可用的场景及数据集

场景 1（医疗）：循证问答 / 医学考试问答（公开数据、易评测）

任务定义：给出医学问题，系统需要基于检索到的医学文献/摘要/指南，输出答案（可带引用）。
优点：公开数据多、评价指标清晰、容易做对比。

可选数据集：

- **PubMedQA**：来自 PubMed 摘要的研究问答（yes/no/maybe + 长答案），含大量未标注与自动生成样本，适合做 RAG（检索摘要 → 回答）。
- **BioASQ-QA**：BioASQ 挑战赛问答数据，强调真实生物医学信息需求，适合做"检索 + 多文档证据"。BioASQ 官方长期维护挑战与数据入口。
- **MedMCQA**：19.4 万+ 医学考试多选题，覆盖 21 学科、附解释文本，适合用 RAG 检索"解释/教材片段/知识库"来做选择或生成解释。
- **MedQA（USMLE 等）**：医学考试多选题，社区维护版本较多，适合做对照实验（不带检索 vs 带检索）。

如何把它做成 RAG 场景：

1. **语料库**：PubMed 摘要（PubMedQA 自带 context）、或 BioASQ 提供的相关材料；也可额外加入公开指南文本（如果时间够）
2. **切分**：按句群/段落 chunk（如 200–400 tokens）+ 保留文章标题、年份、期刊等 metadata
3. **检索**：BM25（传统） vs 向量检索（dense） vs 混合检索（加分）
4. **生成**：LLM 读取 top-k 证据，要求"必须引用证据句/段"
5. **评测**：准确率 + 引用一致性抽检（给现实意义加分）

场景 2（法律）：合同审阅 / 判例推理（最能体现"可追溯"）

任务定义 A（合同审阅）：给合同文本，模型需要定位/抽取关键条款位置或回答条款是否存在。

- **CUAD**：合同审阅数据集，面向法律合同 review，含大量专家标注，适合做"检索条款片段 + 生成结论/定位"。

任务定义 B（判例 holding 选择）：给一段司法判决背景，选择最正确的 holding（类似法律推理）。

- **CaseHOLD**：5.3 万+ 多选题，围绕判决引用与 holding，适合做"检索相似判例段落/法条摘要 + 推理选择"。

法律任务打包式 benchmark：

- **LexGLUE**：法律语言理解基准，整合 7 个法律 NLP 数据集，能快速构造"多任务、多指标"的实验框架（分类/检索/推理等）。

你如何把它做成 RAG 场景：

- CUAD：把合同切 chunk → 检索与问题相关片段 → 输出"条款是否存在/在哪里/风险点"，并引用证据段落
- CaseHOLD：把训练集中出现的判例文本建库（或公开判例摘要）→ 检索相似事实/争点 → 帮助模型选 holding（并输出引用依据）

评测指标建议：

- CUAD：抽取/定位类 F1、EM（看任务设置）；加一个"证据段落是否覆盖标注 span"的覆盖率

- CaseHOLD：Accuracy；加“检索到的证据是否来自同类争点”的人工小样本审计

要交付的“数据集清单 + 场景说明”

医疗（公开、易复现）

1. **PubMedQA**（生物医学研究问答，yes/no/maybe + 长答案）——适合 RAG 检索摘要后回答
2. **BioASQ-QA**（比赛型生物医学问答，多文档证据）——适合做“检索+证据一致性”
3. **MedMCQA**（大规模医学多选题+解释）——适合做“检索教材/解释→提升选择正确率与解释质量”
4. **MedQA（USMLE 等）**（医学考试多选题）——适合做 strong baseline 对照

法律（证据链）

1. **CUAD**（合同审阅、专家标注）——适合条款定位/抽取/风险提示
2. **CaseHOLD**（判例 holding 多选推理）——适合法律推理与检索增强
3. **LexGLUE**（法律 NLP 基准合集）——适合“一个项目跑多任务”展示泛化
4. **ECtHR cases**（欧洲人权法院判例数据）——适合法条/判例预测与证据检索（如果你们想拓展到非美法域）