深度检索 Agent 测试 Benchmark

领域：生物医药管线竞争情报（Pharma Competitive Intelligence）

一、Benchmark 设计理念

基于你的管线价值评估文档，设计 **5 级难度递进** 的测试任务：

级别	类型	典型耗时	核心能力考察
L1	单一事实检索	<1min	精准定位、信息抽取
L2	多源聚合	1-3min	多来源整合、数据清洗
L3	格局分析	3-10min	结构化输出、竞争对标
L4	综合研判	10-30min	多步推理、假设验证
L5	战略决策	30min+	框架应用、情景分析

二、L1 级：单一事实检索（Factual Retrieval）

评估维度

- **准确性**：答案是否正确
- **来源权威性**：是否引用一手来源（FDA/EMA/ClinicalTrials.gov）
- **时效性**：信息是否最新

测试用例

1.1 药物基本信息

```
Q: Keytruda（pembrolizumab）在美国的首次获批日期和适应症是什么？
Expected: 2014年9月4日，晚期黑色素瘤
Source: FDA approval letter
```

1.2 临床试验状态

```
Q: NCT04379596 这个临床试验目前是什么状态？主要终点是什么？
Expected: [需实时检索]
Source: ClinicalTrials.gov
```

1.3 专利信息

```
Q: Humira（adalimumab）的美国化合物专利什么时候到期？
Expected: 2016年12月31日（但生物仿制药受其他专利阻挡至2023年）
Source: Orange Book / USPTO
```

1.4 定价信息

```
Q: Opdivo 在美国的 WAC 价格是多少（按 mg 计）？
Expected: [需实时检索当前价格]
Source: Red Book / SSR Health
```

1.5 监管资格

```
Q: Vertex 的 VX-548 获得了哪些 FDA 特殊审评资格？
Expected: Breakthrough Therapy Designation, Priority Review
Source: FDA press release
```

三、L2 级：多源信息聚合（Multi-Source Aggregation）

评估维度

- **完整性**：是否覆盖所有关键来源
- **一致性处理**：不同来源数据冲突时的处理
- **结构化程度**：输出是否便于后续使用

测试用例

2.1 流行病学数据聚合

```
Q: 请整理非小细胞肺癌（NSCLC）在中国的流行病学数据：
   - 年新发病例数
   - EGFR突变阳性比例
   - 一线治疗后进展的患者比例
   - 5年生存率

Expected Output Format:
| 指标 | 数值 | 来源 | 年份 |
|------|------|------|------|
| 新发病例 | 82.8万 | Globocan | 2022 |
| EGFR+ 比例 | 40-50% | CSCO指南 | 2024 |
| ... | ... | ... | ... |
```

2.2 竞品管线汇总

```
Q: 列出所有针对 KRAS G12C 突变的在研药物（Phase 2及以上），包括：
   - 所属公司
   - 开发阶段
   - 给药方式
   - 临床试验编号

Expected: 结构化表格，至少包含 Sotorasib, Adagrasib, Divarasib 等
Sources: ClinicalTrials.gov, 公司官网, SEC filings
```

2.3 指南推荐整合

```
Q: 比较 NCCN、ESMO、CSCO 三个指南对 HER2+ 乳腺癌一线治疗的推荐差异

Expected: 三栏对比表，标注推荐等级、证据级别、具体方案
```

2.4 专利地图

```
Q: 整理 Ozempic（semaglutide）的专利布局：
   - 化合物专利
   - 制剂专利
   - 用途专利（各适应症）
   - 各专利在美/欧/中的到期时间

Expected: 专利家族表 + 时间线图数据
```

四、L3 级：竞争格局分析（Competitive Landscape）

评估维度

- **分析框架应用**：是否使用合理的分析结构
- **关键洞见提取**：能否识别核心竞争要点
- **可视化建议**：是否给出图表建议

测试用例

3.1 竞争管线 Benchmark

```
Q: 我们正在开发一款 CD19 CAR-T 产品，计划用于 r/r DLBCL 三线治疗。
   请分析当前竞争格局：
   1. 已上市竞品的疗效/安全性数据对比
   2. 在研管线中最具威胁的3个竞品
   3. 我们需要在哪些终点上达到什么水平才能形成差异化？

Expected Output:
- 已上市产品对比表（Yescarta, Kymriah, Breyanzi, Tecartus）
- ORR, CR, CRS发生率, 神经毒性, 中位PFS等关键数据
- 在研管线威胁评估（按时间线+数据强度排序）
- 差异化假设建议
```

3.2 临床设计对标

```
Q: 分析 PD-1/PD-L1 抑制剂在一线 NSCLC（PD-L1≥50%）的关键 Phase 3 试验设计：
   - 主要终点设置
   - 对照组选择
   - 入组标准差异
   - 中位随访时间

   并建议：如果我们现在做这个适应症，试验设计需要注意什么？

Expected: KEYNOTE-024, IMpower110, CheckMate-026 等试验的结构化对比
```

3.3 时间线竞争路径

```
Q: 绘制 ADC 药物在 HER2-low 乳腺癌领域的竞争时间线:
   - 谁已经获批?
   - 谁预计 2025-2027 年获批?
   - 关键数据读出时间点
   - 可能改变指南的里程碑事件

Expected: 时间轴数据（可生成甘特图），标注关键事件
```

3.4 First-in-class vs Best-in-class 分析

```
Q: 分析 BTK 抑制剂的代际竞争:
   - 一代（Ibrutinib）vs 二代（Acalabrutinib, Zanubrutinib）vs 三代
（Pirtobrutinib）
   - 各代的核心差异化点
   - 市场格局演变预测

Expected: 代际对比分析 + 市场份额变化逻辑
```

五、L4 级：综合研判（Comprehensive Assessment）

评估维度

- **多维度整合**：流行病学+竞争+临床+商业的综合
- **假设驱动**：基于用户输入进行情景分析
- **数据支撑**：结论有明确数据来源

测试用例

4.1 资产画像 + SoC 分析

```
Context:
我们有一款口服 SERD（选择性雌激素受体降解剂），
靶向 ESR1，用于 ER+/HER2- 晚期乳腺癌。
临床数据显示：ORR 25%，中位 PFS 7.2 个月。

Q: 请完成以下分析:
```

```
1. 疾病画像：ER+/HER2− 晚期乳腺癌的治疗路径和未满足需求
2. 现有 SoC：内分泌治疗 → CDK4/6i → 化疗的格局
3. 我们的定位机会：口服替代 Fulvestrant 注射？CDK4/6i 耐药后？
4. 与 Elacestrant（已获批口服 SERD）的数据对比

Expected: 完整的资产定位分析报告
```

4.2 市场测算验证

```
Context:
用户输入的市场假设：
- 目标人群：20,000 患者/年
- 渗透率峰值：15%（第5年）
- 年治疗费用：$150,000

Q: 请验证这些假设的合理性：
1. 患者人数是否与流行病学数据一致？
2. 15% 渗透率在同类产品中处于什么水平？
3. 定价与竞品相比是否合理？
4. 指出假设中的风险点

Expected: 假设验证报告 + 敏感性分析建议
```

4.3 TPP 合理性评估

```
Context:
用户提交的 TPP：
- 适应症：EGFR 突变 NSCLC 一线
- 主要终点：PFS ≥ 18 个月
- ORR：≥ 75%
- 安全性：3 级以上 AE < 30%
- 给药：口服 QD

Q: 评估这个 TPP：
1. 与已上市 EGFR TKI（Osimertinib, Lazertinib 等）的数据相比是否可实现？
2. 如果达不到 18 个月 PFS，多少是可接受的最低标准？
3. 安全性目标是否足够差异化？

Expected: TPP 评审意见 + Benchmark 数据支撑
```

4.4 BD 标的评估

```
Q: 评估以下 BD 标的的价值：
   - 某公司的 IL-23 抑制剂，Phase 2 数据刚读出
   - 适应症：中重度银屑病
   - PASI 90 应答率：72%（12周）
```

```
   - 给药：SC Q12W

需要分析：
1. 与 Skyrizi, Tremfya 等已上市 IL-23i 的数据对比
2. 差异化空间在哪里？
3. 合理的估值区间参考（参考同类交易）

Expected: BD 评估报告 + 可比交易清单
```

六、L5 级：战略决策支持（Strategic Decision）

评估维度

- **框架完整性**：是否覆盖文档定义的所有分析步骤
- **结论可行性**：Go/No-Go 建议是否有充分依据
- **战略视角**：是否考虑公司能力匹配

测试用例

5.1 完整管线立项评估

```
Q: 完成以下管线的完整立项评估：

候选资产：一款 Claudin 18.2 ADC
适应症：胃癌 2L+
当前阶段：Pre-IND
公司背景：中型 Biotech，有 ADC 平台，无肿瘤销售团队

请按以下框架输出：
1. Step 1: 疾病画像 + 市场规模
2. Step 2: SoC & Unmet Need
3. Step 3: 竞争管线分析（重点：Astellas 的 Zolbetuximab）
4. Step 4: TPP 建议 + 敏感性分析
5. Step 5: 峰值销售预测（3 种情景）
6. Step 6: Go/No-Go 建议 + 战略选项

Expected: 完整的立项评估报告（包含所有 Step）
```

5.2 LCM 战略规划

```
Context:
我们有一款已上市的 PD-1 抑制剂，当前适应症：
- 黑色素瘤（已获批）
- NSCLC（已获批）
- 胃癌（Phase 3 进行中）

Q: 制定 LCM 战略：
1. 下一步应该拓展哪些适应症？（按：市场规模 × 成功率 × 竞争强度 排序）
2. 是否应该开发联合用药方案？与哪些机制联合？
3. 是否需要开发新剂型（如皮下制剂）？

Expected: LCM 路线图 + 优先级矩阵
```

5.3 风险预警情景模拟

```
Q: 模拟以下情景对我们管线的影响：

情景1：竞品 A 的 Phase 3 头对头试验结果显示优于 SoC
情景2：FDA 更新指南，要求 OS 作为主要终点
情景3：主要竞品降价 30%

对于每个情景：
1. 对我们资产价值的影响（定量估算）
2. 应对策略建议
3. 需要提前准备的工作

Expected: 情景分析表 + 应对预案
```

七、评估标准（Rubric）

7.1 单项评分维度

维度	权重	评分标准（1-5分）
准确性	25%	事实是否正确，数据是否准确
完整性	20%	是否覆盖所有要求的信息点
来源质量	15%	是否引用权威一手来源
结构化	15%	输出是否便于阅读和后续使用
洞见深度	15%	是否提供超越事实的分析见解

维度	权重	评分标准（1-5分）
时效性	10%	信息是否为最新

7.2 分级通过标准

级别	最低通过分	关键要求
L1	4.0	准确性必须 ≥ 4.5
L2	3.8	完整性必须 ≥ 4.0
L3	3.5	结构化必须 ≥ 4.0
L4	3.5	洞见深度必须 ≥ 3.5
L5	3.2	所有维度必须 ≥ 3.0

7.3 特殊扣分项

- **幻觉**：捏造不存在的临床试验/数据 → 该项直接 0 分
- **来源错误**：引用来源与内容不符 → 来源质量 -2 分
- **过时信息**：使用 2 年以上的数据（未标注）→ 时效性 -2 分
- **关键遗漏**：遗漏最主要的竞品/风险 → 完整性 -2 分

八、澄清模块专项测试

8.1 测试澄清是否"问到点子上"

Case 1：模糊的适应症

```
User: "帮我分析一下 PD-1 抑制剂的市场"

Good Clarification:
- "您关注的是哪个具体适应症？（肺癌/黑色素瘤/胃癌/泛瘤种）"
- "是全球市场还是特定区域（中美欧日）？"
- "是整体市场格局还是某个特定产品的竞争分析？"

Bad Clarification:
```

- "您想了解哪些方面？"（太宽泛）
- "请提供更多信息"（无引导性）

Case 2：缺少关键参数

```
User: "评估一下我们这个 ADC 管线的价值"

Good Clarification:
- "请提供靶点和适应症"
- "当前开发阶段？"
- "是否有临床数据可以分享？"
- "评估目的是内部立项还是 BD 定价？"

Missing Critical:
- 不问靶点就开始分析 → 失败
```

Case 3：隐含的决策问题

```
User: "Claudin 18.2 这个靶点怎么样？"

识别真实意图的追问：
- "您是在考虑进入这个靶点还是评估现有投资？"
- "已经有在研项目了吗？还是在做早期调研？"
- "主要关注胃癌还是其他适应症？"
```

8.2 澄清效率评估

指标	目标值
平均澄清轮数	≤ 2 轮
首轮澄清命中率	≥ 80%（首轮问到关键槽位）
过度澄清率	≤ 10%（问了不影响结果的问题）
澄清后任务成功率提升	≥ 30%

九、数据源验证清单

9.1 核心数据源（必须能访问）

数据类型	权威来源	验证方式
临床试验	ClinicalTrials.gov	检索 NCT 编号
FDA 审批	FDA.gov	Orange Book, Purple Book
EMA 审批	EMA.europa.eu	EPAR 文档
流行病学	Globocan, GBD	引用年份确认
指南	NCCN, ESMO, CSCO	版本号确认
定价	Red Book, MEPS	日期确认
专利	USPTO, EPO, CNIPA	专利号验证

9.2 来源优先级

```
一手来源（优先）：
├── 监管机构官网（FDA, EMA, NMPA）
├── 临床试验注册库（ClinicalTrials.gov）
├── 公司官方披露（SEC filings, Press Release）
└── 学术期刊原文（NEJM, Lancet, JCO）

二手来源（谨慎使用）：
├── 行业数据库（IQVIA, Evaluate Pharma）
├── 投行研报
└── 新闻媒体

避免使用：
├── 无来源的博客
├── 非官方数据库
└── 过时的综述（>3年）
```

十、Benchmark 测试流程

10.1 测试执行步骤

```
1. 随机选取每级 3-5 个测试用例
2. 记录：
   - 澄清轮数
   - 总耗时
   - 工具调用次数
```

```
   - 来源数量
3. 人工评分（按 Rubric）
4. LLM-as-Judge 辅助评分
5. 计算各维度得分和总分
```

10.2 基准对比

建议将你的 Agent 与以下基准对比：

基准	说明
人类分析师	有 3 年以上 CI 经验的分析师完成同样任务
Claude Research	使用 Claude Research 模式
Perplexity Pro	使用 Deep Research 模式
GPT-4 + 手动搜索	GPT-4 + 人工辅助检索

10.3 迭代优化指标

```
核心指标：
- L1-L3 综合通过率 → 目标 ≥ 85%
- L4-L5 综合通过率 → 目标 ≥ 70%
- 平均任务完成时间 → 比人类快 5x+
- 幻觉率 → 目标 < 2%

效率指标：
- Token 使用效率（准确率 / Token 数）
- 工具调用效率（信息量 / 调用次数）
```

十一、测试数据集构建建议

11.1 Golden Set（必须 100% 通过）

选取 20-30 个**高频**、**关键**的查询，这些是用户最常问的问题：

```
golden_set = [
    # L1: 基础事实
```

```
    {"query": "Keytruda 2024年全球销售额", "answer": "$25B+", "source": "Merck
10-K"},
    {"query": "Tagrisso 的适应症", "answer": "EGFR突变NSCLC一线/辅助",
"source": "FDA label"},

    # L2: 多源聚合
    {"query": "中国NSCLC患者数", "expected_sources": ["Globocan", "CSCO指
南"]},

    # L3: 竞争分析
    {"query": "KRAS G12C 抑制剂竞争格局", "must_include": ["Sotorasib",
"Adagrasib"]},
]
```

11.2 动态测试集

每周/每月更新，跟踪实时事件：

```
dynamic_tests = [
    "上周 FDA 批准了哪些新药？",
    "最近一个月 Phase 3 读出的肿瘤药有哪些？",
    "2024 年最大的 Pharma BD 交易是什么？",
]
```

11.3 对抗测试集

故意设置陷阱，测试幻觉和错误处理：

```
adversarial_tests = [
    # 不存在的药物
    {"query": "Xantellix 的临床数据", "expected": "无法找到该药物信息"},

    # 过时信息
    {"query": "Herceptin 的专利状态", "trap": "不能只说已过期，需说明生物仿制药已上
市"},

    # 歧义查询
    {"query": "PD-1 最好的药是哪个？", "expected": "需要澄清评价维度"},
]
```

附录：快速启动测试清单

第一阶段（1周）：验证基础能力

- ▢L1 测试：10 个单一事实查询
- ▢来源验证：检查是否引用权威来源
- ▢幻觉检测：5 个对抗测试

第二阶段（2周）：验证聚合能力

- ▢L2 测试：5 个多源聚合任务
- ▢结构化输出：检查表格/对比格式
- ▢数据一致性：多来源冲突处理

第三阶段（2周）：验证分析能力

- ▢L3 测试：3 个竞争格局分析
- ▢L4 测试：2 个综合研判任务
- ▢澄清测试：10 个模糊查询的澄清质量

第四阶段（2周）：验证端到端能力

- ▢L5 测试：1-2 个完整管线评估
- ▢与人类基准对比
- ▢效率和成本评估