05	175B	3000 万亿 tokens	-	-
无	GShard	2020.06	600B	1 万亿 tokens	2048 TPU v3	4 天
无	PanGu-alpha)	2021.04	13B	1.1TB	2048 Ascend 910	-

是否有 ckpt	模型	发布时间	大小	预训练数据规模	硬件	训练时间
无	Codex)	2021.07	12B	1000 万亿 tokens	-	-
有	mT5	2020.10	13B	1 万亿 tokens	-	-

16.1.2 数据集

llm 中文数据集：https://juejin.cn/post/7238921093553438779

16.2 RETRO Transformer

参数量仅为 4%，性能媲美 GPT-3：开发者图解 DeepMind 的 RETRO

http://jalammar.github.io/illustrated-retrieval-transformer/

Improving language models by retrieving from trillions of tokens

16.3 WebGPT

WebGPT: Browser-assisted question-answering with human feedback

https://openai.com/blog/webgpt/

16.4 llm 应用合辑

- ChatGPT 聚合站：https://hokex.com
- 游戏生成站：https://latitude.io/
- 家庭作业辅助站：https://ontimeai.com/
- 文字转语音站：https://www.resemble.ai/
- 艺术作画站：https://starryai.com/
- logo 制作站：https://www.logoai.com/
- ai 写作站：https://www.getconch.ai/
- 音乐制作站：https://soundraw.io/
- 声音模拟站：https://fakeyou.com/
- 一句话生成一段视频：https://runwayml.com/
- 文字转语音：https://murf.ai/

16.5 nanogpt

简化版的 gpt，tiktoken：gpt2 中使用的开源分词工具，比 huggingface 的 tokenizer 快得多

```
import tiktoken
enc = tiktoken.get_encoding("gpt2")

# 字节对编码过程，我的输出是 [31373, 995]
encoding_res = enc.encode("hello world")
print(encoding_res)

# 字节对解码过程，解码结果：hello world
raw_text = enc.decode(encoding_res)
print(raw_text)
```

16.6 达摩院大模型技术交流

https://developer.aliyun.com/live/248332

ppt：链接 密码：5yyf