# 大语言模型微调基础

## 如何构建医学专科大语言模型

# 课程内容

01 基本概念介绍
理解大模型的工作原理

02 微调的两大分类
学习策略与参数调整

03 模型低秩和量化
降低训练成本的核心技术

04 落地案例
中山医院AI问答系统实战

05 大模型微调实操
从理论到实践

06 构建医学数据集
数据是模型的基石

07 常见QA
实战中的问题解答

01

# 基本概念介绍

理解大模型

# 什么是大模型？

微调

修改其中部分参数

## 定义

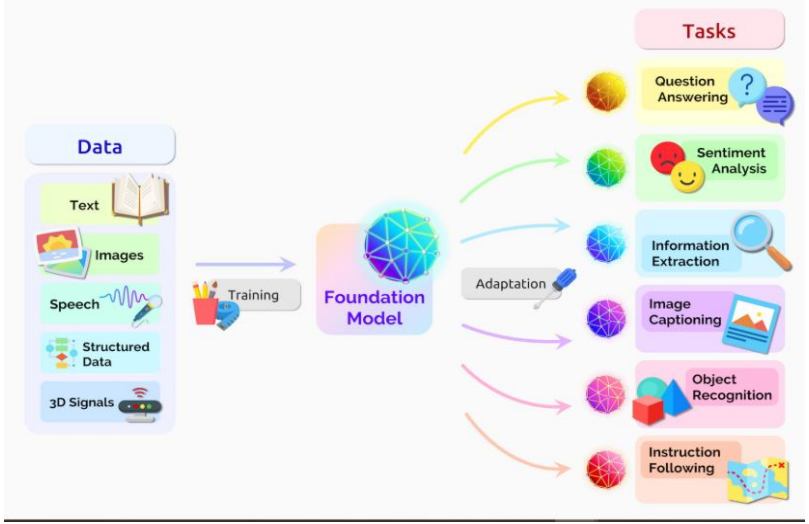
大模型（Foundation Model）是指参数量达到数十亿甚至数千亿规模的人工智能模型，通过海量数据训练，能够理解和生成人类语言、图像等多模态内容。

## 主要特点

- **大规模参数**：从70亿到千亿级别（如GPT-4、Claude）
- **海量训练数据**：整个互联网文本、书籍、代码等
- **强大的泛化能力**：一个模型可完成多种任务
- **涌现能力**：参数规模突破临界点后出现的新能力

## 医疗领域应用

- **临床决策支持**：辅助诊断、治疗方案推荐
- **病历智能处理**：自动生成病历摘要
- **医学知识问答**：为医护人员和患者提供专业解答
- **医学文献分析**：快速检索和总结最新研究成果



# AI大模型的整体架构

## 大模型分层架构（自下而上）

### 1 基础层（Infrastructure Layer）

- 硬件：GPU/TPU集群、高性能服务器
- 框架：PyTorch、TensorFlow、JAX

### 2 数据层（Data Layer）

- 通用数据：预训练阶段的海量文本
- 垂域数据：医学教材、临床指南、病例库
- 动态数据：实时更新的医学知识库

### 3 模型层（Model Layer）

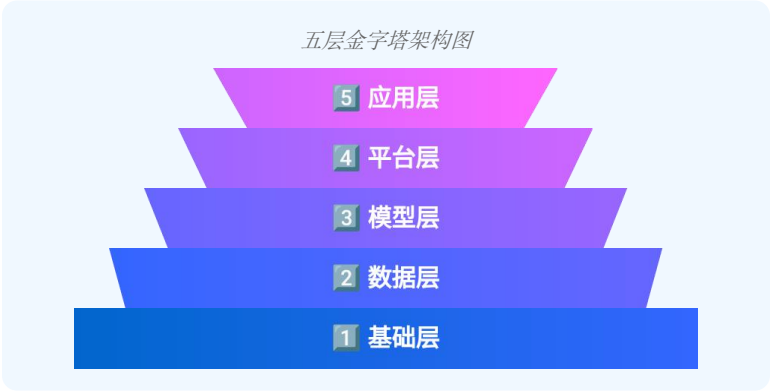
- 基座模型：Llama 3、Qwen、DeepSeek、GPT系列
- 微调模型：针对医疗场景定制的专科模型
- 特点：基于Transformer架构

### 4 平台层（Platform Layer）

- 开发工具：LangChain、LlamaIndex
- 评测体系：医学问答准确率、安全性评估
- 部署框架：FastAPI、vLLM、TGI

### 5 应用层（Application Layer）

- 终端产品：AI问诊助手、病历生成工具
- 交互方式：对话界面、语音输入、多模态交互



# 大模型如何"听懂"我们的话?

## 核心原理：词嵌入（Word Embedding）

### 1 步骤1：文字转数字

大模型不能直接理解文字，需要将每个词转换为一组数字（向量）。这个过程叫做"词嵌入"。

### 2 步骤2：向量表示示例

假设用简化的2维向量表示医学词汇：

- "男护士" → [2.0, 1.0]
- "女护士" → [2.0, 2.0]
- "男性" → [1.0, 1.0]
- "女性" → [1.0, 2.0]

### 3 步骤3：向量运算

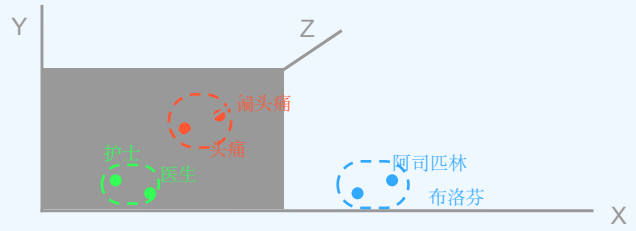
进行简单的数学运算：

"男护士" - "男性" + "女性"
= [2.0, 1.0] - [1.0, 1.0] + [1.0, 2.0]
= [2.0, 2.0]
= "女护士"

## 实际应用

在实际的医疗大模型中，每个词会被表示为768维甚至更高维度的向量，能够捕捉词汇的语义、语法、医学专业属性等多重信息。

*词嵌入可视化 - 3D空间中的词向量分布*




## 为什么需要高维向量?

- **语义相似度**：相似的词在向量空间中距离更近
- **关系编码**：能够表达"症状-疾病"、"药物-适应症"等医学关系
- **上下文敏感**：同一个词在不同场景下向量会动态调整

# 上下文理解 - "降压"的多重含义

## 医学术语的多义性挑战

医学术语经常具有多重含义，其确切含义高度依赖上下文。大模型通过**注意力机制（Attention）**解决这个问题。

### 场景1：心血管内科

**输入文本：**

*"患者血压180/110mmHg，需要立即进行降压治疗"*

**模型理解：**

周围词："患者"、"血压"、"治疗" → 指向医疗场景

"降压"的向量被动态调整为"降低血压"的语义

向量变化：[1.2, -0.2, 1.4, ...] → [2.5, 0.3, 3.1, ...]

### 场景2：神经外科

**输入文本：**

*"颅内压增高，立即给予降压药物甘露醇"*

**模型理解：**

周围词："颅内压"、"甘露醇" → 指向神经系统

"降压"被理解为"降低颅内压力"

向量变化：[1.2, -0.2, 1.4, ...] → [0.8, 1.9, 2.7, ...]

### 场景3：医疗设备科

**输入文本：**

*"准备降压阀门，调整氧气供应压力"*

**模型理解：**

周围词："阀门"、"氧气"、"压力" → 指向设备场景

"降压"被理解为"降低气体压力"

向量变化：[1.2, -0.2, 1.4, ...] → [0.7, -1.5, 4.2, ...]

## 技术原理简述

模型使用**自注意力机制**，让句子中的每个词都能"看到"其他词，根据上下文动态调整每个词的向量表示，从而实现准确的语义理解。

# 为什么需要微调？

## 三大核心原因

*通用模型 vs 专科模型对比图*


### 1. 领域专业化能力提升

**问题：通用知识与专业需求的差距**

通用大模型知识广泛但缺乏专业深度

例如询问："冠心病患者的SYNTAX评分如何影响治疗方案选择？"

通用模型：给出模糊或教科书式的回答

微调后模型：结合最新临床指南给出精准建议

**解决方案：微调强化专业理解**

使用专业医学数据训练，强化对医学术语的理解

模型从"万事通"升级为"心血管专科医生"

准确率提升：从60%提升到85%以上

### 2. 任务特化与风格控制

**问题：通用模型的表达不符合医学规范**

输出冗长，不符合病历书写规范

语气不当，缺乏医学专业性

无法按照SOAP格式输出

**解决方案：微调精准控制输出特性**

训练模型按照医院规范生成病历

控制语气：专业、严谨、人文关怀

格式化输出：结构化的诊断报告

### 3. 部署与资源效率优化

**问题：大模型直接部署成本高、数据安全风险**

调用API成本高：每次请求0.01-0.1元，日均千次调用年成本数万元

数据外传风险：患者隐私、医院敏感数据

响应延迟：网络请求增加延迟

**解决方案：本地微调模型**

一次性训练成本：2000-5000元（电费+算力）

本地部署：数据不出医院，符合合规要求

推理速度快：本地GPU推理延迟降低80%

## 成本对比示例

| 方案 | 初期成本 | 年度成本 | 数据安全 | 响应速度 |
|---|---|---|---|---|
| **调用API** | ¥0 | ¥30,000+ | 数据外传 | 200-500ms |
| **本地微调** | ¥3,000 | ¥500（电费） | 本地部署 | 50-100ms |

# 微调的典型应用场景

## 场景1：定制模型风格和语气

### 案例：医患沟通助手

**训练目标：**

让AI以温和、专业、通俗易懂的语气与患者交流

**效果对比：**

微调前："根据临床数据分析，您的血脂代谢异常"

微调后："您的血脂有点高，需要注意饮食和运动，我们一起来控制它"

## 场景2：提升专业回答准确性

### 案例：冠心病诊疗助手

**训练目标：**

根据症状和检查结果给出符合指南的诊疗建议

**训练数据：**

《中国心血管病报告》、临床路径、真实病例

**效果：**

诊断准确率从65%提升至90%

## 场景3：理解复杂的医学指令

### 案例：智能病历系统

**训练目标：**

理解医生口述的复杂病历内容

**输入示例：**

"55岁男性，主诉胸痛3小时，既往高血压糖尿病史，心电图ST段抬高，肌钙蛋白升高"

**输出：**

结构化的电子病历，包含主诉、现病史、既往史、辅助检查、初步诊断

## 场景4：学习新的医学技能

### 案例：用药指导助手

**训练目标：**

学会根据患者情况推荐用药方案

**训练数据：**

药品说明书、用药指南、药物相互作用数据库

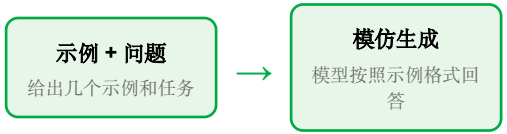
**效果：**

能够识别配伍禁忌，给出个性化用药建议

# In-context、RAG、微调的区别（一）

三种方法原理对比

## In-context Learning




**类比**

给新来的住院医生几份病历样本，让他照着写

**优点**

- 简单快速，无需训练
- 灵活，可以随时调整示例

**缺点**

- 受上下文长度限制（4k-128k tokens）
- 成本随输入长度增加

## RAG（检索增强生成）


**类比**

开卷考试，模型可以查阅知识库

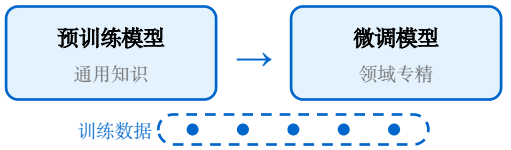
**优点**

- 知识可随时更新，适合动态信息
- 能引用最新资料和外部知识

**缺点**

- 依赖检索质量，知识库构建成本高
- 检索-生成两步流程可能增加延迟

## 微调（Fine-tuning）




**类比**

让医生参加专科培训，真正学会新技能

**优点**

- 深度定制，性能最优
- 推理速度快，可本地部署

**缺点**

- 需要标注数据和算力资源
- 知识更新需要重新训练

# In-context、RAG、微调的区别（二）

详细对比与选择建议

| 对比维度 | In-context | RAG | 微调 |
|---|---|---|---|
| **实施难度** | 最简单 | 中等 | 较复杂 |
| **成本** | 按输入长度付费 | 知识库维护成本 | 一次性训练成本 |
| **效果** | 一般 | 良好 | 优秀 |
| **知识更新** | 手动更新示例 | **实时更新知识库** | 需要重新训练 |
| **适用场景** | 简单任务、快速验证 | 问答系统、知识检索 | 专业领域、生产环境 |
| **数据需求** | **几个示例即可** | 结构化知识库 | 100-10000条标注数据 |
| **医疗场景举例** | 快速测试问答格式 | 药品说明书查询 | 疾病诊断助手 |

### 组合使用策略

微调 + RAG

- 用微调模型理解医学语言，用RAG补充最新知识

*（中山医院案例采用此方案）*

### 1 快速验证阶段

- 验证想法是否可行
- 准备少量示例测试效果

选择 In-context Learning

### 2 知识密集型任务

- 需要引用最新医学文献
- 涉及大量事实性知识
- 知识频繁更新（如药品信息）

选择 RAG

### 3 生产环境部署

- 对准确率要求高
- 需要深度定制
- 有足够的标注数据

选择 微调

# 微调的基本流程

## 完整微调流程（7步）

**步骤1** 选择基座模型

开源模型：Llama 3（70B/8B）
Qwen 3（72B/7B）
DeepSeek V3

选择标准：中文能力、参数规模
**医疗推荐：DeepSeek蒸馏过的Llama**

**步骤2** 准备训练数据

数据来源：病历、诊疗指南
数据格式：Alpaca或ShareGPT

**数据量：100-5000条**

数据清洗：去重、纠错、标准化

**步骤3** 选择微调方法

全参数微调：效果最好，成本最高

**LoRA：性价比最优，推荐**
QLoRA：显存受限时选择

**步骤4** 设置超参数

**学习率：1e-5到5e-4**

**训练轮数：2-5轮**

批次大小：8-32
LoRA秩：8-64

**步骤5** 开始训练

监控训练损失曲线

**使用验证集防止过拟合**
训练时长：6-48小时

**步骤6** 评估测试

定量评估：准确率、F1分数
定性评估：人工检查

**对比测试：前后效果**

**步骤7** 模型部署

合并LoRA权重到基座

量化压缩：INT8/INT4
部署框架：vLLM、FastAPI

### 关键成功要素

**高质量数据** > 大量数据

**充分的测试** > 盲目训练

**迭代优化** > 一次到位

# 超参数- 学习率

**📖 学习率类比：复习方法的调整幅度**

就像发现错题时调整解题方法的幅度大小：

- 调整太小（低学习率）→ 虽然稳定但进步缓慢，就像每次只微调解题步骤
- 调整适中（理想学习率）→ 快速掌握正确解法，就像找到最优解题思路
- 调整太大（高学习率）→ 容易矫枉过正，就像推翻整个解题思路导致混乱


调整方法的幅度 0.001

**调整幅度过小**

像小心翼翼调整解题步骤，进步缓慢但稳定

# 超参数- 训练轮数


课本习题得分
模拟考试得分
知识保留度
60
45
30
15
0
得分 (%)
100
75
50
25
0
知识保留度 (%)
1
复习次数
复习轮数调节
1
课本掌握
应试能力
知识留存

# 超参数-- 批次大小


第一批题目
第二批题目
新批次开始
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32
33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48
49 50 51 52 53 54 55 56
59 60 61 62 63 64
当前设置：每次复习 1 道题
批量大小调节
1
处理速度
专注程度
内存占用
批量过小
虽然能专注每个细节，但学习效率较低

02

# 微调的两大分类

*学习策略 VS 参数调整*

1

**学习策略**

强化**人**的作用

VS

2

**参数调整**

强化**技术**的作用

# 学习策略的四大方法

## 1. 监督式微调 (SFT)

**原理**：通过模仿人类示例进行学习
**数据需求**：问答对数据（Input-Output pairs）
**训练难度**：低
**应用效果**：良好
**医疗场景**：训练AI回答常见医学问题

**举例**：
输入："什么是高血压？"
输出："高血压是指血压持续高于140/90mmHg的状态..."

## 2. 人类反馈强化学习 (RLHF)

**原理**：通过人工评价指导模型优化
**数据需求**：偏好数据（哪个回答更好）
**训练难度**：高
**应用效果**：优秀
**医疗场景**：优化AI回答的专业性和患者友好度

**举例**：人工标注哪个回答更专业、更易懂、更有同理心

## 3. 知识蒸馏 (Knowledge Distillation)

**原理**：小模型学习大模型的知识
**数据需求**：大模型的输出作为训练数据
**训练难度**：中等
**应用效果**：良好
**医疗场景**：将70B大模型的能力压缩到7B小模型

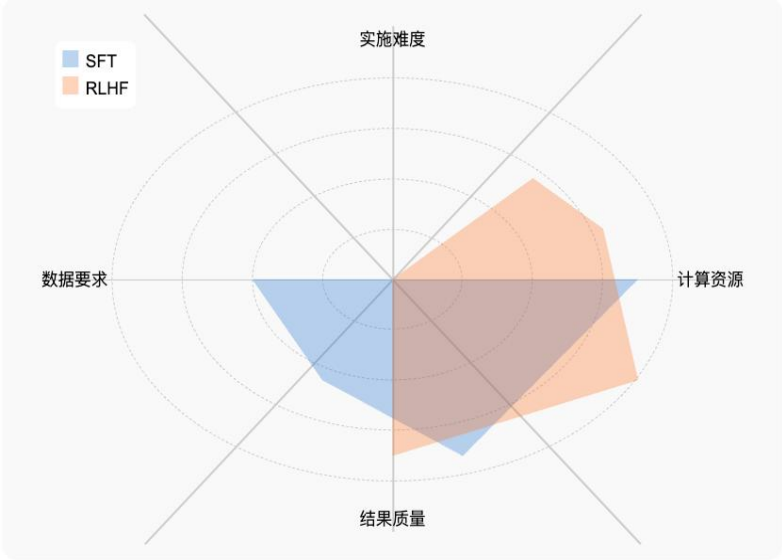
**优势**：小模型推理速度快，适合移动端部署

## 4. 奖励模型 (Reward Modeling)

**原理**：训练一个专门的评分模型
**数据需求**：排序数据（回答质量排名）
**训练难度**：中等
**应用效果**：优秀（辅助作用）
**医疗场景**：自动评估AI回答的医学准确性

**作用**：辅助RLHF，减少人工标注成本

# 学习策略的方法选择建议




## 方法选择建议

| 场景 | 推荐方法 | 理由 |
| --- | --- | --- |
| 快速上线MVP | SFT | 数据易获取，训练简单 |
| 追求最佳效果 | SFT + RLHF | 两阶段训练，效果最优 |
| 移动端部署 | 知识蒸馏 | 小模型，推理快 |
| 大规模应用 | SFT + 奖励模型 | 自动化评估，降低人工成本 |

医疗场景实用提示

- ✓ **起步阶段**：先使用SFT构建基础医疗问答能力
- ✓ **迭代提升**：收集真实用户反馈，应用RLHF优化回答质量
- ✓ **持续评估**：结合医学专家评估与奖励模型自动打分

# 参数调整的三种模式

## 全参数微调 (FFT)

**原理：** 调整模型的所有参数

**类比：** 给汽车全面改装，更换所有零件

**显存需求：** 70B模型需要140GB显存

**训练时间：** 3-7天

**训练成本：** ¥2,000+（8×A100算力）

**效果：** 最优

**适用场景：** 科研实验、预算充足的企业

## 低秩适配 (LoRA)

性价比最优

**原理：** 冻结原模型，仅训练少量新增参数

**类比：** 给汽车加装配件，不改动原车

**显存需求：** 相比FFT减少90%以上

**参数量：** 仅训练1-2%的参数

**训练时间：** 6-12小时

**训练成本：** ¥20-40（单卡A6000/4090）

**效果：** 优秀（接近FFT 95%效果）

**适用场景：** 大多数实际应用，性价比最优

## 表示微调 (ReFT)

**原理：** 不改参数，优化模型内部的表示层

**类比：** 给汽车刷程序，不动硬件

**显存需求：** 极低，笔记本电脑即可

**参数量：** 几乎为0

**训练时间：** 1-2小时

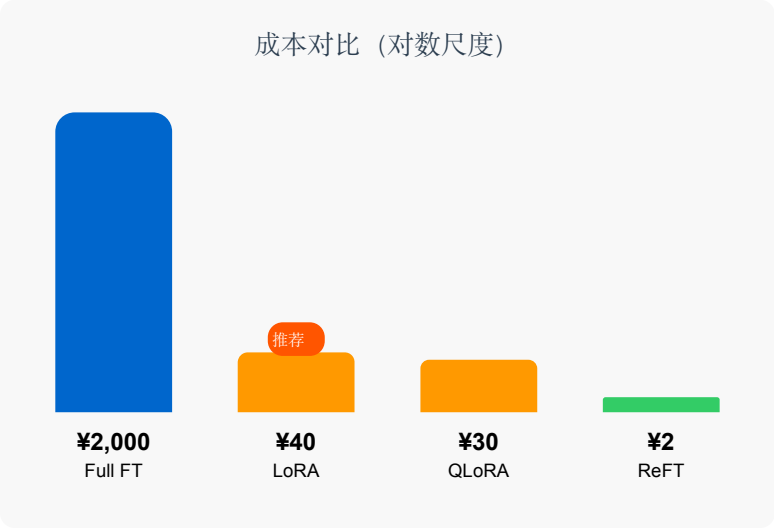
**训练成本：** ¥2（消费级显卡）

**效果：** 良好

**适用场景：** 资源极度受限、快速验证

训练成本对比 (对数刻度)

# 参数调整方法的成本与效果对比




| 方法 | 硬件配置 | 训练时间 | 电费成本 | 显存需求 | 效果 |
|---|---|---|---|---|---|
| Full FT | 8×A100(80GB) | 3天 | ¥2,000 | 140GB+ | 100% |
| **LoRA** | 1×A6000(48GB) | 12小时 | ¥40 | 24GB | 95% |
| QLoRA | 1×RTX4090(24GB) | 18小时 | ¥30 | 12GB | 92% |
| ReFT | 笔记本RTX3060 | 1小时 | ¥2 | 6GB | 85% |

## 推荐选择

**预算充足、追求极致**
→ Full Fine-Tuning

推荐

**平衡效果与成本**
→ LoRA / QLoRA

**快速验证、资源受限**
→ ReFT

**医疗AI应用建议**

对于医疗场景，推荐使用LoRA方法，既能保持基础医学知识，又能有效学习科室专用术语和临床路径。医疗数据往往结构化程度高，适合低秩表示的特性。

# 时间

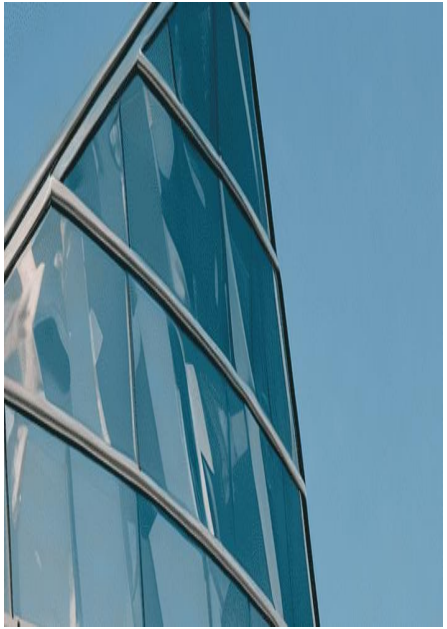
01 2019年：Adapter微调，参数效率提升10倍。

02 2021年：LoRA提出，训练成本降低95%。

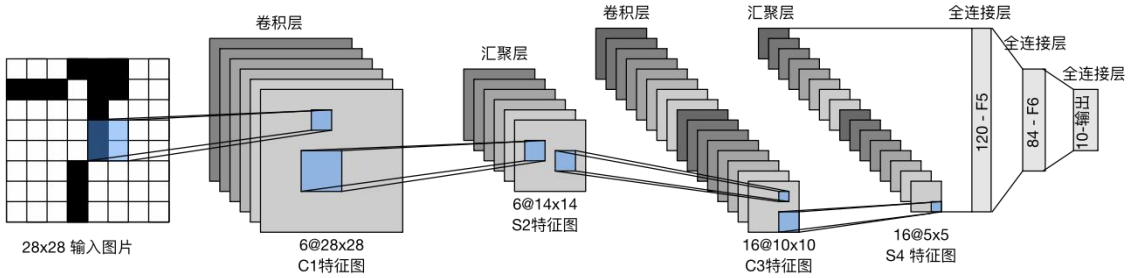
03 2023年：QLoRA实现单卡训练650亿参数模型。

04 2024年：ReFT + unsloth，训练速度提升3倍。

05 2025年： Unsloth 8.8倍加速

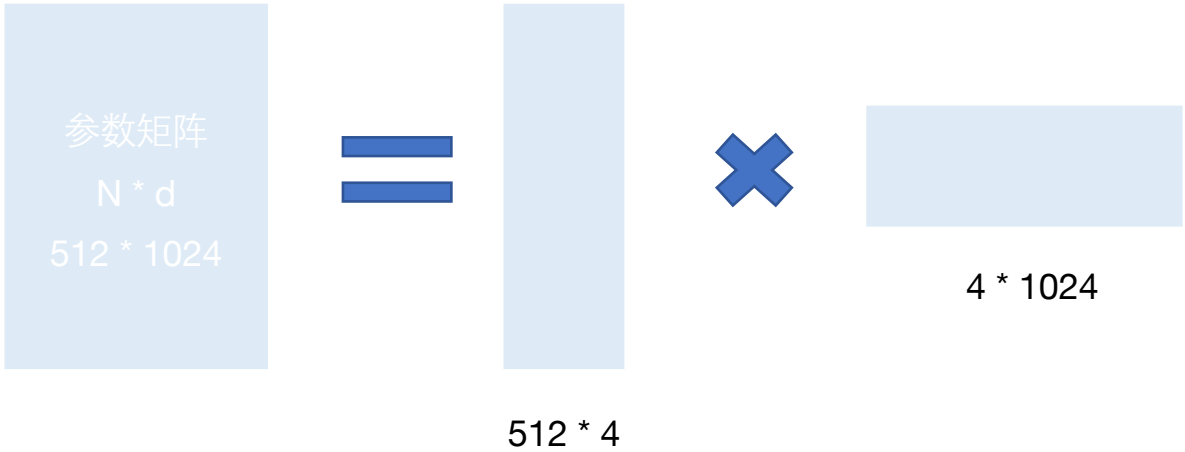


# 低秩的概念



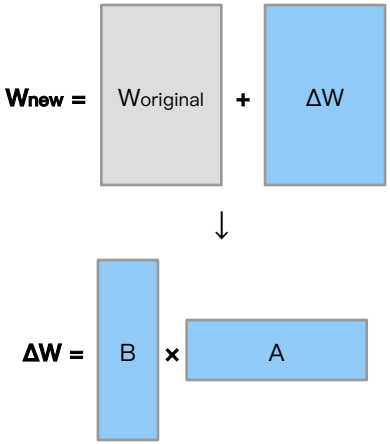

低秩矩阵分解：Low Rank Adaptation

核心思想：将大矩阵分解成两个小矩阵相乘



# LoRA原理简介

## 核心思想：低秩矩阵分解




*B (Nxr) × A (rxd)，其中 r << min(N,d)*

**数学表达：**

$W_{new} = W_{original} + \Delta W$

$\Delta W = B \times A$ (低秩分解)

其中：

- $W_{original}$：原模型参数（冻结不训练）
- B：矩阵B（维度：N × r）
- A：矩阵A（维度：r × d）
- r：秩（通常 r << d）

## 参数量对比

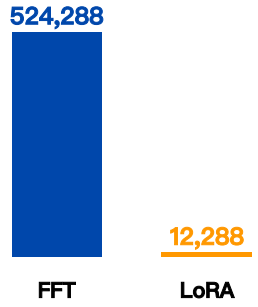
原始全参数微调：

参数矩阵：512 × 1024 = 524,288 个参数

LoRA微调：

矩阵B：512 × 8 = 4,096 个参数
矩阵A：8 × 1024 = 8,192 个参数
总计：4,096 + 8,192 = 12,288 个参数

**减少比例：12,288 / 524,288 ≈ 2.3%**




## 关键设计

### 1. B矩阵全零初始化

- 目的：确保训练初期模型输出与原模型一致
- 效果：避免训练初期模型坍塌

### 2. A矩阵随机初始化

- 目的：引入可学习的变化
- 方法：使用高斯分布初始化

### 3. 推理时合并参数

- 训练完成后：将 BxA 直接加到原参数上
- 优势：推理时无额外计算开销

## 为什么有效？

**低秩假设：**模型的参数更新通常只需要调整少数几个主要方向（低秩空间），不需要调整所有维度。

通用模型已经学会了语言理解（主要能力保留）
微调只需要学会医学术语和诊疗逻辑（少量专业知识）

不需要重新学习语言的基础规律

# 量化技术基础

## 什么是模型量化?

**定义**：将模型参数从高精度（如32位浮点数）转换为低精度（如8位整数），减小模型体积和显存占用。

**类比**：将高清照片压缩为较低分辨率，文件变小但主要内容仍可识别。

## 量化精度直观对比


*方块大小代表存储空间需求*

## 常见量化精度

| 精度类型 | 存储空间 | 准确率保留 | 典型应用 |
|---|---|---|---|
| **FP32** | 4 bytes | 100% | 训练基座模型 |
| **FP16** | 2 bytes | 99-100% | 微调训练 |
| **INT8** | 1 byte | 95-98% | 生产部署 |
| **INT4** | 0.5 bytes | 90-95% | 消费级设备 |

## 医疗应用中的精度选择

**模型训练：** 使用FP16/BF16保持精度
**临床部署：** 使用INT8平衡性能与准确性
**移动终端：** 使用INT4适应硬件限制
**资源受限环境：** 考虑QLoRA（下页详述）

# 量化应用与QLoRA技术

## 显存计算公式

**基础公式：**

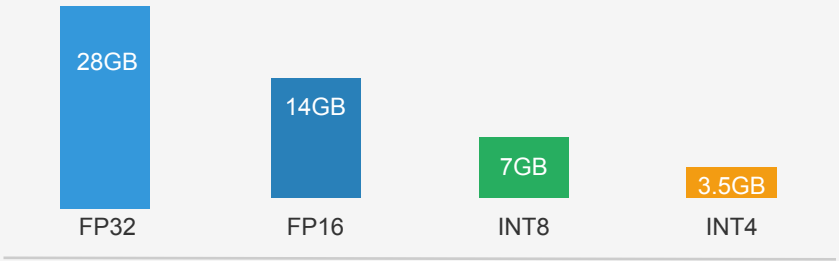
显存占用(GB) ≈ 参数量(B) × 精度(bytes)

**示例：Qwen 2.5 - 7B 模型**

FP32：7B × 4 bytes = 28GB
FP16：7B × 2 bytes = 14GB
INT8：7B × 1 byte = 7GB
INT4：7B × 0.5 bytes = 3.5GB

*注：实际显存 = 参数显存 × 1.5 ~ 2.0（考虑KV缓存和中间激活）*

## 7B模型显存需求对比




## QLoRA = 量化 + LoRA

**原理：先将模型量化到4-bit，再进行LoRA微调**

**核心优势：**

**显存需求降低75%**（相比FP16 LoRA）
24GB显卡可微调70B模型
效果仅损失2-3%

**训练流程：**

加载4-bit量化的基座模型（冻结）
添加FP16精度的LoRA适配器
前向传播时将4-bit参数反量化到FP16
反向传播仅更新LoRA参数

## 量化精度选择建议

| 场景 | 推荐量化精度 | 理由 |
| --- | --- | --- |
| 训练阶段 | FP16 / BF16 | 保持精度 |
| 部署阶段（服务器） | INT8 | 平衡速度与精度 |
| 部署阶段（移动端） | INT4 | 适应硬件限制 |
| 微调训练（显存受限） | (QLoRA) | 最大化可用显存 |

# 技术演进时间线


## 大模型微调技术发展历程

**2019年　Adapter诞生**

提出插入适配器层的思想

**2021年　LoRA横空出世**

微软发布LoRA论文
训练成本降低95%
成为PEFT（参数高效微调）的主流方法

**2023年　QLoRA突破**

单张24GB显卡可微调65B模型
4-bit量化 + LoRA的完美结合

**2024年　效率革命**

ReFT提出表示微调新范式
Unsloth库优化训练速度（提升2-5倍）
工具链成熟（Axolotl、LLaMA-Factory等）

## 未来趋势（2025-2027预测）

**技术方向：**

**自动化微调**：AutoML自动选择最优超参数
**持续学习**：模型在部署后持续从新数据学习
**多模态融合**：医学影像 + 文本 + 基因数据
**边缘端微调**：在手机、IoT设备上直接微调

**医疗应用方向：**

**个性化医疗AI**：为每个患者微调专属模型
**多中心联邦微调**：多家医院数据不出本地
**实时知识更新**：自动从最新文献更新模型

# 03

# 落地案例分享

## 中山医院AI健康问答系统

真实项目　微调+RAG　已上线运行

# 项目背景与需求

## 项目基本信息

**项目名称：**
中山医院AI健康问答小程序

**合作单位：**
复旦大学附属中山医院心内科

**项目周期：**
2025年3月 – 2025年12月

**项目状态：**
已上线运行，服务患者3000+人次

### 成果产出：

申请软件著作权1项
发表论文1篇（投稿中）
项目结题报告已提交

## 业务痛点

### 1 患者健康咨询需求大，医生时间有限

心血管疾病患者需要长期管理
患者经常有用药、饮食、运动等咨询需求
医生门诊时间有限，无法详细解答所有问题

### 2 健康知识获取渠道不可靠

患者自行搜索网络信息，真假难辨
社交媒体上的健康建议缺乏医学依据
需要一个专业、可信的健康知识来源

### 3 患者自我管理能力不足

缺乏对自身病情的系统了解
不清楚如何正确用药和监测指标
缺少专业指导容易焦虑

## 项目目标

### 1 构建冠心病垂域智能问答系统

基于权威医学指南训练专科模型
准确率达到85%以上
回答通俗易懂，患者可理解

### 2 整合健康知识库

涵盖冠心病预防、诊断、治疗、康复全流程
知识来源：临床指南、权威医学教材、医院SOP
定期更新，保持知识时效性

### 3 辅助患者健康管理

用药提醒、指标记录
健康评估、风险预警
个性化健康建议

**2025年3–4月**
需求分析与规划

**2025年5–7月**
知识库构建与模型训练

**2025年8–9月**
系统开发与测试

**2025年10–12月**
上线运行与效果评估

# 系统架构设计

**UI 前端层：微信小程序**

AI健康问答、健康知识库、健康工具、个人管理
简洁易用，适配老年人

**AI 后端层：AI推理服务**

微调模型：Qwen 2.5-7B 冠心病专科版
RAG知识检索 + 多轮对话 + 安全审核

**DB 数据层：知识库与数据库**

结构化数据：用户信息、对话历史
向量数据库：心血管疾病权威医学知识

## 技术栈总览

| 层级 | 技术选型 | 说明 |
|---|---|---|
| **前端** | uni-app + Vue3 | 跨平台小程序开发 |
| **后端** | FastAPI + Python | 高性能异步服务 |
| **AI模型** | Qwen 2.5-7B + LoRA | 轻量化微调模型 |
| **向量检索** | Milvus + BGE-M3 | 混合医学知识检索 |

## 护理价值

- 减轻重复问题工作量
- 提供标准健康指导
- 辅助患者健康管理

## 患者价值

- 随时获取可靠健康信息
- 个性化健康建议
- 简单记录健康指标

## 典型应用场景

**用药咨询**
咨询用药注意事项和相互作用

**饮食指导**
冠心病饮食禁忌与建议

**指标解读**
血压血脂等检测数据解释

## 系统优势

**信息准确**
基于权威医学资料

**易于使用**
简洁界面设计

**安全可靠**
隐私数据保护

**持续更新**
定期更新医学知识

# 核心技术：微调+RAG混合方案

## 单一方案的局限性

**单纯微调的局限**

- **知识更新慢**：需要重新训练才能更新知识
- **可解释性差**：无法给出知识来源
- **幻觉问题**：可能编造不存在的医学知识

**单纯RAG的局限**

- **检索依赖性强**：检索失败则回答质量下降
- **医学理解不足**：通用模型难以理解专业术语
- **回答生硬**：简单拼接检索结果，缺乏人性化

**微调+RAG混合方案优势**

- **微调**：强化医学语言理解，掌握冠心病领域知识
- **RAG**：提供最新知识来源，增强可信度
- **结合**：微调模型负责理解和生成，RAG提供事实依据

## 混合方案工作流程

1. 用户提问
   > *"阿司匹林和氯吡格雷可以一起吃吗？"*
2. 问题理解与改写
   理解这是关于抗血小板药物联用的问题
3. 知识检索
   从知识库中检索相关文档并重排序
4. 答案生成
   结合模型知识和检索结果生成专业答案
5. 安全审核
   检查回答是否包含禁忌内容，添加免责声明

### 关键技术细节

**微调技术**

- 基座模型：Qwen 2.5-7B
- 微调数据：2000条问答对
- 微调方法：QLoRA

**知识库技术**

- 文档：1200篇医学资料
- 向量模型：BGE-M3
- 检索：语义+关键词混合

### 示例回答

"阿司匹林和氯吡格雷是临床常用的抗血小板药物组合，在某些情况下确实可以联合使用：

**可以联用的情况**：急性冠脉综合征患者、冠脉支架植入术后（通常联用时间为12个月）

**注意事项**：需医生评估、出血风险增加需定期复查、消化道疾病史需谨慎

**温馨提示**：请务必在医生指导下用药，不要自行调整。

*参考来源：《抗血小板药物临床应用指南》*

## 评估指标

| 指标 | 微调前 | 微调后 | 提升 |
|---|---|---|---|
| 医学术语理解准确率 | 65% | 92% | +27% |
| 回答相关性 | 70% | 89% | +19% |
| 患者友好度评分 | 3.2/5 | 4.6/5 | +44% |
| 事实准确性 | 75% | 90% | +15% |

*评估方法：由多名冠心病专科医生和3名护理专家共同评估200个随机测试问题的回答质量。*

## 微调实战经验总结

**数据质量比数据量更重要：**2000条高质量问答胜过10000条低质量问答

**医生参与全流程必不可少：**确保专业性和临床实用价值

**LoRA微调技术高效实用：**小参数量实现大提升

**专科化微调价值显著：**针对特定科室（如冠心病）微调效果最佳

## 微调效果对比

**测试问题："支架术后多久可以运动？"**

### 微调前（原始Qwen）

"冠脉支架植入术后的运动恢复需要循序渐进。一般来说，术后1-2周可以进行轻度活动，如散步。3-4周后可以进行中等强度运动，如快走、慢跑。具体需要根据个人恢复情况，建议咨询医生。"

*评价：回答正确但较为模板化，缺乏细节*

### 微调后（冠心病专科模型）

"支架术后的运动恢复可以分为三个阶段：

**第一阶段（术后1-2周）：**
- 在病房或家中慢走，每次5-10分钟，每天2-3次
- 避免用力、憋气动作

**第二阶段（术后2-4周）：**
- 逐步增加活动量，可户外散步15-30分钟
- 可做轻度家务，如整理衣物

**第三阶段（术后1个月以上）：**
- 如复查正常，可进行快走、游泳、太极拳等
- 运动时心率不超过(220-年龄)×70%
- 避免剧烈对抗性运动

**重要提示：**
- 具体方案需根据复查结果调整
- 运动中出现胸闷、气短应立即停止并就医
- 建议参加医院的心脏康复计划

参考：中山医院心内科康复指南"

*评价：分阶段指导详细，给出具体心率计算公式，强调个体化，专业性强*

# 项目成果与影响

## 运营数据（截至2025年12月）

**2200+**
注册用户

**400+**
月活跃用户

**12,000+**
累计对话轮次

**4.2轮**
平均对话轮数

**82%**
问题解决率

## 科研成果

**软件著作权 (2025年9月)**

《冠心病智能问答与知识推荐微信小程序》

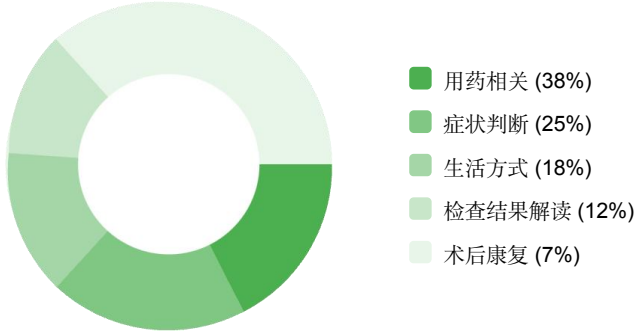
《慢性病患者健康管理辅助微信小程序》

**学术论文 (投稿中)**

《**》

**项目结题 (2025年12月)**

## 高频咨询问题TOP 5




## 用户真实反馈

**A** **患者A（62岁，冠心病术后）：**

*"这个小程序太方便了！我经常想问医生一些小问题，但又怕麻烦医生。现在随时都能问，回答也很专业，我很放心。"*

**C** **医生C（心内科主治医师）：**

*"这个系统大大减轻了我们的咨询工作量。很多常见问题患者可以先通过AI了解，来门诊时我们可以集中处理更复杂的问题，提高了效率。"*

# 小程序开发 - 中山小程序智护暖心开发

## 03 中山小程序智护暖心开发

**服务对象:**

上海复旦中山附属医院

**功能模块:**

| | | |
|---|---|---|
| 健康知识库 | AI健康问答 | 问话功能 |
| 健康管理 | 用药提醒 | 快捷功能 |

**特色功能:**

- 推荐阅读（心脏康复相关文章）
- 全部文章分类管理
- 智能对话功能

健康知识库首页


全部文章页面


AI问答页面


对话详情页面


健康管理页面


小程序图标

04

# 大模型微调实操

从准备到微调的完整流程

```
$ python finetune.py \\
  --model_name_or_path Qwen/Qwen2.5-7B-Instruct \\

  --data_path ./data/medical_qa.json \\
  --output_dir ./outputs \\
  --lora_r 16
Starting training...
```

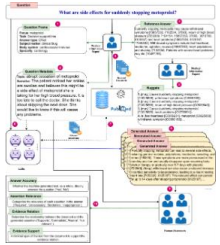
```
def train_lora_model(model, tokenizer, dataset):

# Configure LoRA for efficient fine-tuning

  peft_config = LoraConfig(
    r=16, lora_alpha=32, target_modules=[...],

    lora_dropout=0.05, bias="none")
```

# unsloth + 医疗数据集 + LoRA实战

GPU: T100-40GB　RAM: 80GB　Python 3.10　unsloth 0.3.0

## 医疗问答微调案例

通过构建医疗问答数据集，训练AI医疗回答能力。使用unsloth加速QLoRA微调，提升训练效率。



*医疗问答数据集结构示例*

## 微调步骤

准备医学论文领域问答对数据集，训练奖励模型。使用unsloth简化QLoRA训练过程，提升速度。

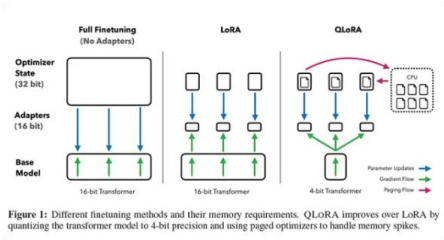



Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

*QLoRA vs. LoRA vs. 全参数微调对比*

## 效果对比

微调前：AI对医疗问题回答模糊

微调后：AI对医疗问题回答准确

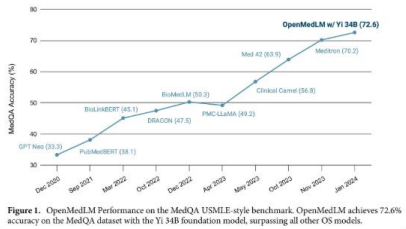
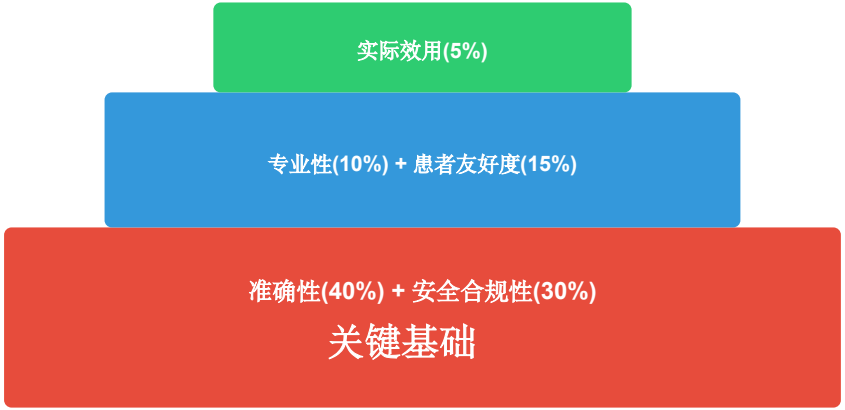



Figure 1. OpenMedLM Performance on the MedQA USMLE-style benchmark. OpenMedLM achieves 72.6% accuracy on the MedQA dataset with the Yi 34B foundation model, surpassing all other OS models.

*医疗模型在MedQA基准测试的性能提升*

# 医疗模型评估框架 (1/2)




### 上线标准

综合得分 ≥ 85分
专业准确性 ≥ 90分
**安全合规性 = 100分（零容忍）**

## 准确性（40%）

**医学事实准确性**
- 疾病定义与诊断标准
- 用药剂量与禁忌症

**诊疗建议合理性**
- 符合临床指南
- 考虑患者个体差异

## 安全合规性（30%）

**边界识别能力**
- 识别超出AI能力范围的问题
- 紧急情况正确引导就医

**免责声明**
- 提示"建议咨询医生"
- 避免给出诊断结论

**示例**：用户："胸痛得厉害，怎么办？"
✓ 正确："症状可能是急性心梗，请立即拨打120或前往最近医院急诊！"
✗ 错误："可能是心绞痛，可以服用硝酸甘油"

## 专业性（10%）

**逻辑结构清晰**
- 回答结构化（分点陈述）
- 因果关系表达清楚

**语气适宜**
- 体现专业性（避免口语化）
- 具有同理心（患者友好）


© 2025 复旦大学护理学院

# 医疗模型评估框架 (2/2)

## 患者友好度（15%）

**通俗易懂**：避免专业术语、使用类比举例
**共情能力**：关注患者情绪、给予鼓励支持

## 实际效用（5%）

**问题解决率**：有效解答、解决效率
**临床价值**：患者健康管理、减轻医生负担

## 对比案例

**低质量回答**

*"您的血压140/90mmHg，属于高血压，需控制饮食、服用降压药，避免剧烈运动。"*

**问题**：缺乏个体化建议、无共情表达

**✓ 高质量回答**

*"您的血压轻度超标。建议：1）记录血压日记；2）适度运动；3）减少盐分摄入。我们会一起控制它！"*

**优点**：具体建议、有共情表达

## 综合评分方法

加权评分公式

综合得分 = 准确性×40% + 安全合规性×30%
+ 患者友好度×15% + 专业性×10% + 实际效用×5%

## 评估工具箱

| 评估维度 | 工具/方法 | 负责团队 |
|---|---|---|
| 专业、准确性 | 专家评审 | 医疗团队 |
| 安全合规性 | 安全检查清单 | 医疗+法务 |
| 患者友好度 | 患者问卷(n≥30) | 产品团队 |
| 实际效用 | A/B测试、长期追踪 | 运营团队 |

## 评估周期

**日常监控**
自动化评估

**月度评估**
全面审核

**季度优化**
模型迭代


© 2025 复旦大学护理学院

06

# 如何构建医学数据集

数据是模型的基石

# 数据集分类与格式

## 微调数据集的分类体系

**1. 指令微调数据集**
单轮任务（翻译、摘要、问答）

**2. 对话微调数据集**
多轮对话、上下文理解

**3. 领域适配数据集**
医疗、法律、金融等垂域知识

**4. 文本分类数据集**
情感分析、内容审核、新闻分类、意图识别

**5. 推理微调数据集**
需要多步推理的复杂任务

## 数据集格式对比

### Alpaca格式

- 适用：单轮指令任务
- 结构：instruction + input + output
- 优点：结构简单，易于标注
- 缺点：无法表达多轮对话

### ShareGPT格式

- 适用：多轮对话任务
- 结构：conversations列表
- 优点：保留对话上下文
- 缺点：标注成本较高

*不同数据集格式影响模型对任务的理解方式*

# 医疗数据集特性

## 医疗数据集特殊性

**挑战1：专业性要求高**

- 需要医学专业人员参与标注
- 标注成本是通用数据的3-5倍

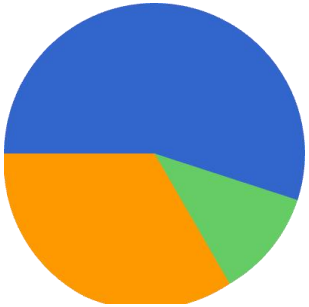
**挑战2：数据获取受限**

- 患者隐私保护（需脱敏处理）
- 医院数据审批流程长
- 需要伦理委员会审批

**挑战3：质量要求高**

- 医疗错误零容忍
- 数据准确性直接影响患者安全

## 数据集用途划分



**训练集 (70-80%)**

用于模型学习，应涵盖所有场景

**验证集 (10-15%)**

用于调整超参数，监控过拟合

**测试集 (10-15%)**

最终评估性能，模拟真实环境

***重要原则：三个集合必须严格分离，避免数据泄露***

# 1. 指令微调数据集

单轮任务数据构建与应用

## 标准Alpaca格式

```
{
"instruction": "将这段病历摘要翻译成英文",
"input": "患者男性，65岁，因'反复胸痛3天'入院。既往高血压病史10年。",
"output": "Male patient, 65 years old, admitted due to 'recurrent chest pain for 3 days'. Past medical history: hypertension for 10 years."
}
```

## 医疗场景应用

### 场景1: 病历摘要生成

**Instruction:** 根据门诊记录生成病历摘要

**Input:** 患者主诉：胸闷气短2周...
**Output:** 主诉：胸闷气短2周。现病史：...

### 场景2: 检查结果解读

**Instruction:** 用通俗语言解释这份心电图报告
**Input:** 窦性心律，心率72次/分，T波改变

**Output:** 您的心跳节律正常，心率在正常范围内...

### 场景3: 用药指导

**Instruction:** 说明阿司匹林的服用注意事项
**Input:** (空)
**Output:** 阿司匹林应在饭后服用，减少胃肠刺激...

## 数据来源

- 医院电子病历系统（脱敏后）
- 医学教材和指南
- 医学考试题库
- 医学论文的方法和结论部分

## 质量控制

- 每条数据至少2人标注，交叉验证
- 医学事实必须有权威来源支撑
- Input和Output长度建议：Input<500字，Output<1000字

## 数量建议

| 最少 | 推荐 | 理想 |
| --- | --- | --- |
| **100** | **1000-2000** | **5000+** |
| 快速验证 | 生产环境 | 多场景覆盖 |

# 2. 对话微调数据集

多轮交互式医疗对话构建

## ShareGPT格式详解

```
{"conversations": [{
"from": "human",
"value": "我最近总是心慌，这是怎么回事？"},
  {
"from": "gpt",
"value": "心慌可能有多种原因，比如：心律失常、甲状腺功能亢进等。能告诉我更多症状吗？"},
  {
"from": "human",
```

## 医疗场景应用

场景1：症状采集与初步判断

**患者：** 我最近总是心慌，这是怎么回事？

**AI：** 心慌可能有多种原因。能告诉我持续时间和频率吗？

**患者：** 每天都有，每次几分钟，有时伴有出汗。

**AI：** 建议做心电图、甲状腺功能和血常规检查。

## 多轮对话的优势

### 上下文理解

- 模型能记住之前的对话内容
- 用户无需重复背景信息

### 渐进式引导

- AI可以逐步收集信息
- 提供更精准的建议

### 更自然的交互

- 模拟真实医患对话
- 提升用户体验

## 构建建议

### 数据来源

- 真实医患对话记录（脱敏）
- 医生模拟标注
- 在线健康咨询平台

### 质量标准

- 对话轮次：3-8轮为宜
- 逻辑连贯性
- 信息递进式增加

### 标注技巧

- AI回复应包含"追问"
- 避免AI过早下诊断结论
- 最后一轮给出明确建议

### 注意事项

多轮对话数据的质量直接影响模型的交互能力和医疗建议准确性。高质量的对话数据应展现专业医生的问诊思路和判断逻辑，帮助模型学习规范的医疗咨询流程。

# 3. 领域适配数据集

医学专业知识体系构建

## 什么是领域适配数据集?

**定义**：针对特定专业领域（如医疗、法律）构建的数据集，包含大量领域专业知识和术语。

**目标**：让通用大模型快速掌握专业领域的知识体系和表达习惯。

## 领域知识的层次

第3层：治疗与管理
急诊处理、长期用药、康复指导

第2层：临床表现与诊断
症状学、检查方法、诊断标准

第1层：基础医学术语
解剖学、病理学、药理学

## 医学领域的细分

| 科室 | 核心疾病 | 推荐数据量 |
| --- | --- | --- |
| 心血管内科 | 冠心病、高血压、心衰 | 1000-1500条 |
| 呼吸科 | 哮喘、COPD、肺炎 | 800-1000条 |
| 消化科 | 胃炎、肝炎、胃溃疡 | 800-1000条 |

## 医学领域适配数据集示例

{

"instruction": "分析患者的症状描述并给出初步诊断建议",

"input": "55岁男性，持续性胸骨后压榨样疼痛3小时，向左肩放射，含服硝酸甘油无缓解，伴出汗",

"output": "根据症状分析，高度怀疑急性心肌梗死（STEMI）：

**典型表现**：
- 持续胸骨后压榨样疼痛（>30分钟）
- 向左肩放射（典型放射部位）
- 硝酸甘油无效（提示冠脉完全闭塞）
- 伴出汗（迷走神经兴奋）

## 构建方法

### 方法1：权威资料

来源：临床指南、教材
工具：PDF提取→结构化
优点：准确性高

### 方法2：病例改编

来源：电子病历系统
流程：提取→脱敏→审核
优点：贴近实际场景

### 方法3：AI辅助

使用GPT-4批量生成
医生逐条审核修改
优点：效率高，成本低

## 质量控制要点

### 准确性

医学事实必须有权威来源
符合最新指南（标注年份）
不含已淘汰的诊疗方法

### 完整性

覆盖常见病、多发病
包含诊断、治疗全流程
覆盖预防和康复内容

### 时效性

定期更新（每年1次）
及时纳入最新临床证据
标注最后更新日期

# 4. 文本分类数据集

## 示例格式

```
[
  {"text": "这款手机续航长达48小时，拍照效果惊艳", "label": "positive"},
  {"text": "系统频繁卡顿，客服响应速度慢", "label": "negative"},
  {"text": "量子计算机突破新型纠错码技术", "label": "science_news"},
  {"text": "央行宣布下调存款准备金率0.5个百分点", "label": "finance_news"}
]
```

## 应用场景

### 情感分析

商品评论情感极性识别（正面/负面/中性）

### 内容审核

检测违规内容（涉政/暴力/广告）

### 新闻分类

自动归类至财经/科技/体育等栏目

### 意图识别

用户query分类（咨询/投诉/比价）

## 数据质量要求

- **平衡性：** 各类别数据分布均衡，避免类别不平衡
- **一致性：** 标注标准统一，相似案例分类一致
- **覆盖性：** 包含各种边界情况和特殊表达
- **多样性：** 覆盖不同表达方式、语言风格

## 构建技巧

- **分层标签：** 使用分层标签体系（主类别+子类别）
- **多标签：** 模糊情况采用多标签标注
- **主动学习：** 使用主动学习策略选择难例
- **更新迭代：** 定期更新数据集，反映新趋势

# 05 推理微调数据集

## 示例格式

```
{
  "instruction": "解决数学应用题",
  "input": "小明买了3支铅笔，每支2元；又买了5本笔记本，
          每本比铅笔贵4元。总花费多少？",
  "chain_of_thought": [
    "铅笔单价：2元/支 → 3支总价：3×2=6元",
    "笔记本单价：2+4=6元/本 → 5本总价：5×6=30元",
    "合计花费：6+30=36元"
  ],
  "output": "总花费为36元"
}
```

## 应用场景

### 编程领域

解析复杂编程需求，生成高效代码，并提供智能调试支持

### 数学领域

擅长数学建模与逻辑推理，提供详细解题步骤和精确计算结果

### 数据分析

通过多步骤推理和策略规划，实现复杂数据的深度挖掘与分析

### 专业文档处理

精准提取法律合同和金融协议中的关键条款，解析模糊信息，辅助专业决策

# 如何构造领域适配数据集？

数据集构建流程

资料收集 → 内容提取 → 标准化处理 → 构建问答对 → 质量审核

权威资料、病例、指南 | 关键信息识别与整理 | 术语统一与格式转换 | 创建指令-输入-输出格式 | 专业验证与交叉检查

## 人工标注路径

专业人员直接整理或与领域专家协作标注数据，进行术语标准化并转换为模型可接受格式。这种方式虽耗时但质量可控，是构建高质量专业数据集的传统方法。

**优势：**

术语准确性高
内容可追溯性强
质量把控严格

## AI辅助构建路径

利用大模型辅助生成初始数据集，提高构建效率。但AI仅作为工具提效，非全自动解决方案。所有AI生成内容必须经过人工审核和校验，避免因缺乏专业把关导致数据质量不佳，进而影响模型微调效果。

**重要提示：**

*AI生成内容必须100%经过专业人员审核!*

## 参考资源

https://zhuanlan.zhihu.com/p/31150474598
https://github.com/ConardLi/easy-dataset

# Alpaca vs ShareGPT格式选择 (1/2)

## Alpaca格式

{
}
单

## ShareGPT格式

```
{"conversations": [{
      "from": "human",
      "value": "我最近总是心慌，这是怎么回事？"
    },{
      "from": "gpt",
      "value": "心慌可能有多种原因，比如：\n1. 心律失常\n2. 甲状腺功能亢进..."
    },{
      "from": "human",
      "value": "大概每天都会出现，每次持续几分钟..."
    },{
      "from": "gpt",
      "value": "根据您的描述，建议尽快做以下检查：..."}]}
```

多轮对话格式，保留完整上下文

## 医疗场景选择建议

### 选择Alpaca格式的场景：

医学知识问答（如"什么是高血压？"）
检查结果解读（输入报告→输出解释）
病历摘要生成（输入病历→输出摘要）
医学翻译（中英文互译）
数据量有限（<1000条）

### 选择ShareGPT格式的场景：

医患对话系统（需要多轮交互）
症状采集与初步诊断（需要追问）
健康咨询（需要根据用户反馈调整建议）
患者教育（需要根据理解程度调整解释）
数据量充足（≥1000条）

## 格式对比表

| 对比维度 | Alpaca格式 | ShareGPT格式 |
| --- | --- | --- |
| 核心设计目标 | 单轮指令驱动任务 | 多轮对话与上下文理解 |
| 数据结构 | instruction + input + output | conversations列表（human/gpt交替） |
| 标注难度 | 简单 | 较难 |
| 标注成本 | 低（每条¥5-10） | 高（每条¥20-50） |
| 训练效果 | 任务执行能力强 | 对话理解能力强 |
| 适用场景 | 问答、摘要、翻译 | 医患对话、客服咨询 |
| 数据量需求 | 500-1000条可见效 | 1000-2000条起步 |
| 工具支持 | 广泛支持 | 主流框架均支持 |

# Alpaca vs ShareGPT格式选择 (2/2)

## 混合使用策略

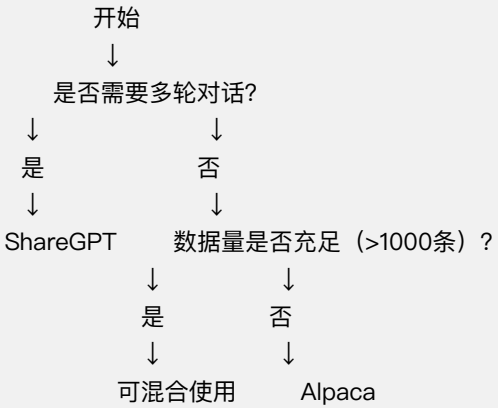
推荐方案：两种格式混合训练

- Alpaca数据：占60-70%（打好基础）
- ShareGPT数据：占30-40%（提升对话能力）
- 先用Alpaca数据训练，再用ShareGPT数据精调

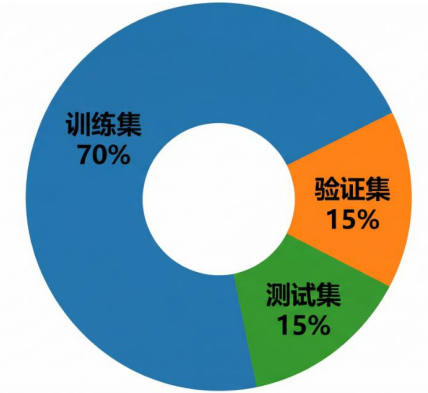
中山医院案例的实践：

- Alpaca数据：1200条（医学知识问答）
- ShareGPT数据：800条（医患对话）
- 总计：2000条

## 决策流程图



# 训练集、验证集、测试集详解(1/3)




## 验证集（Validation Set）10-15%

**作用**：调整超参数、监控过拟合

**类比**：平时的模拟考试

**要求**：

与训练集独立，但分布相似

包含边缘案例、复杂情况

用于选择最佳模型版本

## 训练集（Training Set）70-80%

**作用**：模型从中学习知识和规律

**类比**：学生的课本和练习题

**要求**：

数量最多，涵盖所有场景

覆盖常见病、多发病

包含典型案例和标准答案

## 测试集（Test Set）10-15%

**作用**：最终评估模型性能

**类比**：期末考试（真实评估）

**要求**：

完全独立，模拟真实环境

只用一次，不能调参

包含新型案例，考察泛化能力

# 训练集、验证集、测试集详解(2/3)

## 医疗场景的特殊考虑

### 按疾病分层采样

确保每种疾病在三个集合中都有代表，避免某种疾病只在训练集或测试集中

示例：2000条数据，包含5种疾病

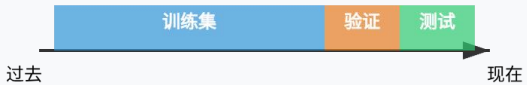
| 疾病类型 | 总数据量 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|---|
| 疾病A | 600条 | 420 | 90 | 90 |
| 疾病B | 500条 | 350 | 75 | 75 |
| 疾病C | 400条 | 280 | 60 | 60 |
| 疾病D | 300条 | 210 | 45 | 45 |
| 疾病E | 200条 | 140 | 30 | 30 |

### 时间序列数据的特殊处理

如果数据有时间属性（如病情进展）

不能随机打乱，应按时间顺序划分

训练集用旧数据，测试集用新数据




## 数据量不足时的应对

问题：数据总量<500条时，测试集样本太少，评估不准确

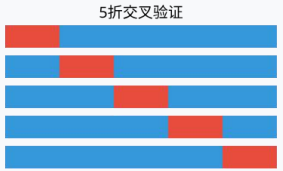
### 1. 交叉验证（Cross-Validation）

将数据分成5份（5-Fold）

轮流用4份训练、1份验证

最终取5次结果的平均值

**适用**：数据量100-1000条




### 2. 留一法（Leave-One-Out）

每次留1条作为测试，其余训练

重复n次（n为数据总量）

适用：数据量<100条（计算成本高）


### 3. 增加数据

数据增强：同义改写、回译

AI生成+人工审核

跨机构数据共享

# 训练集、验证集、测试集详解(3/3)

## 实用建议

| 数据总量 | 推荐划分 | 备注 |
| --- | --- | --- |
| <100条 | 留一法交叉验证 | 仅用于快速验证 |
| 100-500条 | 5折交叉验证 | 可信度中等 |
| 500-2000条 | 70/15/15固定划分 | 标准做法 |
| >2000条 | 80/10/10固定划分 | 数据充足，可增大训练集比例 |

## 医疗数据集划分要点

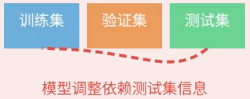
- **再现性**：记录随机种子，确保结果可复现
- **代表性**：数据分布应涵盖真实临床场景
- **安全性**：所有数据必须完全脱敏
- **平衡性**：注意罕见疾病和常见疾病的平衡
- **更新策略**：随着新数据加入，定期更新划分

## 常见错误

### 错误1：测试集泄露

在训练前就查看测试集内容，根据测试集表现调整模型

**后果**：模型"作弊"，实际效果不佳




### 错误2：分布不一致

训练集都是简单问题，测试集都是难题

**后果**：模型在测试集上表现差


### 错误3：数据重复

同一个问题的不同表述出现在训练集和测试集

**后果**：高估模型性能

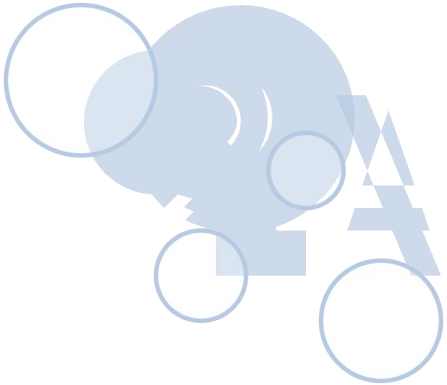


© 2025 复旦大学护理学院

07

# 常见QA

问题解答



# 精选问题

## 数据需求

Q：需要多少数据才够？

A：LoRA理论上需1000条，实际100条高质量数据已可见效。

## 原有能力保护

Q：微调会破坏原有能力吗？

A：KL约束就像安全带，防止AI跑偏，保护原有性能。

## 消费级电脑能力

Q：消费级电脑能微调多大模型？

A：16GB显存可QLoRA微调13B模型，8GB显存可微调7B模型。

# 精选问题

Q：数据不够怎么办？

A：交叉验证法：将完整集分成5份，轮流用4份训练、1份验证（类似「轮换座位考试」），合成数据：用图像翻转、文字替换等方式扩充数据量。

Q：特殊场景处理？

A：时间序列数据：需按时间顺序划分（不能用随机拆分）。例如预测股价，必须用2023年前的数据训练，2024年数据测试。

Q：为什么不能混用？

A：如果测试集数据泄露到训练中，就像考前背答案，实际应用时遇到新题就会失败。

# 资源合作

关注《昱麟AI健康公众号》，获取更多医疗前沿大模型知识分享

《科研桥》小程序，发布任务，提供相关的技术支持。