



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

面向智慧医疗的大语言模型微调技术研究

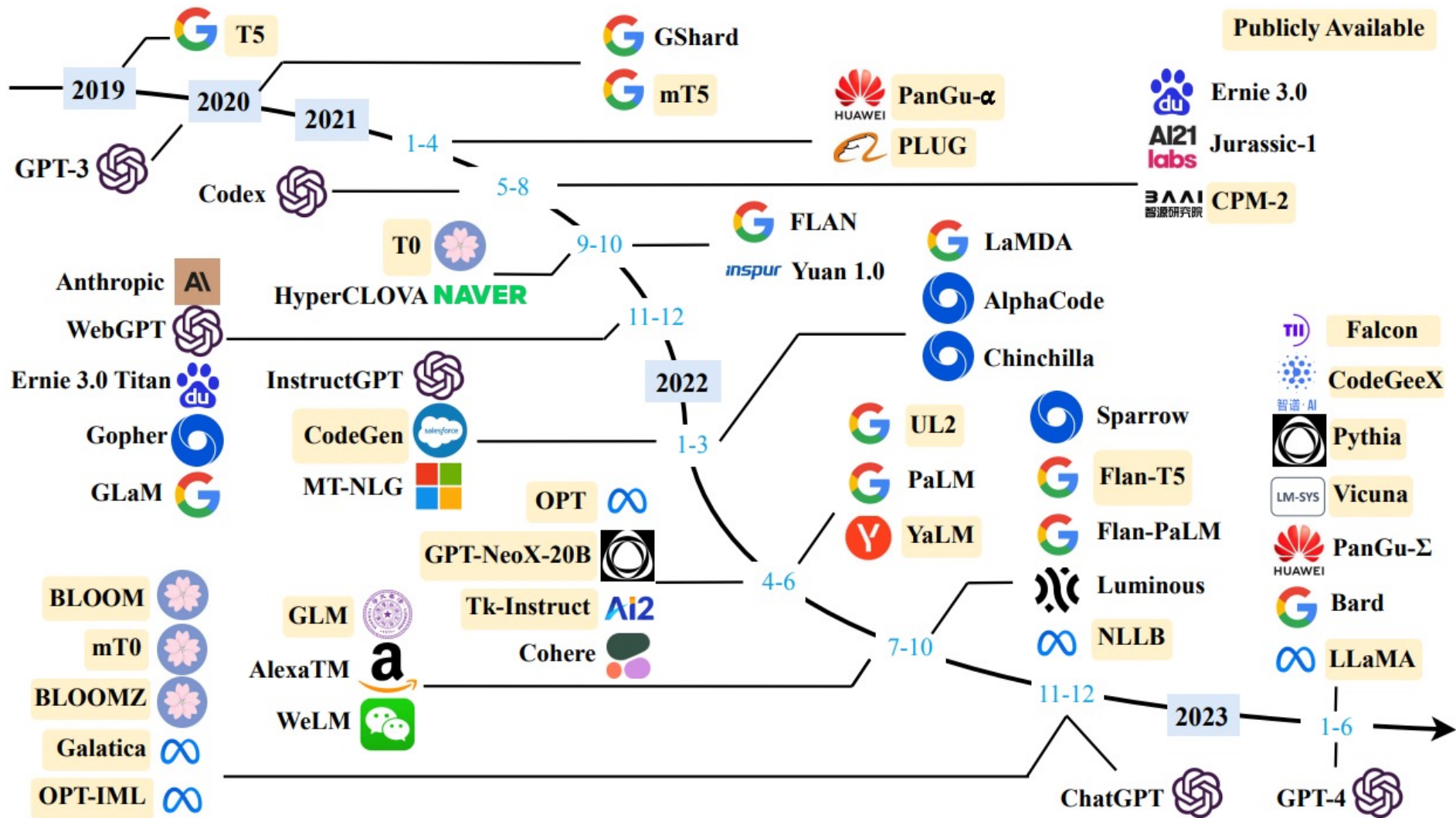
哈尔滨工业大学社会计算与信息检索研究中心

赵森栋

2023年9月24日

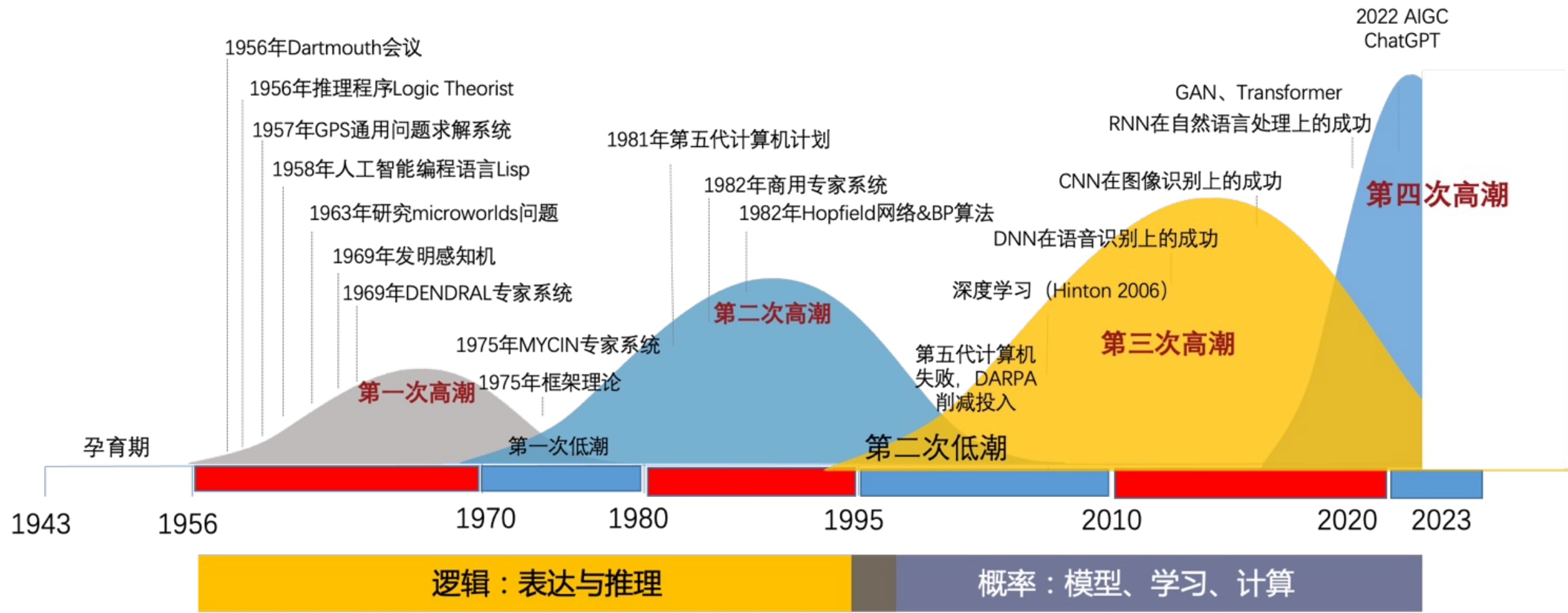
- 一** **大语言模型简介**
- 二** **大语言模型与智慧医疗**
- 三** **大语言模型微调技术**
- 四** **中文医学大模型**

百模大战



“A Survey of Large Language Models”

人工智能的四次高潮



研究大模型的核心是：人工智能——自然语言处理

- **自然语言**指的是人类语言，特指**文本符号**，而非语音信号
- **自然语言处理**（Natural Language Processing, NLP）
 - 用计算机来**理解**和**生成**自然语言的各种理论和方法
- 属于**认知智能**，是人类和动物的主要区别之一
 - 需要更强的**抽象**和**推理**能力
- 更具体地说：ChatGPT是（大）**语言模型**

语言模型-预测下一个词

我昨天晚上吃了

饭	6.5%
饺子	3.7%
面条	3.2%
烧烤	0.83%
烤串	0.17%
小笼包	0.13%
红肠	0.08%
...	
桌子	0.01%
...	

我昨天晚上在哈尔滨吃了

红肠	33.6%
锅包肉	26.3%
铁锅炖	17.8%
小鸡炖蘑菇	14.5%
马迭尔雪糕	9.8%
饺子	2.1%
面条	2.1%
...	
桌子	0.006%
...	

语言模型-根据上下文预测被掩码词

从一句话“我爱北京天安门”构造出4到带标准答案的填空题

填空题1: () 爱北京天安门

答案: 我

填空题2: 我 () 北京天安门

答案: 爱

填空题3: 我爱 () 天安门

答案: 北京

填空题4: 我爱北京 ()

答案: 天安门

我们的目标不是解决填空题，而是在解决填空题的过程中获得**通用语言智能**

语言模型数学定义

- 定义: 在单词序列 w_1, w_2, \dots, w_T 上的**概率分布** P 。
- 用不同假设分解这种**联合概率**会产生不同类型的语言模型。

经典语言模型

联合概率密度最大化求解问题

Bag of words model

$$p(w_1, w_2, \dots, w_T) = \prod_{i=1}^T p(w_i)$$

N-gram model

$$p(w_1, w_2, \dots, w_T) = \prod_{i=1}^T p(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-N})$$

Hidden Markov model

$$p(w_1, w_2, \dots, w_T) = \sum_{h_0, h_1, \dots, h_T \in H} p(h_0) \prod_{i=1}^T p(w_i | h_i) p(h_i | h_{i-1})$$

语言模型数学定义

- 定义: 在单词序列 w_1, w_2, \dots, w_T 上的概率分布 P 。
- 用不同假设分解这种联合概率会产生不同类型的语言模型。

神经网络语言模型

Neural
language
models:

Word embedding model

$$p(w_1, w_2, \dots, w_T)^{2c} = \prod_{i=1}^T \prod_{-c \leq j \leq c, j \neq 0} p(w_{i+j} | w_i)$$

Effectively an embedding layer followed by one-layer fully-connected neural network with softmax activation

RNN, LSTM, Transformer (w/. decoder)

Transformer (w/. encoder)

Generative language model

$$p(w_1, w_2, \dots, w_T) = \prod_{i=1}^T p(w_i | w_1, w_2, \dots, w_{i-1})$$

Masked language model

$$p(w_1, w_2, \dots, w_T) = \prod_{i=1}^T p(w_i | w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_T)$$

大语言模型的关键核心技术

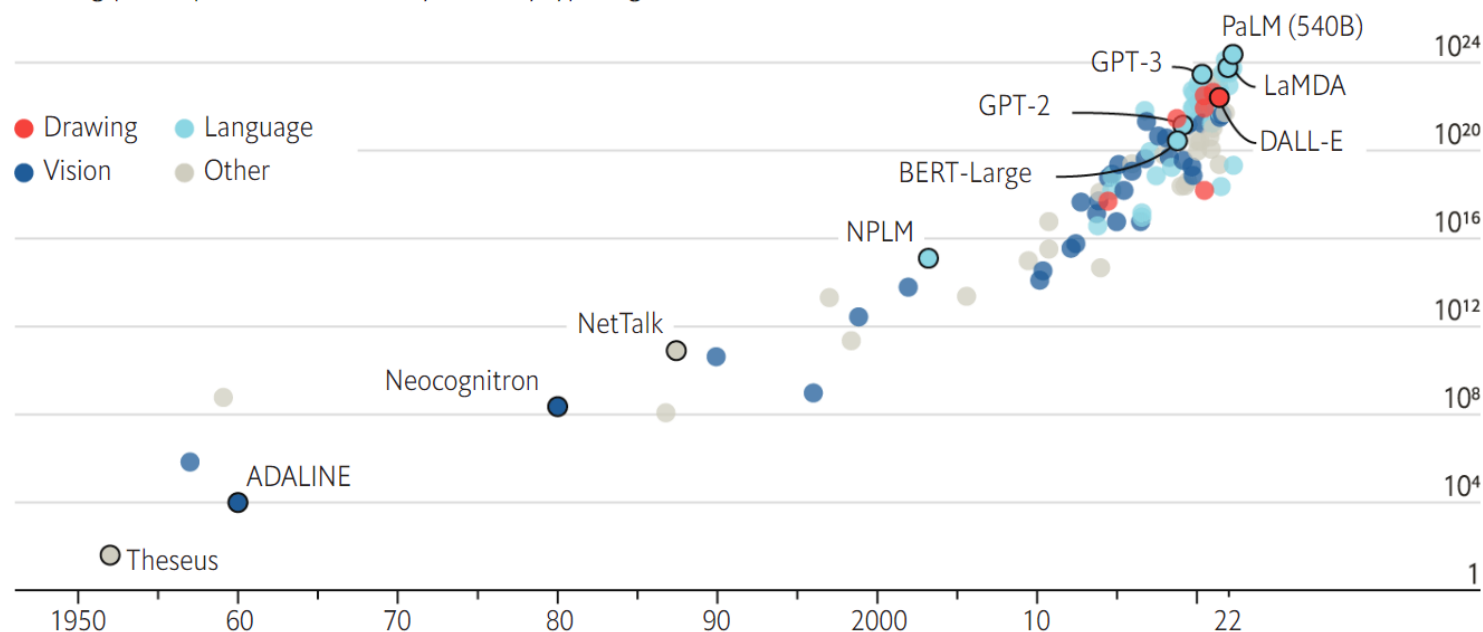
- **大规模预训练模型**
 - 模型规模越大，**能力越强**
- **指令微调 (Instruction Tuning)**
 - 将各种任务用Prompt形式进行统一
 - 在成千上万类型任务的标注数据上微调语言模型 (Instruction Tuning)
 - 模型能够处理未见任务 (Zero-shot)
- **基于人类反馈的强化学习 (Reinforcement Learning from Human Feedback)**
 - 结果更符合人类的预期 (多样性、安全性)
 - 利用真实用户的反馈 (AI正循环)

模型规模越来越大

- 模型规模与表现**正相关**，因此不停追求越来越大的规模
- 随着模型规模越来越大，“**涌现**”出了令人惊讶的“智能”
- 是否一直增强？

The blessings of scale

AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale



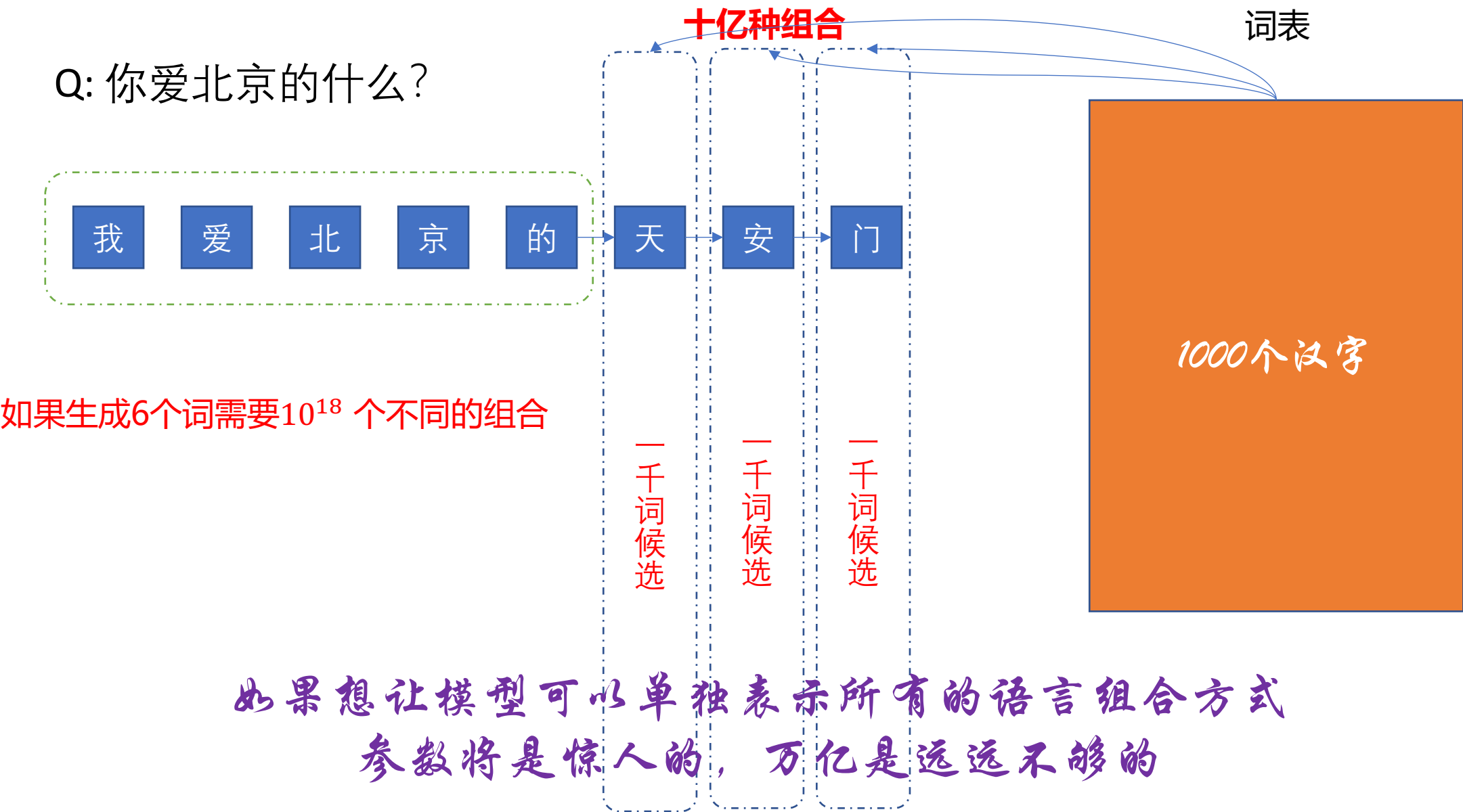
Sources: “Compute trends across three eras of machine learning”, by J. Sevilla et al., arXiv, 2022; Our World in Data

模型规模越来越大

模型	发布时间	参数量	预训练数据量	能力表现
GPT-1	2018年6月	1.17亿	约5GB	生成连贯的文本、回答简单问题、写作短文
GPT-2	2019年2月	15亿	40GB	创造更真实、更有逻辑性的文章、段落和句子
GPT-3	2020年5月	1750亿	45TB	能够执行未训练过的任务
GPT-4	2023年3月	1.8万亿	13万TB	模型幻觉现象大幅度减少

模型规模和训练数据的增加会带来能力的提高

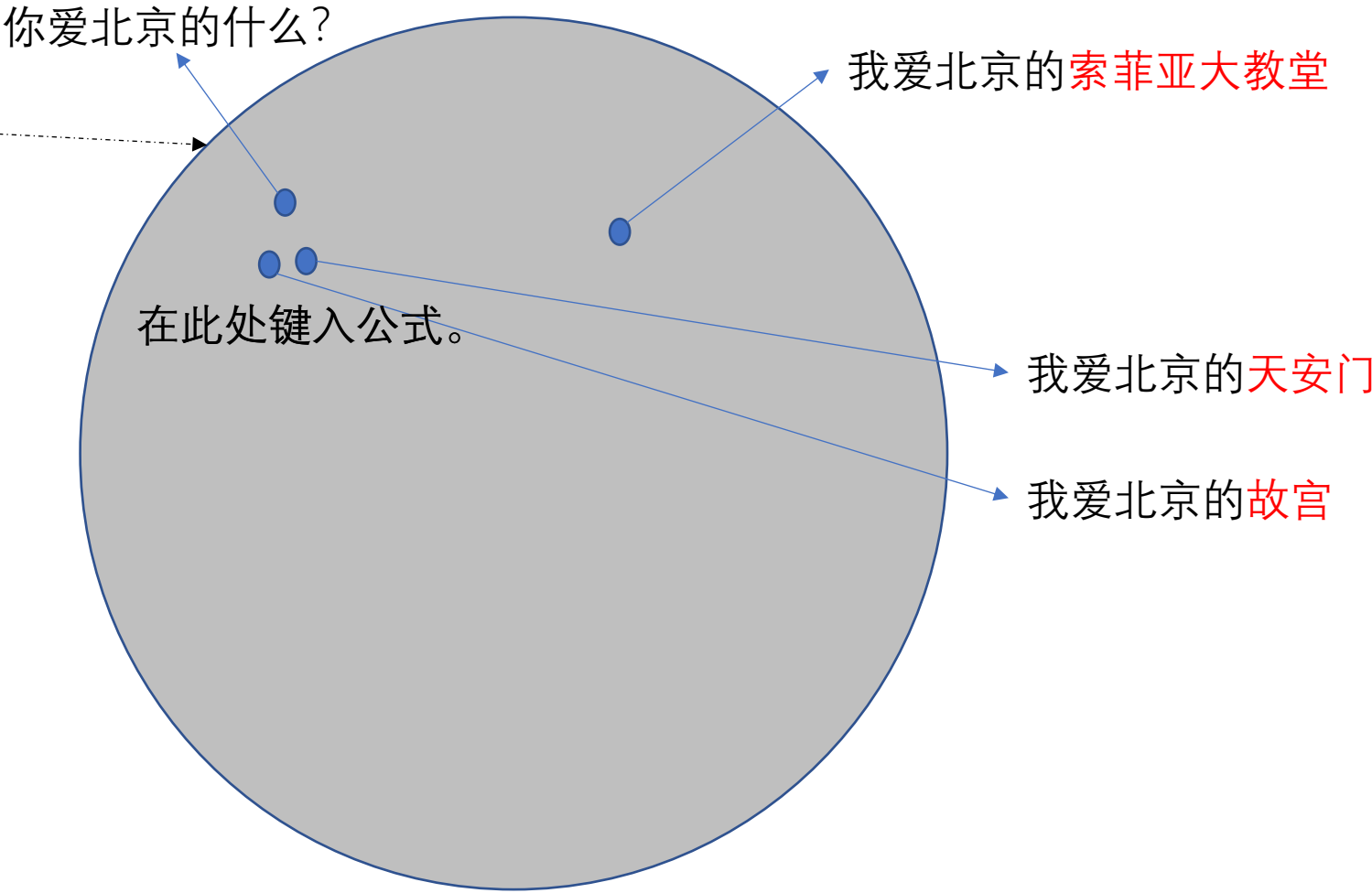
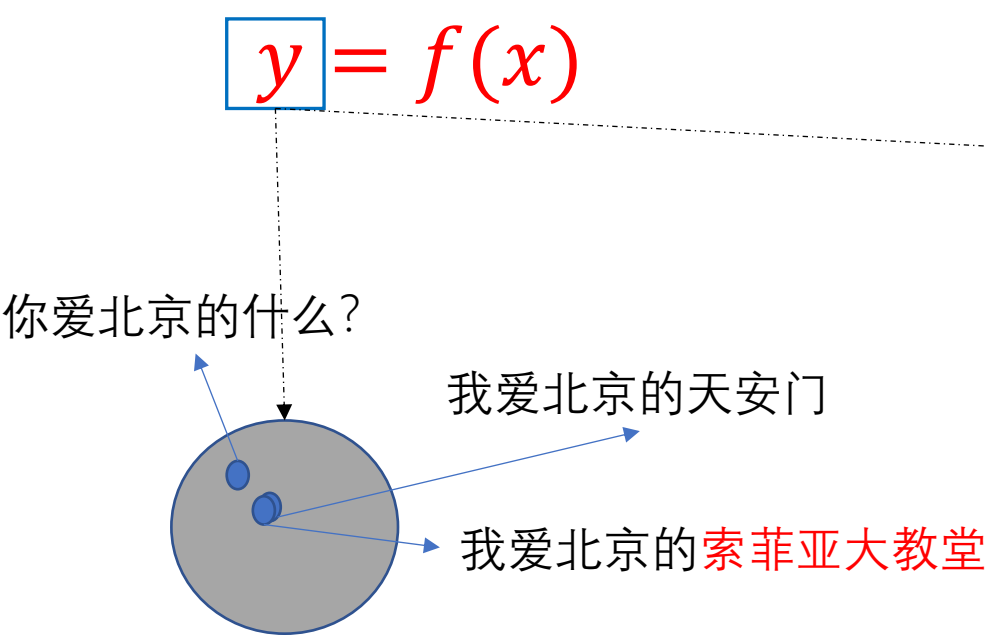
语言模型搜索空间爆炸问题



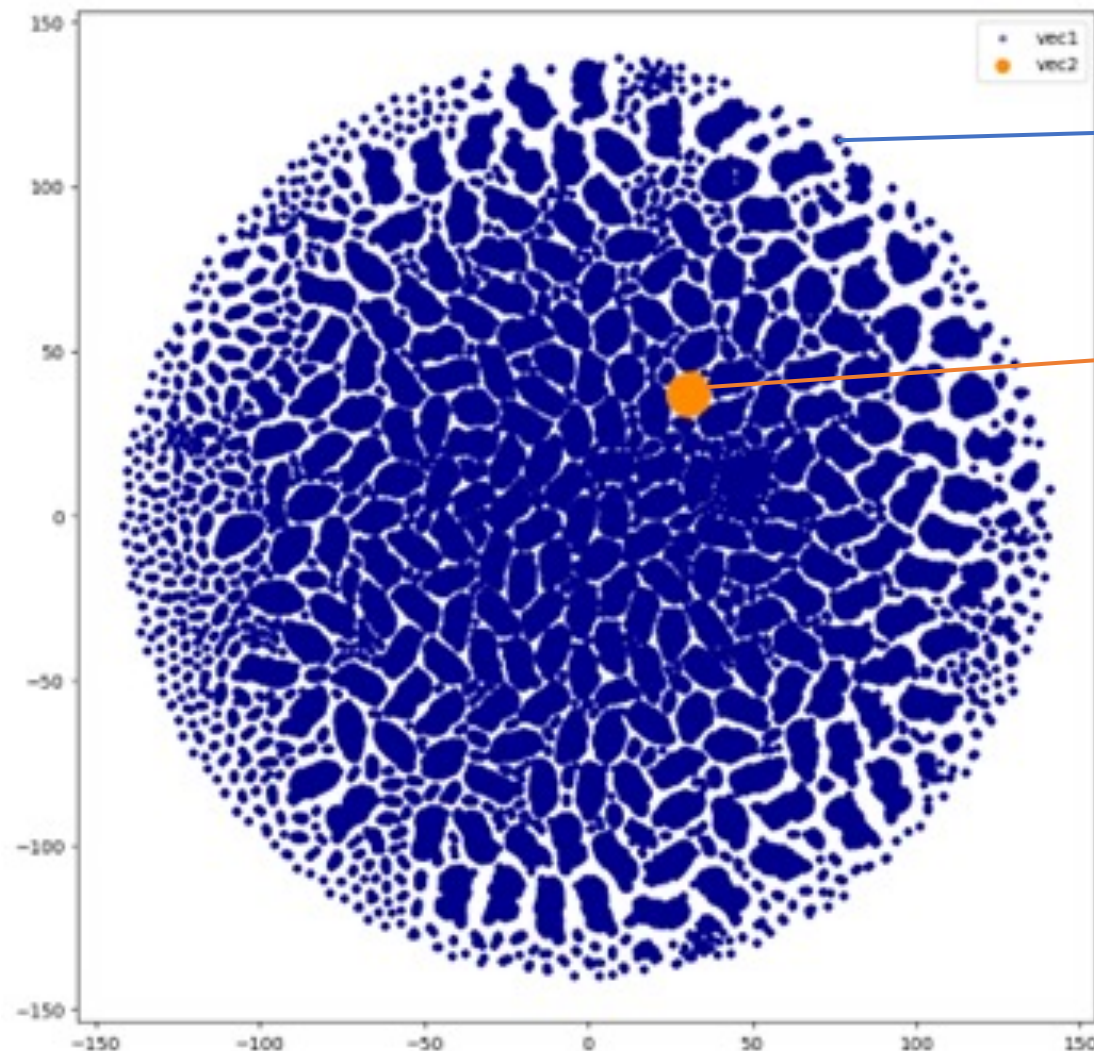
大语言模型如何解决这个问题？

基于神经网络的大模型用向量空间的距离度量替代概率计算

$$y = f(x)$$



不同规模模型的表示空间



200M参数模型

2M参数模型

相同的训练集和测试集

$D(f_{200M}(\dots \text{天安门}), f_{200M}(\dots \text{索菲亚教堂}))$

\gg

$D(f_{2M}(\dots \text{天安门}), f_{2M}(\dots \text{索菲亚教堂}))$

大语言模型应用示例

LA

你了解大语言模型的指令微调技术吗



是的，我熟悉大语言模型的指令微调技术。指令微调是指在训练大语言模型时，通过向模型提供特定的指令或提示来引导其生成更准确、更具定向性的输出。

在指令微调中，通常会为模型提供一些示例输入和期望的输出，以帮助模型理解特定的任务或领域。通过反复调整微调参数和训练数据，可以使模型逐渐适应特定的任务需求，并生成更符合期望的结果。

指令微调技术在许多领域都有广泛的应用，包括自动摘要、翻译、对话生成等。它能够提高模型在特定领域的表现，并使其更好地适应特定的任务要求。

如果你有关于指令微调技术的具体问题，我会尽力为你解答。



利用指令微调技术让大语言模型快速记住大量的医学知识需要以下步骤：

1. 数据收集：收集医学领域的相关文本数据，包括医学书籍、论文、研究报告、临床指南等。这些数据应涵盖广泛的医学专业知识。
2. 数据预处理：对收集到的数据进行预处理，包括分词、去除噪声、去除非医学领域相关的内容等。确保数据质量和一致性。
3. 构建微调数据集：从预处理后的数据中选择一部分作为微调数据集。这些数据集应包含医学领域的输入示例和对应的期望输出，以指导模型生成正确的医学知识。
4. 定义指令和提示：根据任务需求，为模型定义明确的指令和提示。这些指令和提示应包含关键信息，帮助模型理解问题和生成准确的医学知识。
5. 进行微调训练：使用微调数据集和定义的指令，对大语言模型进行微调训练。微调训练的目标是使模型根据指令和提示生成符合期望的医学知识，并逐步提高模型的表现。
6. 评估和调优：在微调训练过程中，定期评估模型的性能和生成结果。根据评估结果进行调优，包括调整微调参数、增加数据量或改进指令等。

请注意，虽然指令微调技术可以帮助模型更好地理解 and 生成医学知识，但模型生成的结果应始终经过专业医生或医学专家的审查和验证。大语言模型仅提供信息，而不应替代医疗专业人员的判断和决策。

如果我只有结构化的专家医学知识，该如何利用大模型微调技术来把这些结构化的医学知识嵌入到大模型里边呢



如果你拥有结构化的专家医学知识，并希望将其嵌入到大模型中，可以采取以下步骤：

1. 将结构化知识转化为文本格式：将专家医学知识转换为自然语言文本的形式。这可以包括将结构化数据转化为句子或段落，并确保信息的准确性和完整性。
2. 构建微调数据集：根据转换后的文本形式，构建一个微调数据集。这个数据集应包含专家医学知识的输入示例和对应的期望输出。
3. 定义指令和提示：根据任务需求和专家医学知识的特点，为模型定义明确的指令和提示。这些指令和提示应帮助模型理解专家知识并正确生成相关的内容。
4. 进行微调训练：使用微调数据集和定义的指令，对大语言模型进行微调训练。微调的目标是使模型能够根据指令和提示生成符合期望的专家医学知识，并逐步提高模型在该领域的表现。
5. 评估和调优：在微调训练过程中，定期评估模型的性能和生成结果。根据评估结果进行调优，包括微调参数的调整、增加更多的结构化医学知识样本或改进指令等。

通过以上步骤，结合大模型微调技术，你可以尝试将专家医学知识嵌入到大模型中，以扩展模型的理解和生成能力。但请记住，任何由模型生成的结果仍应经过专业医生或医学专家的审查和验证，以确保准确性和可靠性。



是的，你可以尝试直接使用结构化知识而不进行转换成自然语言文本的步骤。虽然大语言模型通常在处理自然语言文本方面表现出色，但在某些情况下，直接利用结构化知识可能更为有效和高效。

对于结构化知识的使用，你可以考虑以下方法：

1. 自定义模型架构：构建一个适合处理结构化知识的自定义模型架构。这可以是基于神经网络的模型，结合了适合处理结构化数据的层和算法。
2. 数据预处理：进行必要的预处理，以使结构化知识适合输入到自定义模型中。这可能包括数据的归一化、特征提取、缺失值处理等。
3. 构建训练集：使用结构化知识构建训练集，其中包含输入数据和期望输出。这可以通过将已有的结构化知识作为训练样本，为模型提供输入和目标输出。
4. 模型训练：使用训练集对自定义模型进行训练，以使其能够从结构化知识中学习和推理。
5. 评估和调优：对训练后的模型进行评估和调优，确保其在处理结构化知识方面的准确性和性能。

通过这些方法，你可以尝试直接利用结构化知识来训练和构建自定义模型，以满足特定的医学知识需求。请记住，在应用模型结果时，始终需要对生成的输出进行验证和审查，以确保其准确性和适用性。

一

大语言模型简介

二

大语言模型与智慧医疗

三

大语言模型微调技术

四

中文医学大模型

大语言模型与智慧医疗

大语言模型的对话式交互使自然语言处理的应用价值显著提升

Hi

一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，应该采取哪些治疗方案？



针对肝胆管结石病，常规的治疗方案有以下几种：



1. 药物治疗：如果结石较小，可以使用药物促进结石排出。一般使用利胆药物或胆汁酸药物，如乙胺丁三醇、奥司他韦等。
2. 内镜治疗：如果结石较大或者药物治疗无效，可以考虑采用内镜下胆管镜取石术（ERCP），通过内镜将镜子插入口腔，经食管、胃镜检查，到达胆管，通过口腔内的钳子取出结石。
3. 外科手术：如果内镜治疗无效或者结石过大，可以考虑开腹手术或腹腔镜手术。
4. 激光治疗：对于一些特殊情况，如肝胆管结石过大或位置不便于内镜治疗等，可以考虑采用激光碎石术，通过激光将结石碎成小颗粒后排出。

需要根据患者的具体情况来制定治疗方案。建议患者在医生的指导下选择合适的治疗方案，并严格遵守医嘱进行治疗。

ChatGPT 3.5

大语言模型与智慧医疗

大语言模型在垂直领域，如医学领域上的生成内容知识性有限

HI

一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，请问有什么治疗方案？



针对肝胆管结石病，常规的治疗方案有以下几种：

1. 药物治疗：如果结石较小，可以使用药物促进结石排出。一般使用利胆药物或胆汁酸药物，如**乙胺丁醇**、**奥司他韦**等。
2. 内镜治疗：如果结石较大或者药物治疗无效，可以考虑采用内镜下胆管镜取石术（ERCP），通过内镜将镜子插入口腔，经食管、胃镜检查，到达胆管，通过口腔内的钳子取出结石。
3. 外科手术：如果内镜治疗无效或者结石过大，可以考虑开腹手术或腹腔镜手术。
4. 激光治疗：对于一些特殊情况，如肝胆管结石过大或位置不便于内镜治疗等，可以考虑采用激光碎石术，通过激光将结石碎成小颗粒后排出。

需要根据患者的具体情况来制定治疗方案。建议患者在医生的指导下选择合适的治疗方案，并严格遵守医嘱进行治疗。

- **乙胺丁醇：抑菌性抗结核药物**
- **奥司他韦：甲型、乙型流感病毒治疗预防药物**

大语言模型与智慧医疗

生成内容知识性受限原因

- 药物等术语在模型训练阶段曝光率有限
- 基于自回归的生成式模型难以掌握疾病、药物等实体间的关系

Hi

一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，应该采取哪些治疗方案？

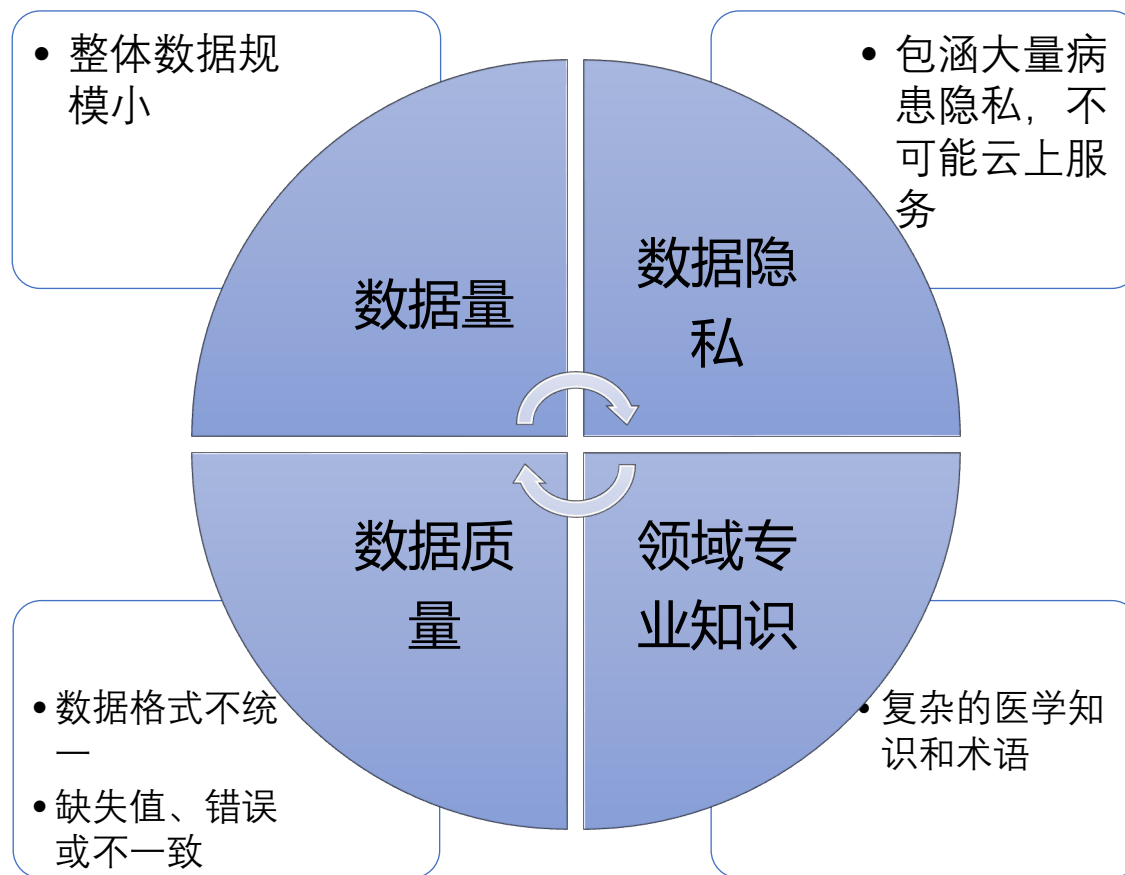


针对肝胆管结石病，常规的治疗方案有以下几种：

1. 药物治疗：如果结石较小，可以使用药物促进结石排出。一般使用利胆药物或胆汁酸药物，如乙胺丁三醇、奥司他韦等。
2. 内镜治疗：如果结石较大或者药物治疗无效，可以考虑采用内镜下胆管镜取石术（ERCP），通过内镜将镜子插入口腔，经食管、胃镜检查，到达胆管，通过口腔内的钳子取出结石。
3. 外科手术：如果内镜治疗无效或者结石过大，可以考虑开腹手术或腹腔镜手术。
4. 激光治疗：对于一些特殊情况，如肝胆管结石过大或位置不便于内镜治疗等，可以考虑采用激光碎石术，通过激光将结石碎成小颗粒后排出。

需要根据患者的具体情况来制定治疗方案。建议患者在医生的指导下选择合适的治疗方案，并严格遵守医嘱进行治疗。

大语言模型与智慧医疗



医疗领域需要数据和算力需求小，客制化和本地部署的私有化大语言模型

一

大语言模型简介

二

大语言模型与智慧医疗

三

大语言模型微调技术

四

中文医学大模型

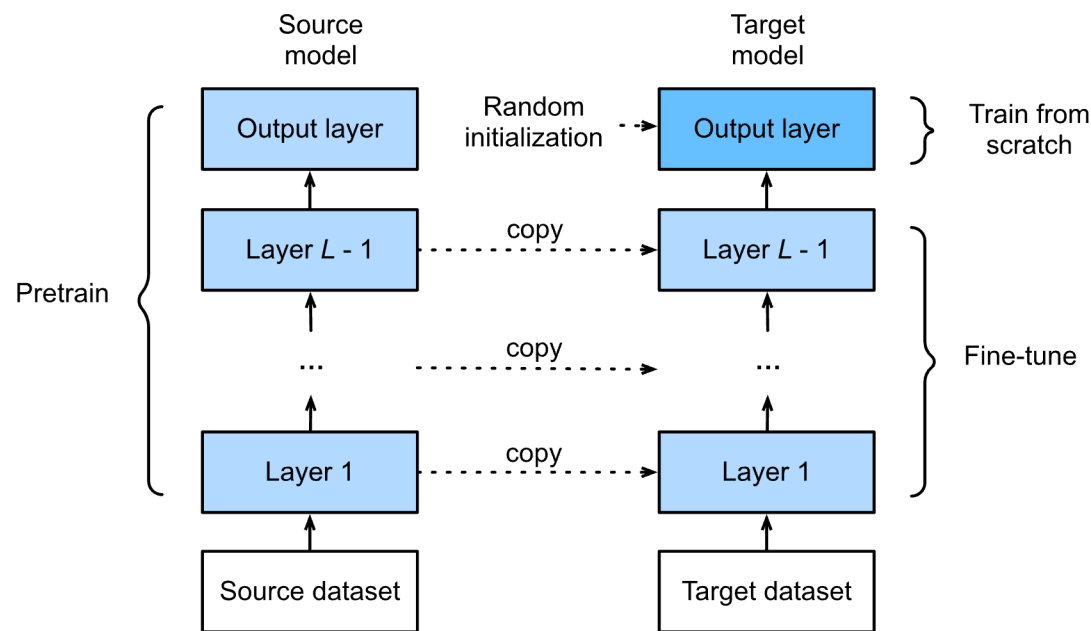
Fine Tuning微调



需要大量的下游特定任务的标注数据

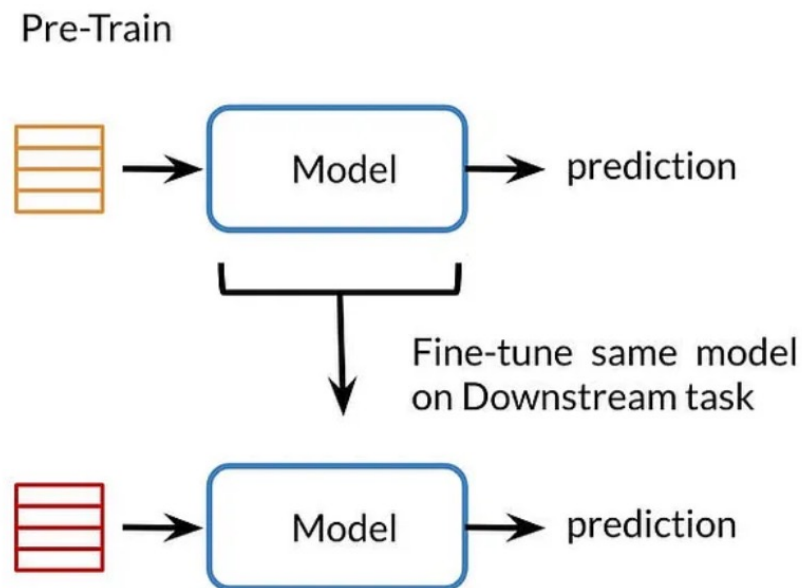
Fine Tuning微调

- 在预训练基础上，针对特定任务或领域做模型参数的更新
 - 在源数据集上做预训练
 - 拷贝每一层的参数
 - 随机初始化输出层参数
 - 在目标数据集上对每一层参数做微调
 - 在目标数据集上从头训练输出层参数



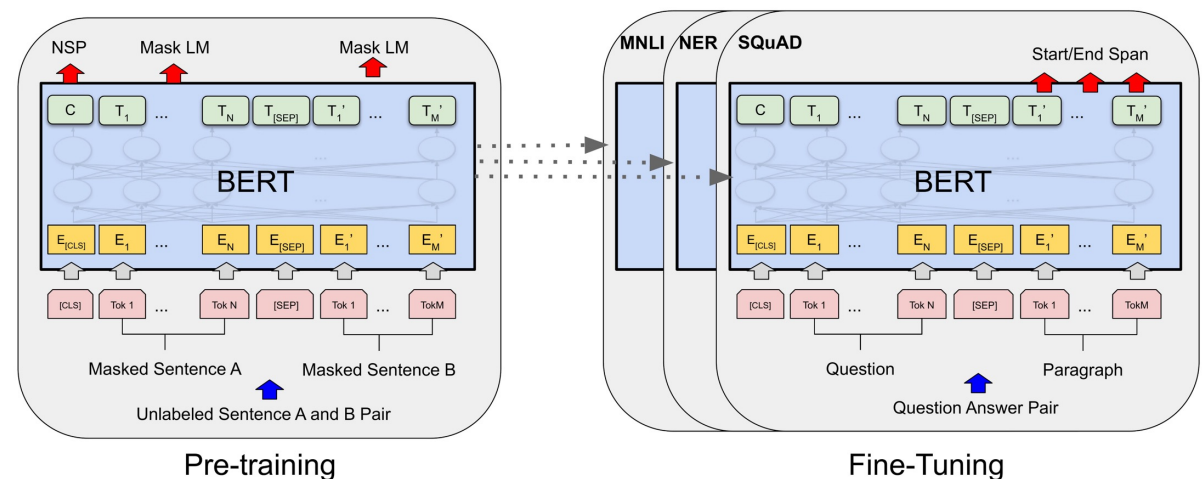
Fine Tuning的优点

- **利用预训练知识：**继承预训练阶段获得的能力
- **快速收敛：**把预训练阶段的知识快速迁移到特定任务中
- **节省计算资源：**显著减少所需的计算资源和时间
- **适应强：**微调阶段可以使用各种类型的任务和标注文本数据
- **改善小样本学习：**在特定任务的小样本数据上可以减轻过拟合问题

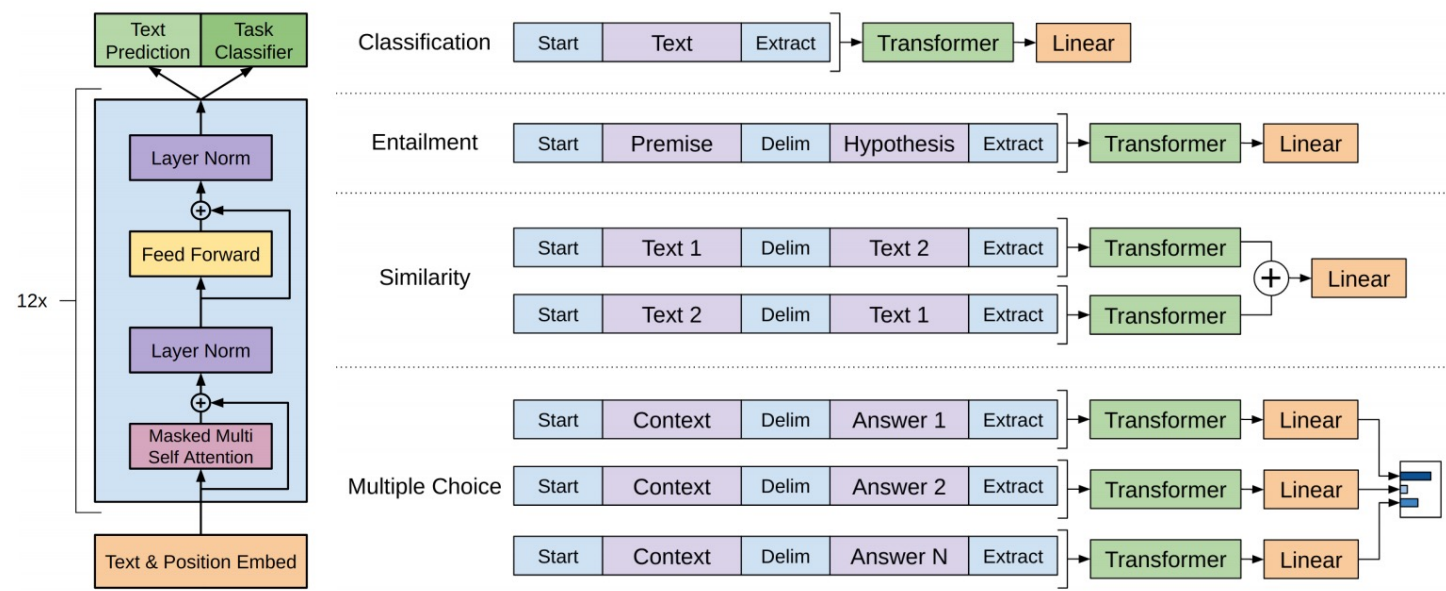


Fine-Tuning应用

BERT



GPT



Fine-tuning的缺点

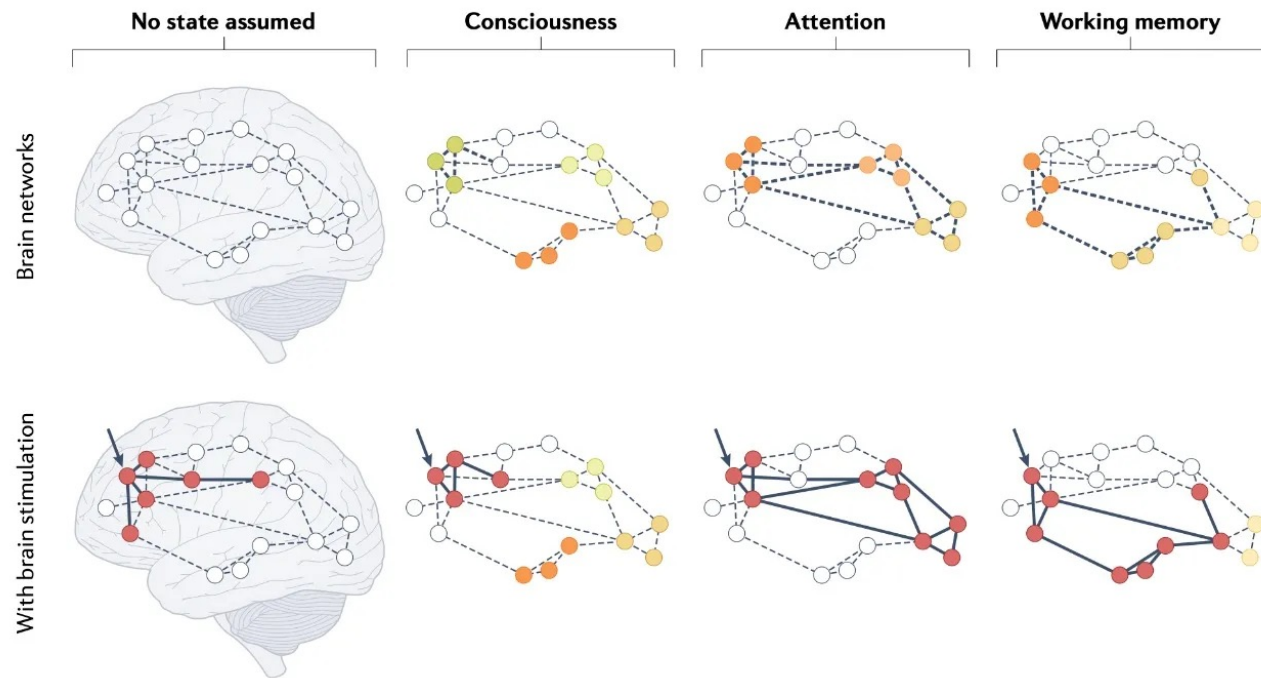
需要微调语言模型所有参数，每个任务都需要保存一遍所有的模型参数

数据需求大： 微调需要大量标注数据来调整模型

灵活性有限： 一旦模型被微调到特定任务，它在其他任务上的表现可能会受到影响。

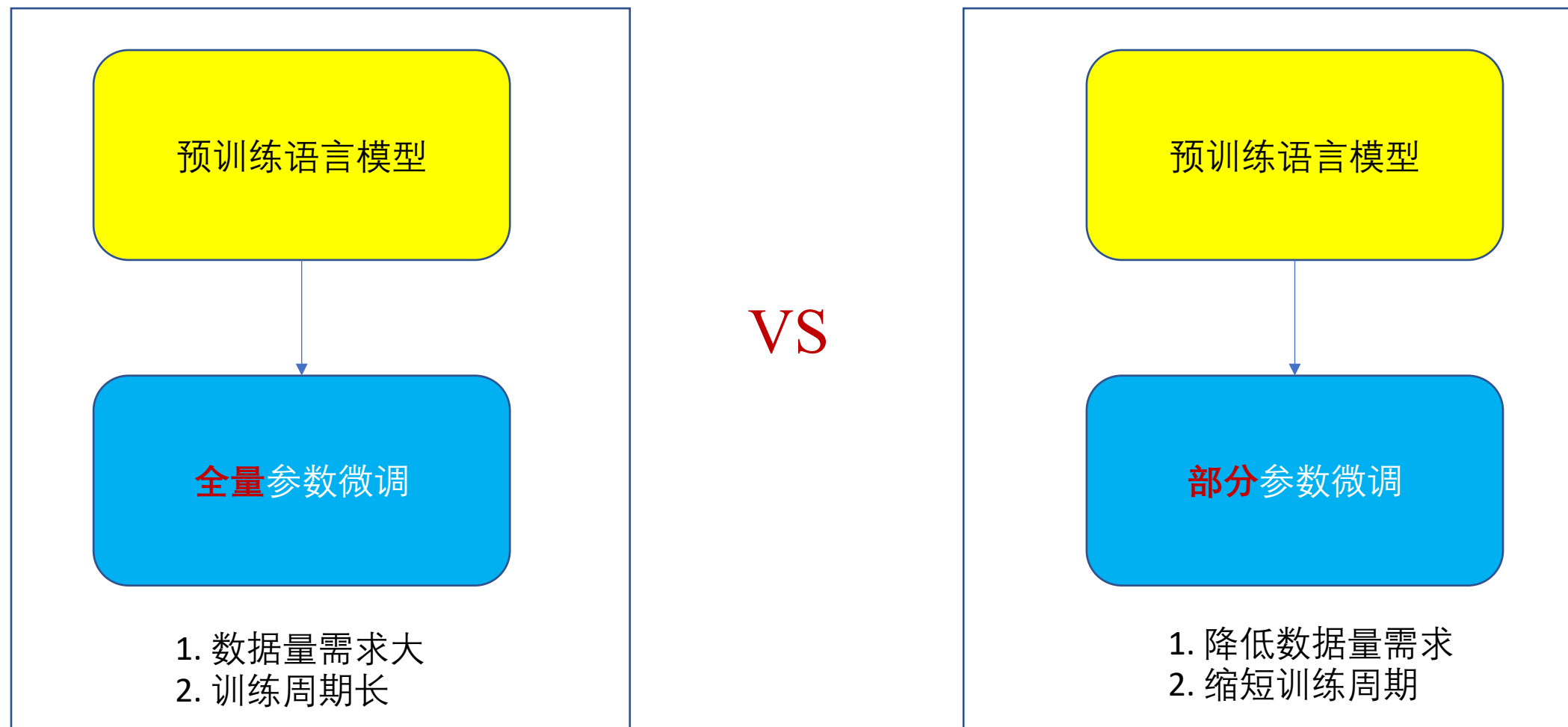
任务迁移困难： 迁移到另一个任务上需要重新进行微调

泛化能力差： 导致模型在特定数据分布上表现出色，但在其他数据分布上的泛化能力较差



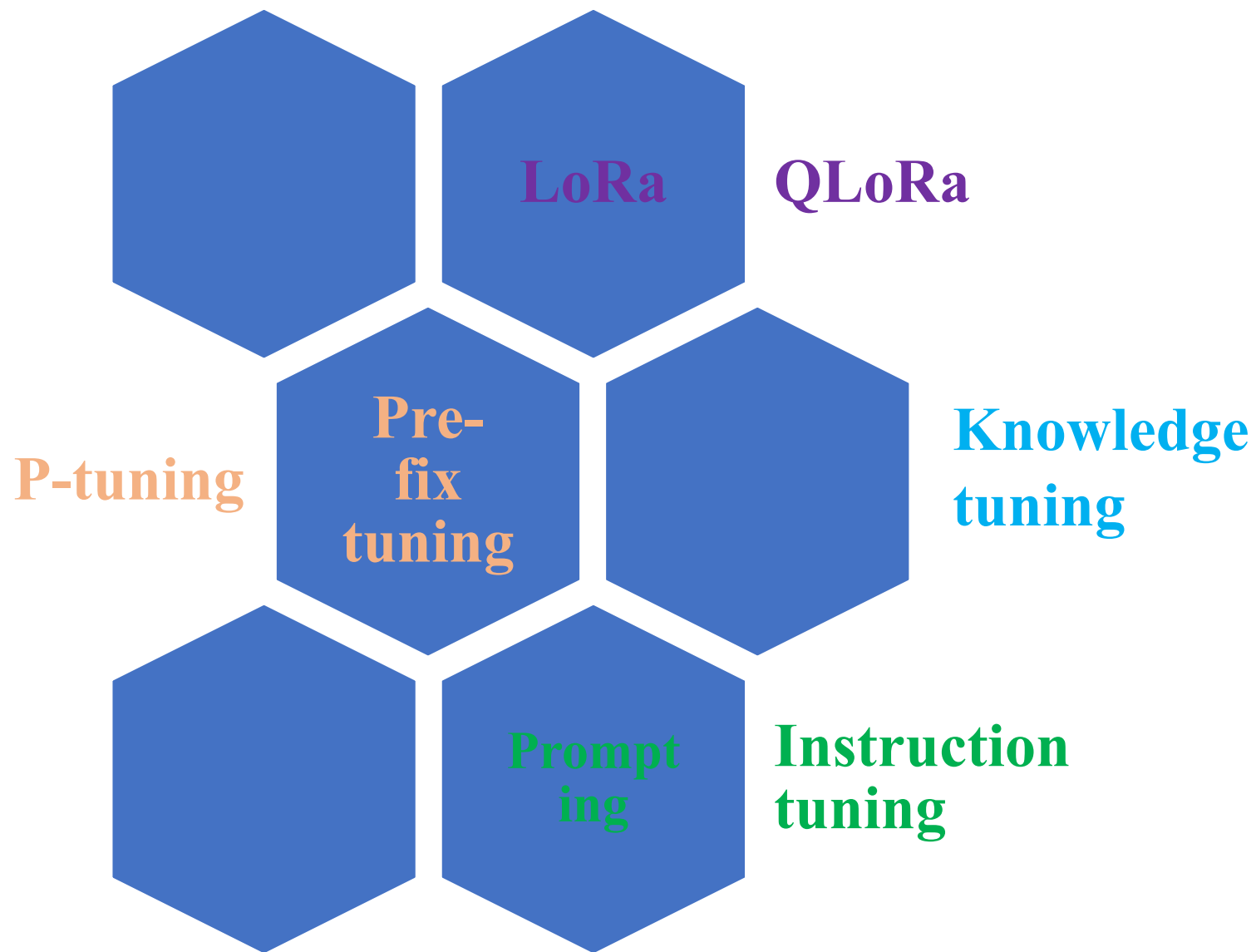
Nature Reviews Neuroscience

全量参数调整和部分参数调整



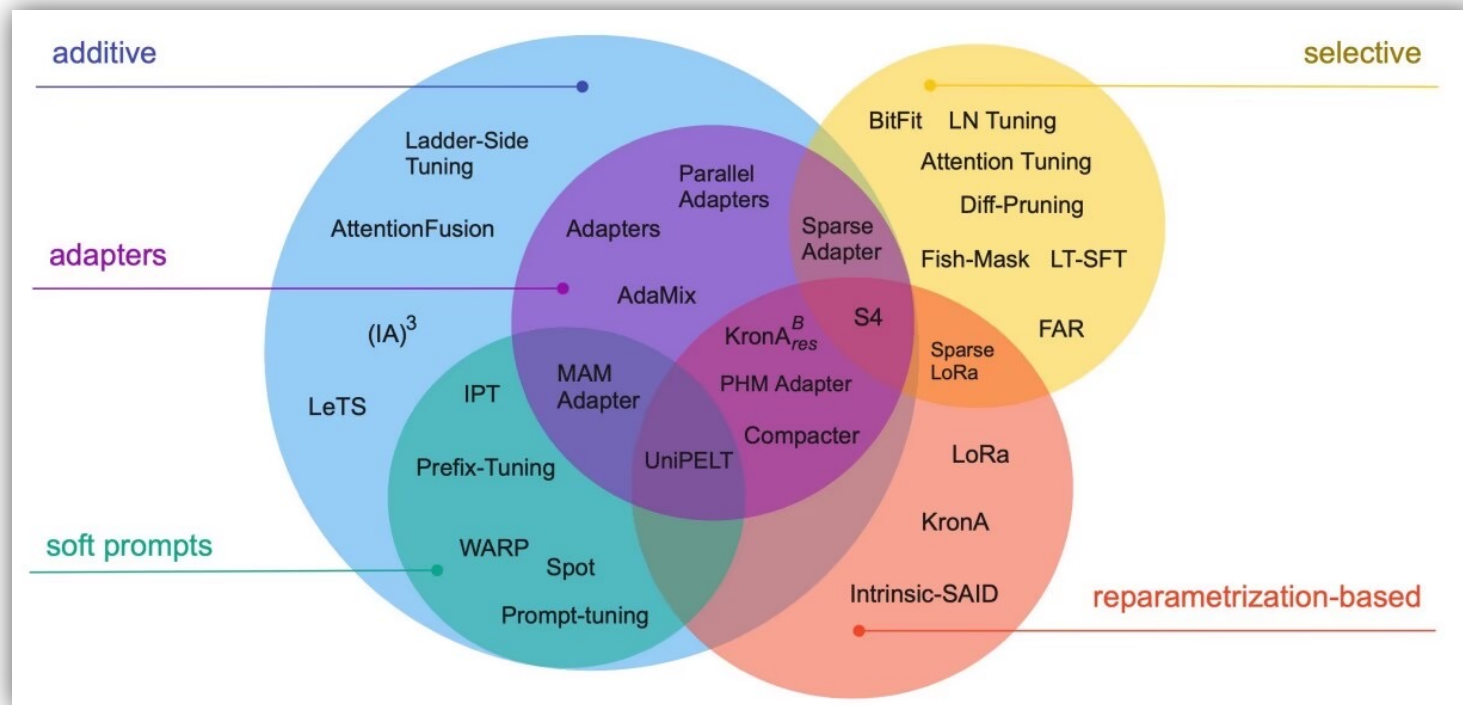
参数高效微调：仅微调模型中的一小部分参数，固定其他参数

参数高效微调方法



P-Tuning 与 Prefix Tuning

随着预训练语言模型(PLM)规模的日益增大,在消费级硬件和有限的标注数据上对整个模型进行全量微调变得不太可行。为了解决这个问题,参数高效迁移学习(Parameter-efficient Transfer Learning, PEFT)领域出现了多种新思路。p-tuning和prefix-tuning正是PEFT中的两大代表技术。



Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning[J]. arXiv preprint arXiv:2303.15647, 2023.

Prefix Tuning

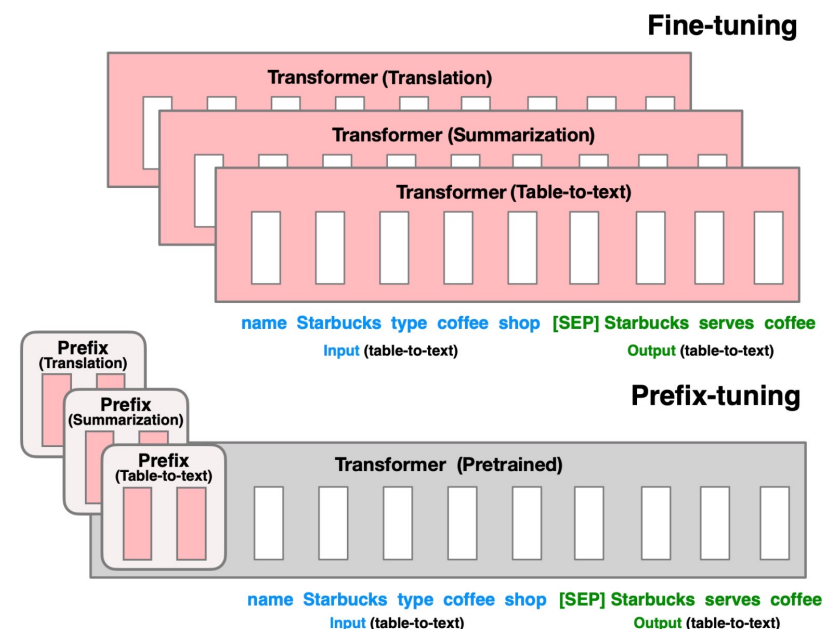
背景：随着大模型如 GPT-3 的迅速流行，进行大模型的预训练或全量微调变得遥不可及。因此，需要一种参数高效的微调技术，让科研人员或开发者有机会尝试微调大模型

- **解决的技术问题：**

- **参数效率：**通过只更新前缀部分的参数，降低了微调的计算成本。
- **训练稳定性：**通过在前缀层前加入 MLP 结构，防止训练不稳定和性能下降。

- **具体原理：**

- **虚拟前缀：**在输入 token 之前构造任务相关的虚拟 token 作为前缀。
- **参数更新：**在训练时只更新前缀部分的参数，其他部分参数固定。
- **结构适配：**针对不同的模型结构，如自回归架构和编码器-解码器架构，构造不同的前缀。



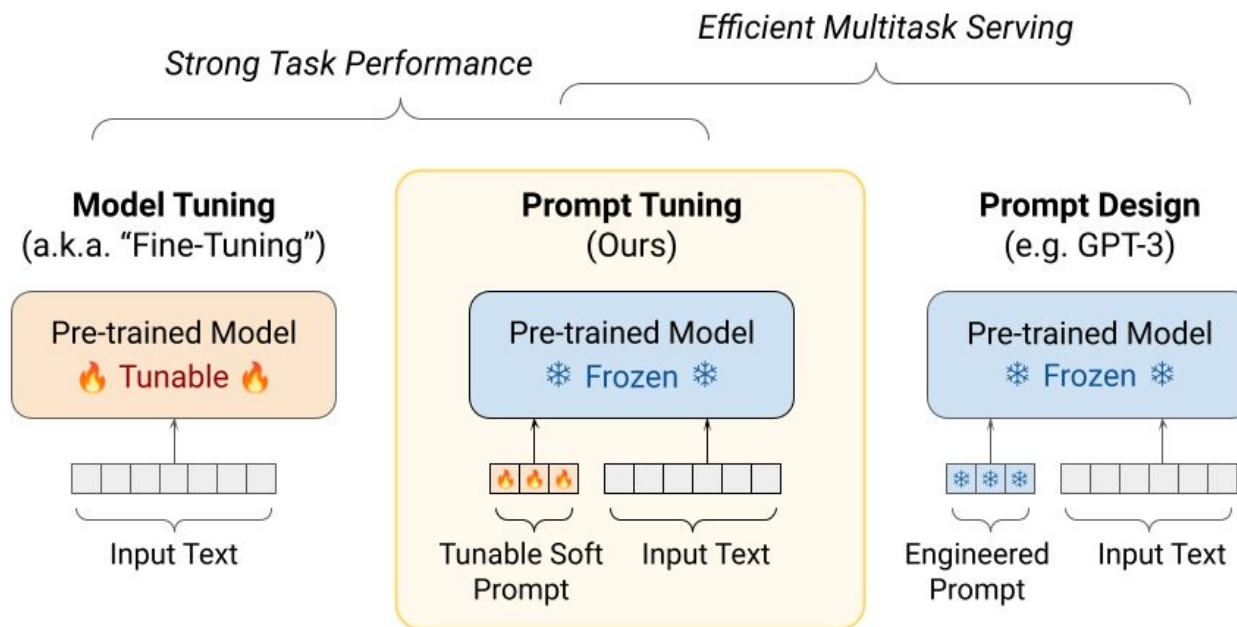
Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

P-Tuning (Prompt Tuning)

背景：P-Tuning 主要解决大模型的 Prompt 构造方式对下游任务效果的敏感性问题。人工设计的模板变化特别敏感，导致性能不稳定。

- 解决的技术问题：

- **模板敏感性**：通过连续可微的 tunable soft prompt 解决人工设计模板的敏感性问题
- **参数效率**：实现了参数高效的微调



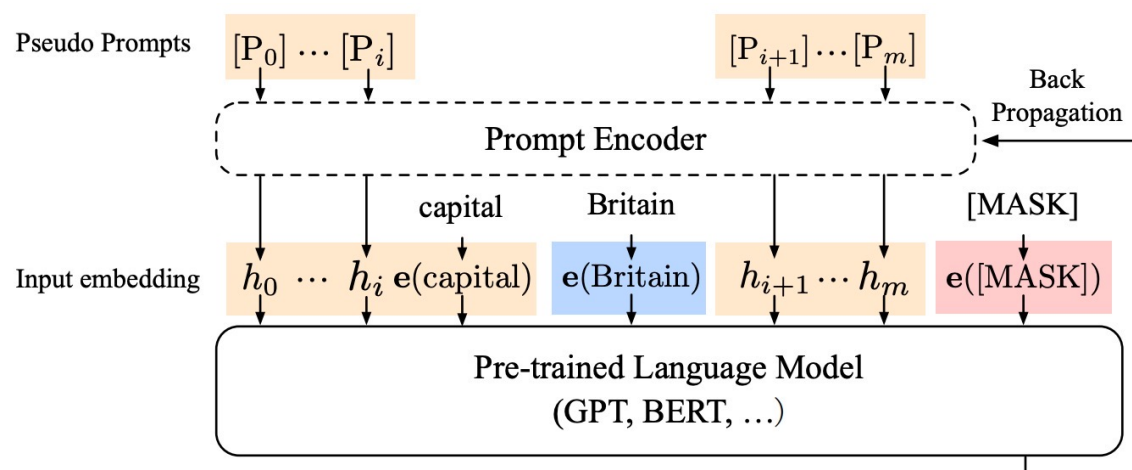
P-Tuning

背景：P-Tuning 主要解决大模型的 Prompt 构造方式对下游任务效果的敏感性问题。人工设计的模板变化特别敏感，导致性能不稳定。

- **具体原理：**

- **连续可微的 tunable soft prompt**：与 Prefix-Tuning 类似，设计了连续可微的 tunable soft prompt，但 P-Tuning 的 tunable soft prompt 仅限于输入层，位置可选，不一定是前缀。
- **Prompt 转换为可学习的 Embedding**：将 Prompt 转换为可学习的 Embedding 层，并用 MLP+LSTM 的方式进行处理

- **与 Prefix-Tuning 的区别**：P-Tuning 主要针对 Prompt 构造方式的敏感性问题，虽然两者都使用了参数可调的伪token，但 P-Tuning 的使用范围和位置与 Prefix Tuning 有所不同



(b) P-tuning

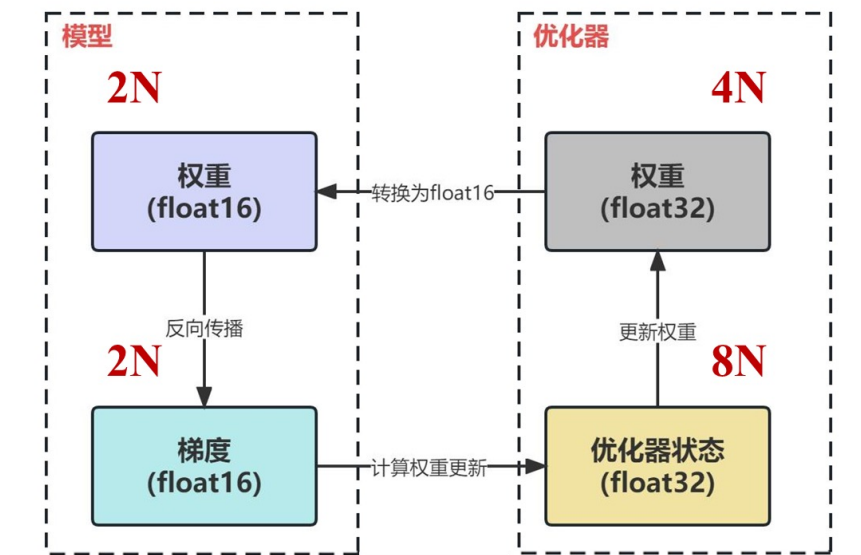
LoRA与QLoRA

□ 研究背景

1. 模型参数越来越大(GPT-3 175B参数)
2. 显存不足严重影响了模型的训练、部署

假设模型参数量为N:

- 权重空间 + 梯度空间 + Adam优化器: $16N$
- 175B的模型需要2.8TB显存



□ 动机与目标

1. 降低可训练的参数量
2. 降低模型权重占用的空间
3. 降低优化器占用的空间



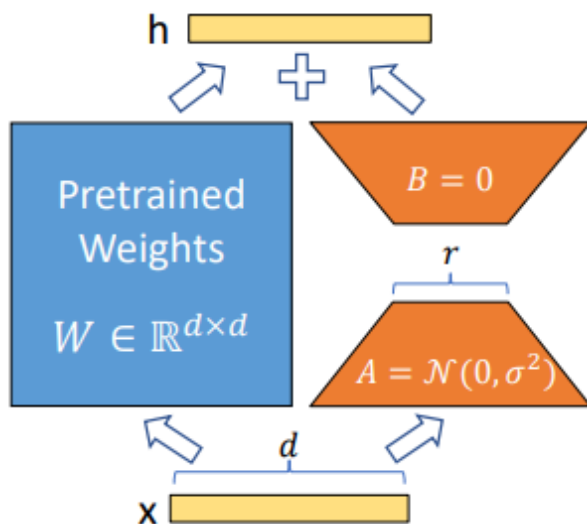
降低模型训练与部署所需要的空间!

让更多的研究人员与企业能训练自己的大模型!

LoRA: Low-Rank Adaptation of Large Language Models

□ 核心思想

1. 模型参数矩阵本质都是低秩矩阵
2. 因此用两个低秩矩阵乘积近似完整的矩阵



$$h = W_0x + \Delta Wx = W_0x + BAx$$

□ 优化效果

- 模型权重 (2N) → 2N
- 模型梯度 (2N) → 0.0002N
- Adam优化器状态 (8N) → 0.0004N
- Adam优化器权重副本 (4N) → 0.0008N

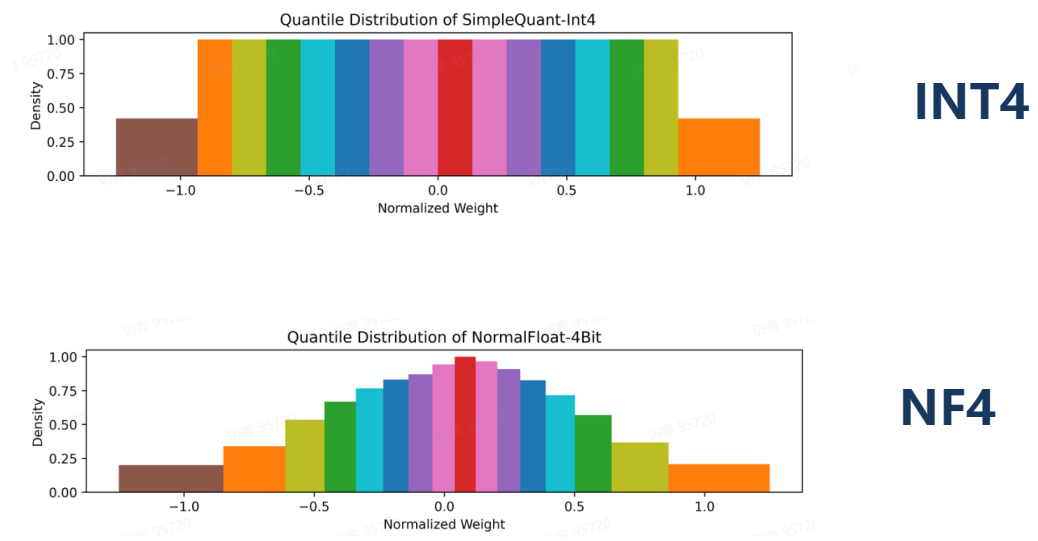
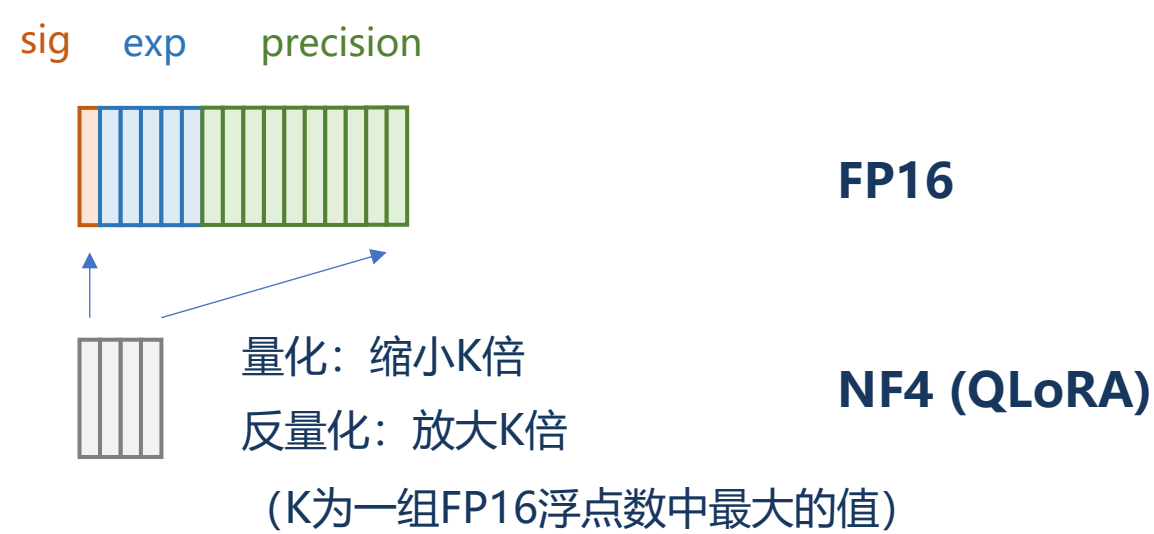
总计: $2N + 0.0012N \approx 2N$

- 1B的模型需要的显存从16GB降低到了2GB
- 65B的模型需要的显存从1040GB降低到了130GB

QLoRA: Efficient Finetuning of Quantized LLMs

□ 核心思想

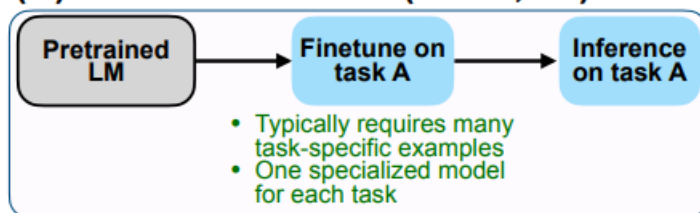
1. 通过将16bit浮点数缩小K倍，转换为4bit下能表示的16个阈值之一。以损失精度为代价，用4bit的空间存储16bit的模型权重，减少了75%的存储空间占用！
2. 提出了NF4(Normal distribution Float 4bit), 理论最优的量化方式：按照正态分布设置16个阈值，与模型参数的分布最为接近，损失的精度最小！



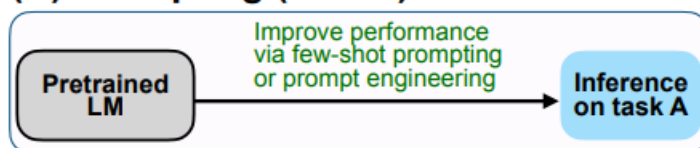
Prompting和Instruction Tuning

- Prompting关注于模型在**单一任务**上的性能优化，侧重**In-Domain**领域
- Instruction Tuning关注于模型在**未见过任务**上的泛化性能优化，侧重**Out-of-Domain**领域

(A) Pretrain–finetune (BERT, T5)



(B) Prompting (GPT-3)



(C) Instruction tuning (FLAN)

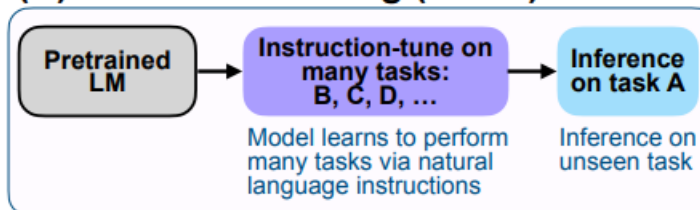
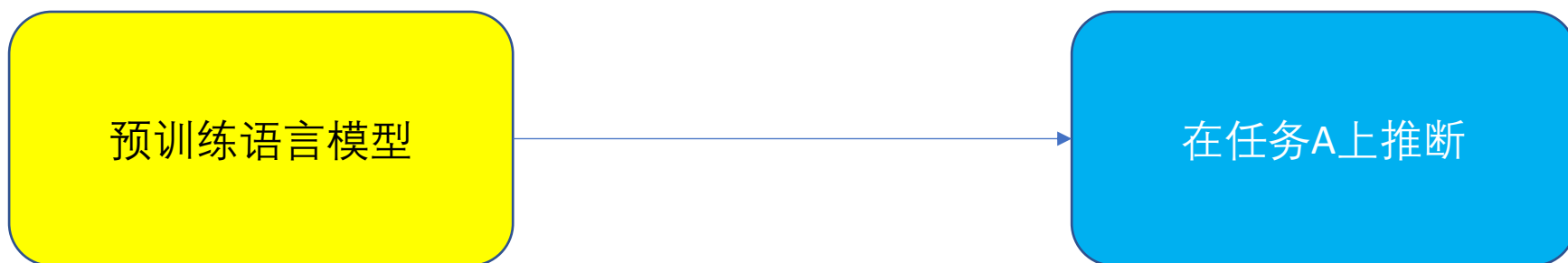


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

Prompt Learning和Instruction Tuning的区别

Prompting



需要少量的下游特定任务的标注数据

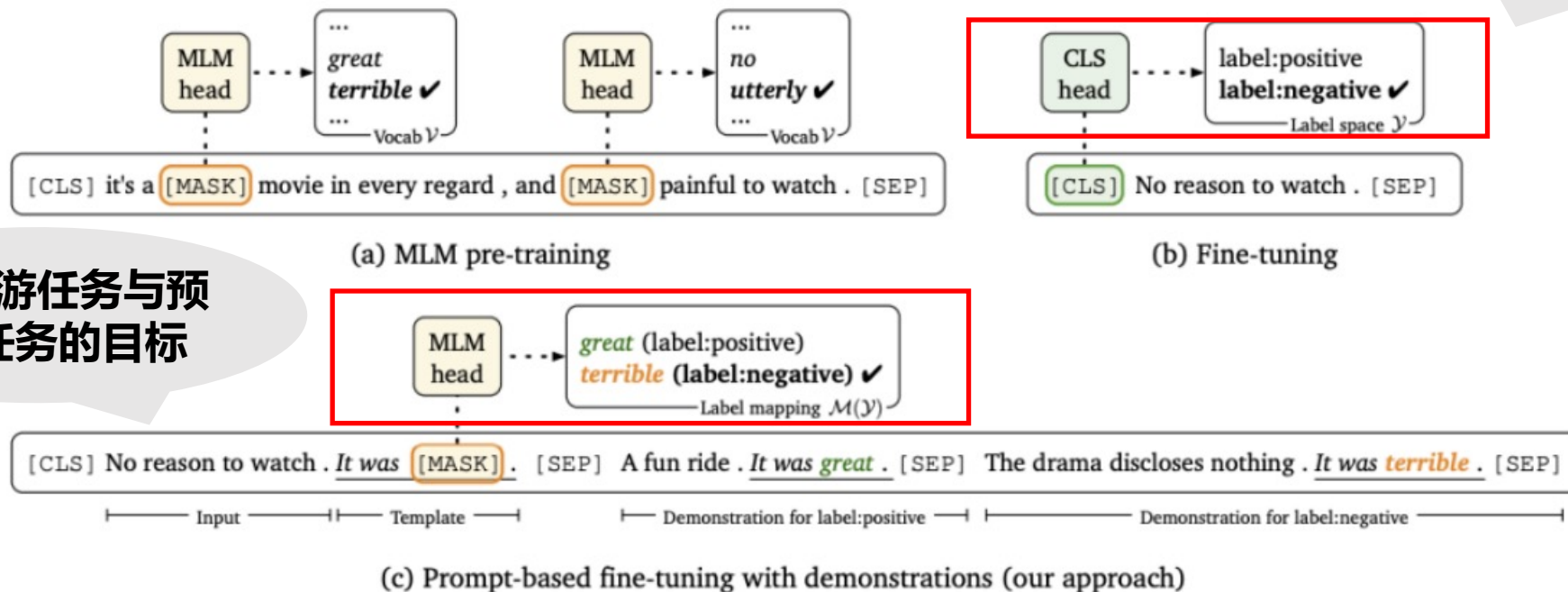
Prompting

- Prompting: 激发模型的**补全能力**，通过添加模板，将下游任务目标转化为预训练任务目标

- 统一下游任务与预训练任务的目标，减轻了语义鸿沟

- 避免额外参数的引入，克服少样本场景下模型容易过拟合的问题

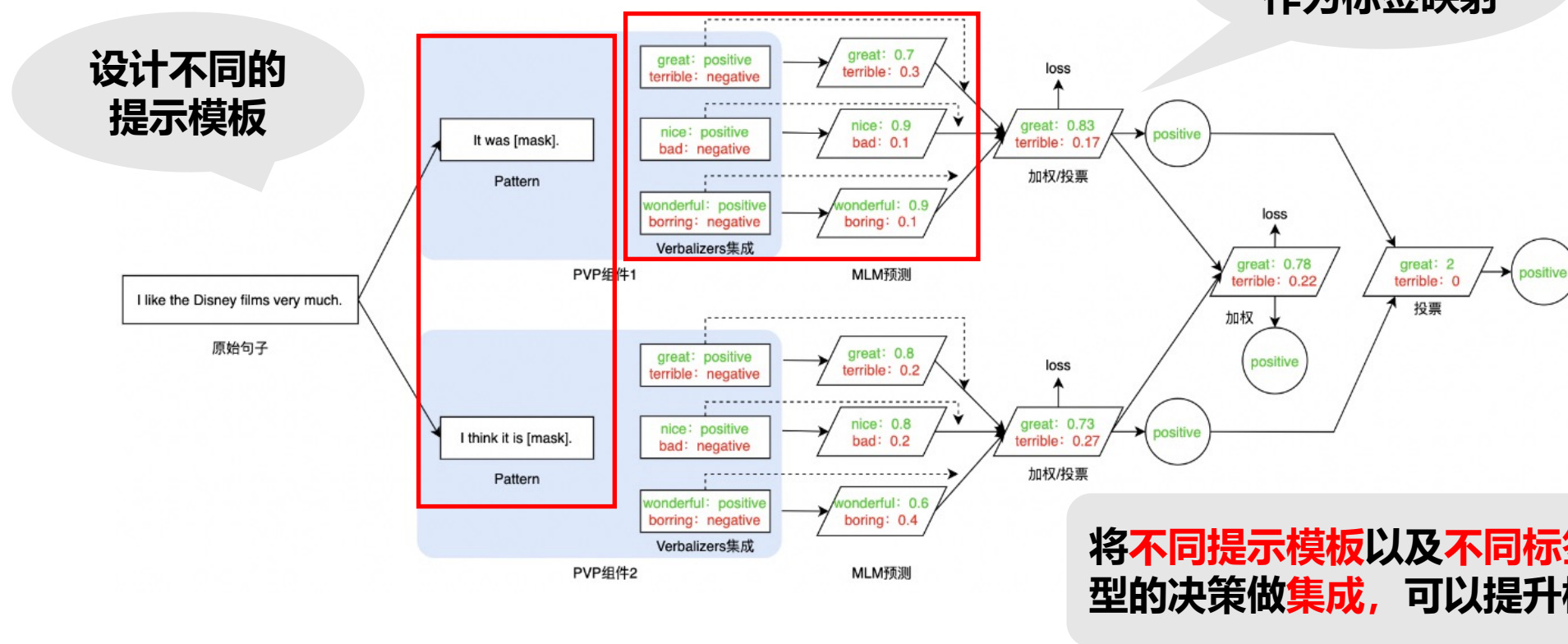
减少了额外参数的引入



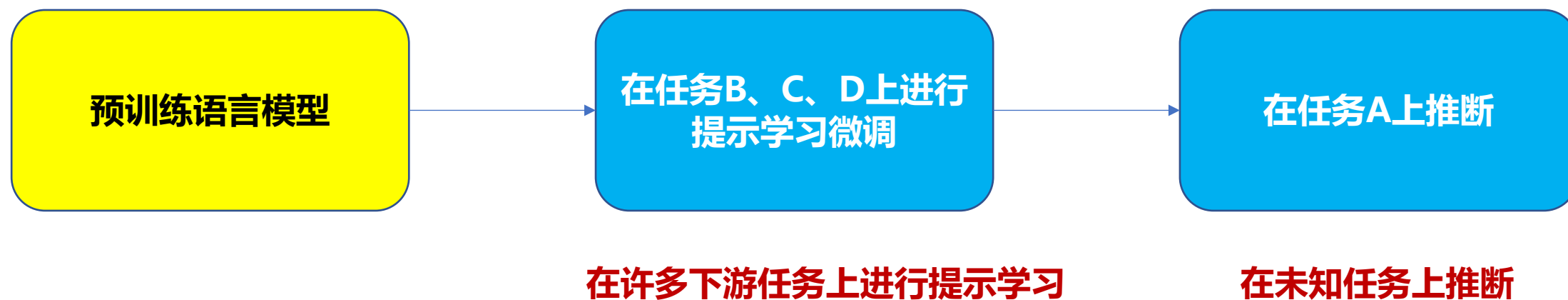
统一下游任务与预训练任务的目标

Prompting

- 代表性工作PET (Pattern-Exploiting Training)
 - 设计重要组件: Pattern-Verbalizer-Pair (PVP)
 - Pattern指提示模板; Verbalizer指标签空间的映射

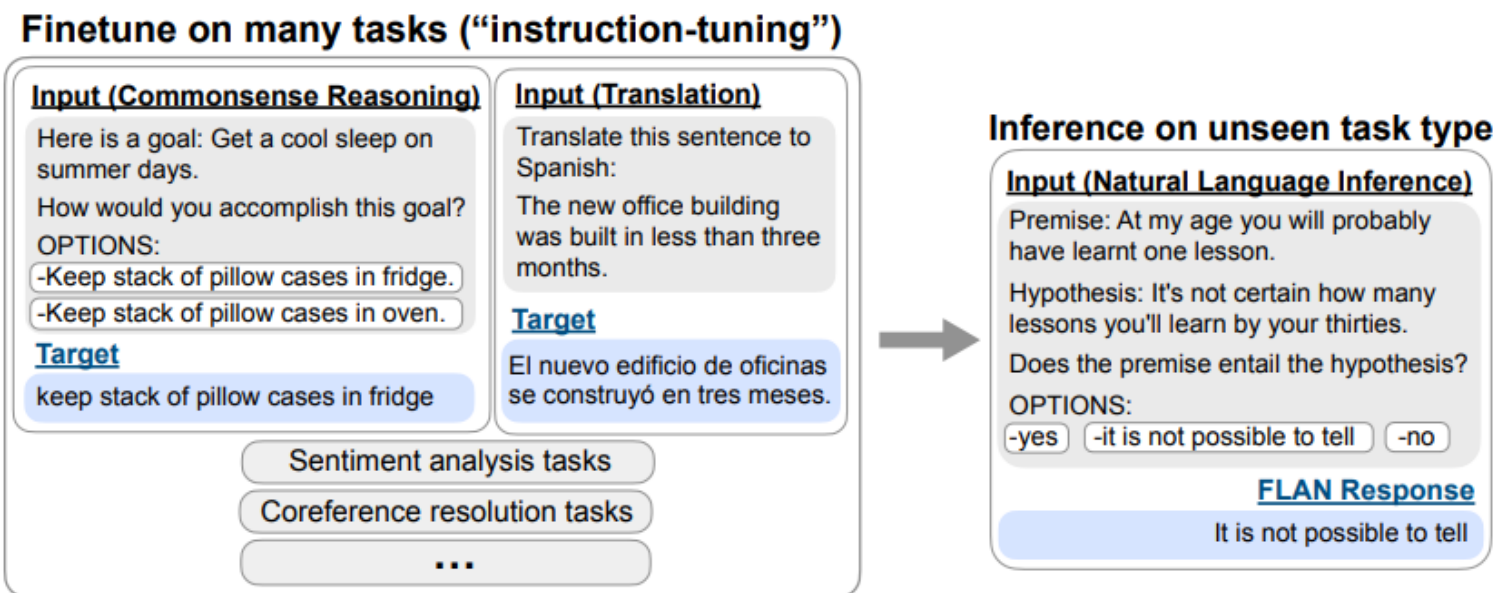


Instruction Tuning



Instruction Tuning

- **Instruction Tuning**: 发掘语言模型**本身具备的知识**，激发模型的**理解能力**，通过给出构建各个任务下的指令，让模型学会理解指令去做出正确的行动
 - 通过给出更明显的指令/指示,让模型去理解并做出正确的回复

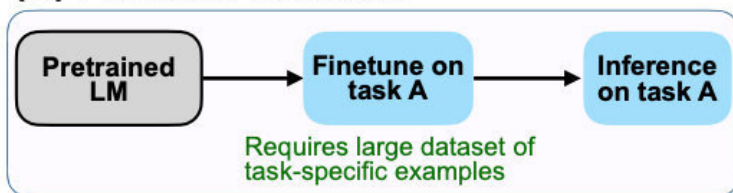


在B、C、D等任务上构建指令微调数据，在任务A上对模型进行测试

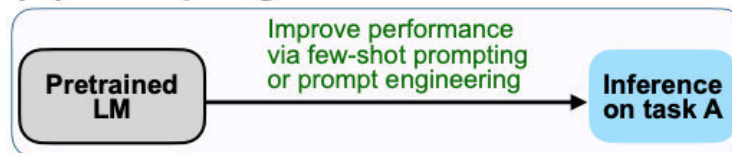
Instruction Tuning

- 经过Instruction Tuning多任务精调后，可以用于其他任务的zero-shot

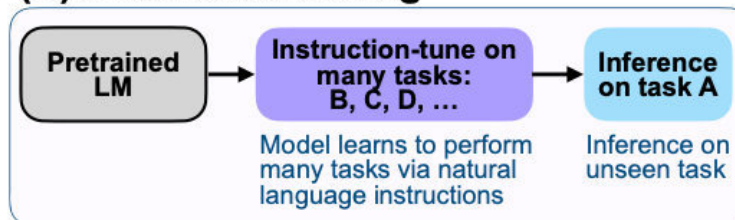
(A) Pretrain-finetune



(B) Prompting



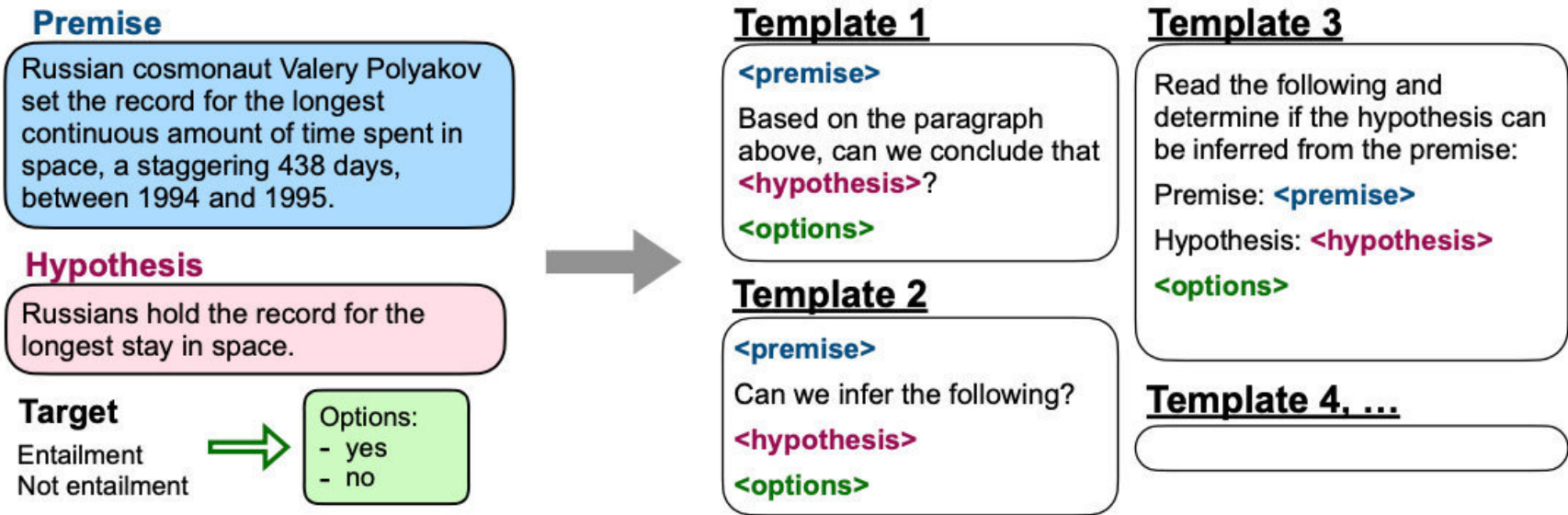
(C) Instruction tuning



- Prompting在没精调的模型上也能有一定效果，而Instruction Tuning则必须对模型精调，让模型知道这种指令模式

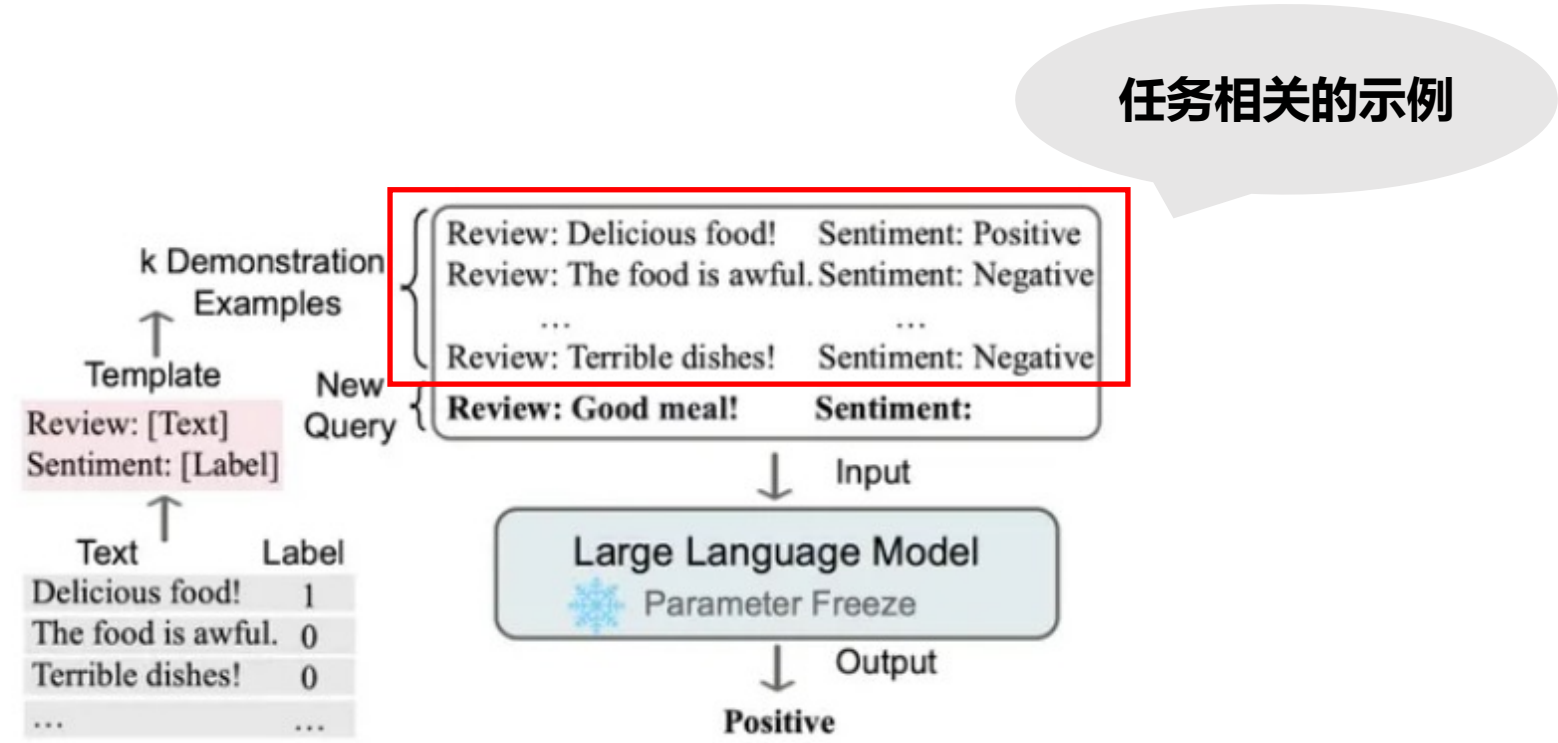
Instruction Tuning

- 需要为每个任务设计指令模版



Instruction Tuning

- 为了帮助模型更好的理解指令，衍生成出In-Context Learning (ICL)
- ICL通过添加任务相关的若干示例，从而帮助模型更好的理解任务，提升其预测效果



一

大语言模型简介

二

大语言模型与智慧医疗

三

大语言模型微调技术

四

中文医学大模型

中文医学大模型“本草”（原名华驼）

- **首个（开源）** 的中文医学大模型
- 首次提出**知识微调**（Knowledge-tuning）方法
 - ✓ 把通用大模型快速改造成行业大模型的方法
- 2023年3月31日在**GitHub**开源
- 在GitHub获得超过**3.7K+** 星标



BenTsao (本草)

HIT-SCIR
Health Intelligence



<https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese>

中文医学大模型“本草”（原名华驼）

➤ 基于大模型并结合医学知识进行指令微调数据集构建

- ✓ Prompt: “给定上述知识，生成一个与知识相关的问题，并进行回答”
- ✓ 基于结构化知识的生成一定程度保证生成内容知识正确性
- ✓ 结构化知识样例：

{ “class”: “百种常见病”, “中心词”: “肝癌”, “药物治疗”: [“瑞格非尼”, “对乙型或丙型肝炎有效的抗病毒药物”, “索拉非尼”], “多发地区”: [“撒哈拉以南的非洲”], “高危因素”: [“肥胖”, “HBV DNA 过高”, “慢性酗酒”, “男性”, “慢性乙型肝炎感染”, “肝癌家族史”, “慢性丙型肝炎肝硬化”, “核心启动子突变”, “肝硬化”, “HCV 重叠感染”, “老年性心瓣膜病”, “乙型肝炎 e 抗原”, “糖尿病”], “化疗方案”: [“化疗栓塞”, “T 细胞免疫治疗”], “遗传因素”: [“一级亲属的肝癌家族史”], “相关导致”: [“发热”], “多发群体”: [“男性”, “40~50 岁男性”], “影像学检查”: [“腹部增强 CT”, “骨扫描”, “肝癌超声监测”, “腹部增强 MRI”, “超声检查”, “胸部 CT 扫描”, “腹部影像”, “肝脏超声”], “发病性别倾向”: [“男性较女性多见”], “ICD-10”: [“C22.9”, “C22.0”], “病因”: [“与肝硬化”, “吸烟”, “酗酒”, “黄曲霉毒素”, “病毒性肝炎以及黄曲霉素等化学致癌物质有关”, “钆造影剂”, “原发性硬化性胆管炎”, “口服避孕药”, “迟发性皮肤卟啉症”, “雄激素类固醇”, “原发性胆汁性肝硬化”, “血色病”, “非酒精性脂肪性肝炎”], “所属科室”: [“腹部损伤”], “并发症”: [“肿瘤破裂导致腹腔内出血”, “贫血”], “治疗方案”: [“核糖核酸染色”, “顺铂”, “NA”, “替加氟”, “氟尿苷”, “氟尿嘧啶”, “阿霉素”, “丝裂霉素”], “临床症状及体征”: [“贫血”, “年龄”, “淋巴结肿大”, “低血糖症”], “筛查”: [“酒精性脂肪性肝炎相关肝硬化”, “遗传性血色病患者”, “乙型和丙型肝炎相关肝硬化患者”, “乙型肝炎病毒 (HBV) 携带者”, “酒精性肝硬化”, “原发性胆汁性肝硬化”] }

中文医学大模型 “本草” （原名华驼）

- 同参数规模条件下，本草模型回答知识准确性显著更强；
- 十亿级参数量的本草在经过指令微调和与千亿级模型有一定可比性。

测试输入	Llama输出 - Facebook (十亿级参数)	Alpaca输出 - Stanford (十亿级参数)	BenTsao本草输出 - 哈工大 (十亿级参数)	ChatGPT输出 - OpenAI (千亿级参数)
一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，应该采取哪些治疗方案？	肝胆管结石病患者应该采取哪些治疗方案？	患者应接受手术和化疗相结合的治疗。	针对肝胆管结石病，治疗方案包括手术治疗、药物治疗和辅助治疗等。具体治疗方案需要根据患者的具体情况而定。	对于肝胆管结石病的患者，治疗方案应该根据具体情况进行选择。一般而言，如果结石较小且没有引起症状，可以采取观察和生活方式改变等保守治疗措施。如果结石较大或引起症状明显，可能需要进行手术治疗，包括腹腔镜下胆囊切除术、经皮穿刺胆管取石术等。在治疗过程中还需注意饮食、控制疼痛、预防并发症等。治疗方案应该由医生根据具体情况制定。
麻风病和儿童哮喘的病因是否一致？	儿童哮喘和麻风病的病因是一致，但是儿童哮喘和麻风病的病因不一致	是的，哮喘和儿童喘息有相似的原因。	不一致，麻风病的病因是麻风杆菌感染，而儿童哮喘的病因是营养不良、吸烟等因素。	麻风病和儿童哮喘的病因不一致。麻风病是由麻风杆菌感染引起的慢性传染病，而儿童哮喘则是由多种因素引起的慢性呼吸道疾病，包括过敏、环境因素和遗传因素等。两者的发病机制和治疗方法也不同，需要根据具体情况进行诊断和治疗。
小张最近感觉身体不适，出现心悸、气促等症状。体检发现心脏扩大、搏动减弱。	问题：小张最近感觉身体不适，出现心悸、气促等症状。体检发现心脏扩大、搏动减弱。	最近，小张一直感觉身体不舒服，有心烦意乱、呼吸急促等症状。体格检查显示心脏增大，肌肉无力。	小张可能患有心肌炎，建议尽快到医院进行检查和治疗。治疗方案包括药物治疗和手术治疗。	小张可能患有心脏疾病，需要进一步的检查和治疗。心脏扩大和搏动减弱是心衰的常见表现，建议尽快就医，并听从医生的建议进行治疗。

知识微调 (Knowledge tuning)

研究问题

在大语言模型的背景下，指令微调成为将大语言模型快速适配垂直领域的解决方案，然而对于知识正确性要求较高的领域，如医学，指令微调后的模型回复依旧可能存在错误

Question

In English:
A patient with persistent distension pain and jaundice has been diagnosed with **hepato-biliary stones** disease after imaging examination. What treatment options should be taken?

In Chinese:
一位患者出现了持续性的胀痛和黄疸，经影像学检查后确诊为肝胆管结石病，应该采取哪些治疗方案？



Answer for the question in English

..... The following are some possible treatment options:
5. Medications: Medications such as **bile acid therapy or ursodeoxycholic acid** may be used to dissolve certain types of gallstones, but they are generally not effective for hepatobiliary stones.

Answer for the question in Chinese

以下是一些可能的治疗方案：
1. 药物治疗：对于较小的结石，可以使用药物来溶解它们，例如**利福平**或酸碱疗法等。.....
(Here are some possible treatment options:
1. Medication: For smaller stones, medications such as **rifampicin** or acid-base therapy may be used to dissolve them.)

解决方案

提出知识微调方法，令大语言模型通过调用知识函数的方法获得输入样例相关的医学知识，用以辅助模型输出

Question:

一个5岁的男童最近有一段时间频繁发生耳痛、听力下降等症状，经检查确诊为急性中耳炎。请问医生应该如何治疗？
(A 5-year-old boy has recently been experiencing symptoms such as frequent ear pain and decreased hearing for some time. After examination, he was diagnosed with acute otitis media. What is the recommended treatment plan for the doctor?)

Key entity: 儿童急性中耳炎
(Pediatric Acute Otitis Media)
Attribute: 治疗方案
(Treatment Plan)

Stage 1: Parameter Filling

`acquire_knowledge(entity='儿童急性中耳炎', attribute='治疗方案')`

```
def acquire_knowledge(entity, attribute):  
    .....  
    return knowledge
```

Stage 2: Knowledge Function Call

Question:

一个5岁的男童最近有一段时间频繁发生耳痛、听力下降等症状，经检查确诊为急性中耳炎。请问医生应该如何治疗？
Knowledge: "儿童急性中耳炎": {

"治疗方案": ["全身治疗", "抗生素治疗", "局部治疗", "观察"]
"药物治疗": ["阿莫西林", "头孢菌素", "大环内酯类", "镇痛药"]
{ "Pediatric Acute Otitis Media": {
"Treatment Options": ["Systemic Treatment", "Antibiotic Treatment", "Local Treatment", "Observation"],
"Medication Treatment": ["Amoxicillin", "Cephalosporins", "Macrolides", "Analgesics"]}}

Output:

根据当前的治疗标准，初期不需要使用药物，而是以观察为主。但如果症状在48小时后没有改善，医生应该考虑使用抗生素进行治疗，通常建议使用阿莫西林等抗生素。

相关知识：儿童急性中耳炎 “: {” 治疗方案 “: {” 全身治疗 “, ” 抗生素治疗 “, ” 局部治疗 “, ” 观察 “, ” 药物治疗 “: {” 阿莫西林 “, ” 头孢菌素 “, ” 大环内酯类 “, ” 镇痛药 “}}

(According to current treatment standards, medication is not necessary for the early stages of acute otitis media, and observation is the main approach. However, if symptoms do not improve after 48 hours, the doctor should consider using antibiotics for treatment. Amoxicillin and other antibiotics are usually recommended.

Relevant knowledge: "Pediatric Acute Otitis Media": { "Treatment Options": ["Systemic Treatment", "Antibiotic Treatment", "Local Treatment", "Observation"], "Medication Treatment": ["Amoxicillin", "Cephalosporins", "Macrolides", "Analgesics"] }

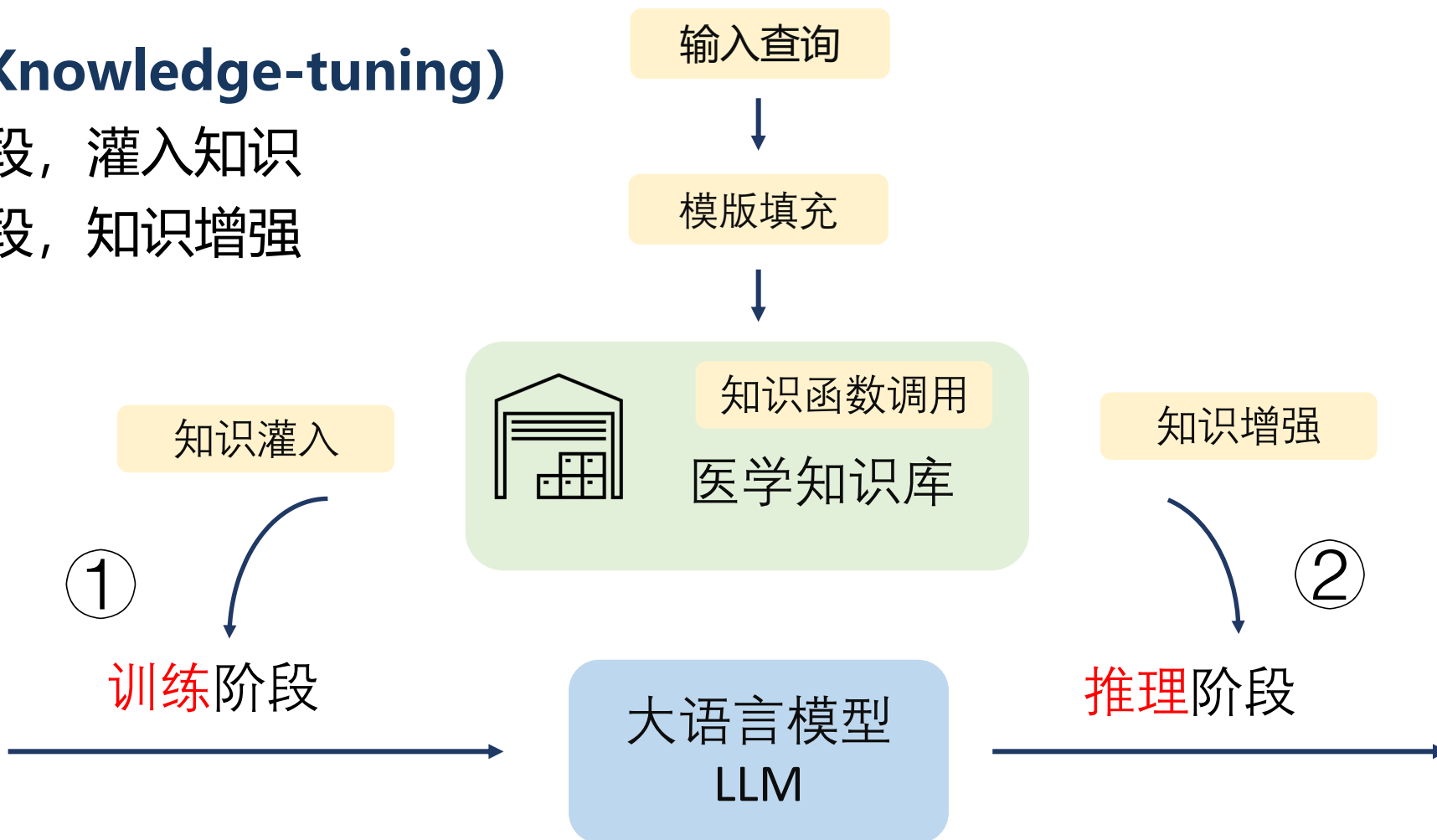
Stage 3: Response Generation with Knowledge

实验证明知识微调可以显著提升模型回复知识性，并可以提供模型输出相应的知识依据

知识微调 (Knowledge tuning)

➤ 知识微调 (Knowledge-tuning)

- ✓ 训练阶段，灌入知识
- ✓ 推理阶段，知识增强



通过引入**领域知识**，提升模型**输出内容**中的**知识精确性**

方法：训练阶段

- 结合医学**知识**的指令微调**数据生成**
 - 医学知识**：知识图谱，医学文献，.....
 - Prompt**：“给定上述知识，生成一个与知识相关的问题，并进行回答”
- 医学**知识**的指导可提升生成**指令微调数据**的**知识正确性**

知识

训练阶段

LLMs

结合知识的有监督微调/指令微调
(SFT: Supervised Fine-tuning / IT: Instruct-tuning)

Question:

A 5-year-old boy has recently been experiencing symptoms such as frequent ear pain and decreased hearing for some time. After examination, he was diagnosed with acute otitis media.
What is the recommended treatment plan for the doctor?

Key entity:

Pediatric Acute Otitis Media

Attribute:

Treatment Plan

Stage 1: Parameter Filling

acquire_knowledge

(entity=Pediatric Acute Otitis Media, attribute=Treatment Plan)

```
def acquire_knowledge(entity, attribute):  
    .....  
    return knowledge
```

Stage 2: Knowledge Function Call

Question:

A 5-year-old boy has recently been experiencing symptoms such as frequent ear pain and decreased hearing for some time. After examination, he was diagnosed with acute otitis media.
What is the recommended treatment plan for the doctor?

Knowledge:

```
{ "Pediatric Acute Otitis Media": {  
    "Treatment Options": ["Systemic Treatment", "Antibiotic Treatment", "Local Treatment", "Observation"],  
    "Medication Treatment": ["Amoxicillin", "Cephalosporins", "Macrolides", "Analgesics"]} }
```

Output:

According to current treatment standards, medication is not necessary for the early stages of acute otitis media, and observation is the main approach. However, if symptoms do not improve after 48 hours, the doctor should consider using antibiotics for treatment. Amoxicillin and other antibiotics are usually recommended.

Relevant knowledge: "Pediatric Acute Otitis Media": { "Treatment Options": ["Systemic Treatment", "Antibiotic Treatment", "Local Treatment", "Observation"], "Medication Treatment": ["Amoxicillin", "Cephalosporins", "Macrolides", "Analgesics"] }

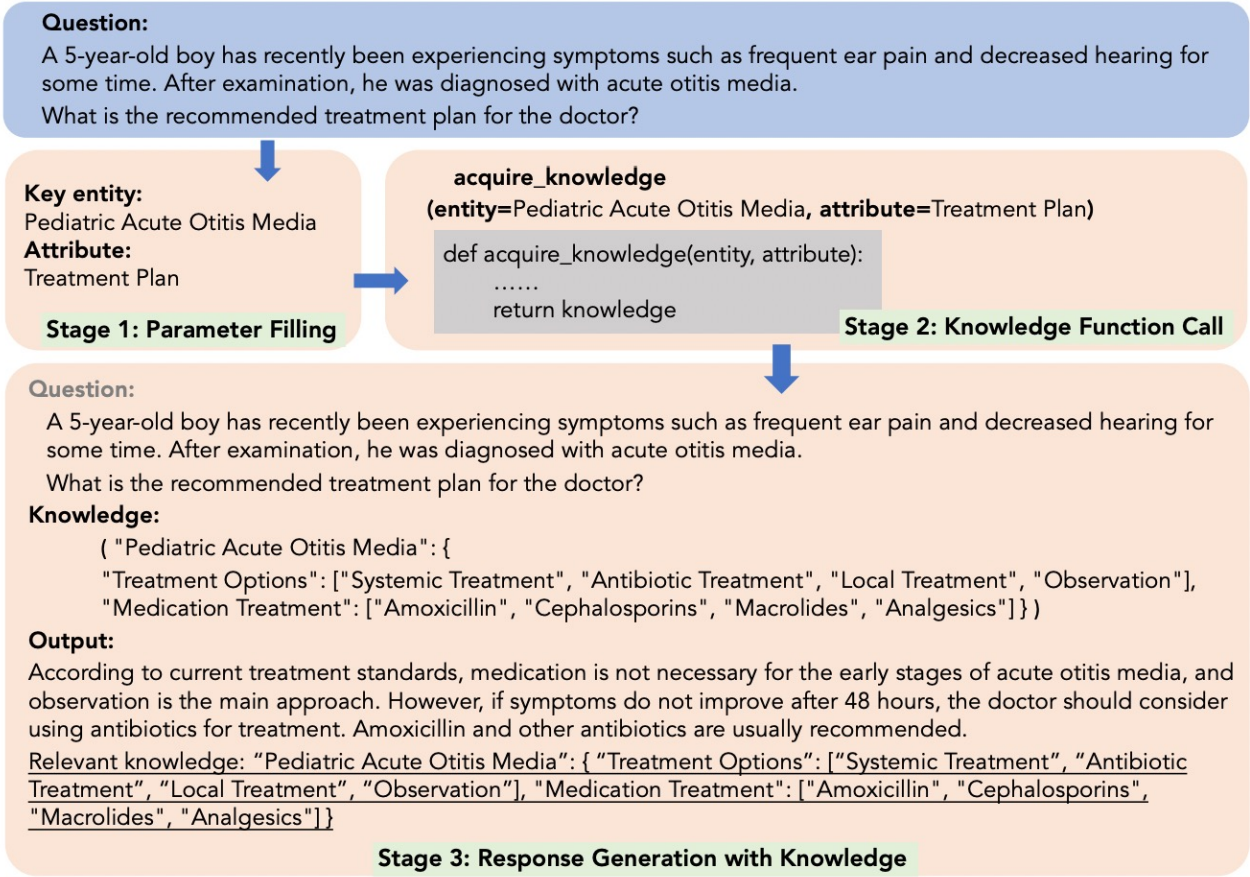
Stage 3: Response Generation with Knowledge

方法：推理阶段

- 结合医学知识的回复生成
 - 医学知识：知识图谱、医学文献、.....
 - 训练大语言模型利用医学知识库：根据输入内容查询相关知识
- 结合知识的生成可提升回复内容的知识正确性



结合知识的回复生成



实验效果

Base Model	H ₂		ChatGPT
	h ₁	h ₂	
Alpaca	1.78	1.98	2.68
Alpaca + <i>i-tuning</i>	1.80	1.92	2.64
Alpaca + <i>k-tuning</i>	2.25	2.40	2.69
Alpaca + <i>k-tuning</i> [†]	2.63	2.59	2.71
Bloom	1.82	1.83	2.61
Bloom + <i>i-tuning</i>	1.81	1.88	2.47
Bloom + <i>k-tuning</i>	2.41	2.42	2.74
Bloom + <i>k-tuning</i> [†]	2.72	2.57	2.79
ChatGPT	2.71	2.59	-

h1代表回复的有帮助程度， h2代表回复没有误导的程度

Medical Knowledge: Entity: cicatricial pyloric obstruction, Attribute: symptom, Knowledge content: “vomiting during afternoon and night, abdominal pain during the night and after eating.”
Question: When should we suspect a cicatricial pyloric obstruction based on the symptoms?
Golden response: When patients have upper abdominal pain, postprandial vomiting, early satiety, and abdominal distension, with or without weight loss, one should suspect cicatricial pyloric obstruction.
Bloom: When the following symptoms are present, suspect cicatricial pyloric obstruction: (1) Upper abdominal pain, (2) Nausea, (3) Vomiting, (4) Weight loss, (5) Belching, (6) <i>Black stools</i> , (7) <i>Jaundice</i> ,
Bloom with instruction-tuning: When the following symptoms are present, suspect scar-induced pyloric obstruction: abdominal pain, vomiting, <i>acid reflux</i> , <i>jaundice</i> , <i>heartburn</i> , <i>belching</i> , <i>nausea</i> , <i>loss of appetite</i> , <i>weight loss</i> ,
Bloom with knowledge-tuning: (1) Abdominal pain and bloating: cicatricial pyloric obstruction often occurs after eating, especially at night. The upper abdominal pain worsens after meals and progresses to diffuse upper abdominal distension or discomfort as gastric retention occurs. (2) Vomiting: The most prominent symptom, vomiting mainly occurs in the afternoon and evening.

样例展示

Knowledge-tuning Large Language Models with Structured Medical Knowledge Bases for Reliable Response Generation in Chinese

Haochun Wang, Sendong Zhao, Zewen Qiang, Zijian Li, Nuwa Xi, Yanrui Du, MuZhen Cai, Haoqiang Guo, Yuhan Chen, Haoming Xu, Bing Qin, Ting Liu

Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China {hcwang, sdzhao}@ir.hit.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable success in diverse natural language processing (NLP) tasks in general domains. However, LLMs sometimes generate responses with the *hallucination* about medical facts due to limited domain knowledge. Such shortcomings pose potential risks in the utilization of LLMs within medical contexts. To address this challenge, we propose knowledge-

tical question in English and Chinese respectively. When answering the question in English, ChatGPT provides reasonable medications for “hepatobiliary stones”. However, given the identical one in Chinese, ChatGPT recommends “Ri-fampicin”, which is an antibiotic medicine to treat mycobacterial infections and not effective for hepatobiliary stones. Such hallucinations in the responses generated by the large language models can lead to sub-optimal drug recommenda-

Wang et al. Knowledge-tuning. Arxiv 2023

The CALLA Dataset: Probing LLMs’ Interactive Knowledge Acquisition from Chinese Medical Literature

Yanrui Du¹, Sendong Zhao¹, Muzhen Cai¹, Jianyu Chen¹ Haochun Wang¹, Yuhan Chen¹, Haoqiang Guo¹, Bing Qin¹

¹Harbin Institute of Technology, Harbin, China { yrdu, sdzhao, yhchen, mzcai, jychen, hcwang, yhchen, hqguo, qinb }@ir.hit.edu.cn

Abstract

The application of Large Language Models (LLMs) to the medical domain has stimulated the interest of researchers. Recent studies have focused on constructing Instruction Fine-Tuning (IFT) data through medical knowledge graphs to enrich the interactive medical knowledge of LLMs. However, the medical literature serving as a rich source of medical knowledge remains unexplored. Our work introduces the CALLA dataset to probe LLMs’ interactive knowledge ac-

The golden fact

在鉴别腺瘤性息肉与胆固醇性息肉，相比于常规超声造影，高帧频超声造影会提供更准确的影像学依据。(Compared with conventional CEUS, high frame frequency CEUS can provide more accurate imaging evidence in differentiating adenomatous polyps from cholesterol polyps)

Ask a question consistent with the fact

人：在鉴别腺瘤性息肉与胆固醇性息肉，相比于常规超声造影，高帧频超声造影会提供更准确的影像学依据吗？(Does high frame frequency CEUS provide more accurate imaging evidence than conventional CEUS in

Du et al. CALLA. Arxiv 2023



哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY



谢谢!

E-mail: sdzhao@ir.hit.edu.cn