# RAG知识库多版本评测分析报告

评测基准: 30个医疗相关问题的测试集，以下是本测试使用的问题集：

# 症状查询

"什么是偏头痛，它和普通头痛有什么区别？": ["偏头痛", "头痛"],

"经常感觉疲劳乏力是怎么回事？": ["疲劳", "乏力"],

"晚上睡觉时小腿抽筋是什么原因？": ["抽筋", "肌肉痉挛"],

"皮肤上出现红疹还很痒，可能是哪些情况？": ["红疹", "皮疹", "瘙痒"],

"除了感冒，还有什么病会引起喉咙痛？": ["喉咙痛", "咽痛"],

"如何缓解胃酸反流（烧心）的症状？": ["胃酸反流", "烧心"],

# 疾病查询

"请详细介绍一下2型糖尿病。": ["2型糖尿病", "糖尿病"],

"高血压的诊断标准是什么？": ["高血压"],

"如何预防骨质疏松？": ["骨质疏松"],

"胆结石必须要做手术吗？有哪些治疗方法？": ["胆结石"],

"抑郁症的早期症状有哪些？": ["抑郁症"],

"荨麻疹的病因是什么？能根治吗？": ["荨麻疹"],

# 药品查询

"布洛芬缓释胶囊是用来做什么的？": ["布洛芬"],

"阿莫西林的用法用量和注意事项是什么？": ["阿莫西林"],

"蒙脱石散有哪些副作用？": ["蒙脱石散"],

"请问“川芎口服液”的成分是什么？": ["川芎口服液"],

"哺乳期妇女可以服用对乙酰氨基酚吗？": ["对乙酰氨基酚"],

"降压药需要终身服用吗？": ["降压药", "高血压"],

# 检查与治疗查询

"核磁共振（MRI）检查是用来做什么的？": ["核磁共振"],

"什么是靶向治疗？": ["靶向治疗"],

"做胃镜之前需要做哪些准备？": ["胃镜"],

"“血常规”检查能查出什么问题？": ["血常规"],

# 复合与场景化查询

"我最近总是头晕，还恶心想吐，会是什么病？": ["头晕", "恶心", "呕吐"],

"孩子发烧咳嗽，可以吃点头孢吗？": ["发烧", "咳嗽", "头孢"],

"运动后膝盖疼，是应该冷敷还是热敷？": ["膝盖疼", "关节痛"],

"我妈妈有高血压，饮食上应该注意些什么？": ["高血压"],

"感冒和流感有什么不一样，怎么区分？": ["感冒", "流感"],

"长期失眠应该怎么办，有什么推荐的治疗方法？": ["失眠"],

"吃完海鲜后身上起了很多红点，非常痒，我该怎么办？": ["过敏", "荨麻疹", "红点"],

"体检发现尿酸高，需要吃药吗？平时要注意什么？": ["尿酸高", "痛风"],

核心指标: Hit Rate @ 3 (返回的前3个结果中，来源文档标题包含目标关键词的命中率)

## 1. 总体性能对比

### 经过对六个不同版本知识库的自动化评测，得到了清晰的性能排序。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **排名** | **实验版本 (Experiment)** | **命中率 (Hit Rate)** | **命中数/总数** | **核心策略** |
| 🥇 **1** | **v6\_recursive\_bge-large-zh-v1.5** | **86.67%** | **26 / 30** | **大尺寸递归切分 + 高性能模型** |
| 🥈 2 | v4\_recursive\_bge-large-zh-v1.5 | 76.67% | 23 / 30 | 中尺寸递归切分 + 高性能模型 |
| 🥉 3 | v2\_recursive\_text2vec\_base | 73.33% | 22 / 30 | 中尺寸递归切分 + 基准模型 |
| 🥉 3 | v3\_recursive\_m3e\_base | 73.33% | 22 / 30 | 中尺寸递归切分 + M3E模型 |
| 5 | v5\_recursive\_bge-large-zh-v1.5 | 63.33% | 19 / 30 | 小尺寸递归切分 + 高性能模型 |
| 6 | v1\_section\_text2vec\_base | 56.67% | 17 / 30 | 按分区切分 + 基准模型 |

## 2. 影响准确率的因素总结与深度分析

通过对不同版本性能差异的归因分析，可以得出关于如何优化RAG检索器的三个关键结论。

### 影响因素一：文本切分 (Chunking) 策略

**结论：** 对于本医疗数据集，**递归字符切分（Recursive Splitting）策略的性能显著优于按分区切分（Chunk by Section）策略。**

* **数据支撑**: v2 (73.33%) 相较于 v1 (56.67%)，在**使用相同模型**的情况下，仅因改变了切分策略，命中率就**提升了16.66个百分点**。
* **原因分析**:
  + **信息密度**: “按分区切分”产生的知识片段（Chunks）大小不均。一个内容较长的分区，其包含的“平均语义”可能较为发散，降低了与精确用户问题的匹配度。
  + **检索聚焦**: “递归切分”通过控制Chunk的最大尺寸，将长文本切分成了更聚焦的片段。这些片段的向量在语义空间中指向更明确，从而更容易被用户的查询向量检索到。

### 影响因素二：向量模型 (Embedding Model)

**结论：** **向量模型的能力是决定检索性能上限的关键因素。BAAI/bge-large-zh-v1.5 在本次评测中表现出明显的性能优势。**

* **数据支撑**:
  + v4 (76.67%) 相较于 v2 (73.33%) 和 v3 (73.33%)，在**采用相同的切分策略和尺寸**下，仅因将模型从base级更换为bge-large，性能就获得了可观的提升。
  + v6 (86.67%) 则在v4的基础上，通过优化Chunk尺寸，进一步发挥了bge-large模型的性能潜力。
* **原因分析**:
  + **语义理解深度**: bge-large 作为一个更大规模的模型，能够更精确地捕捉中文医学术语之间的语义关系，从而在相关性排序上表现更佳。
  + **模型对比**: text2vec-base 和 m3e-base 在本次评测中性能表现相近，均可作为有效的基准模型，但与bge-large相比，在检索准确率上存在差距。

### 影响因素三：Chunk尺寸 (Chunk Size)

**结论：** 对于本医疗数据集和bge-large模型，**较大尺寸的Chunk（512）表现最佳，而过小的Chunk（128）则导致性能明显下降。**

* **数据支撑**: 在**使用相同的bge-large模型和recursive切分策略**下，不同尺寸的Chunk性能差异显著：
  + **v6 (Chunk Size = 512): 86.67%**
  + v4 (Chunk Size = 250): 76.67%
  + v5 (Chunk Size = 128): 63.33%
* **原因分析**:
  + **上下文完整性**: 此结果表明，医疗领域的知识片段需要**足够的上下文**来保证信息的完整性。一个过小的Chunk（如128字符），可能只包含部分信息而丢失了关键主语（如药品名称），从而降低了作为检索结果的有效性。
  + **尺寸与性能的平衡**: 512字符的尺寸，在“信息密度”和“上下文完整性”之间为本数据集找到了一个较好的平衡点，使得检索效果最优化。

## 3. 最终结论与战略建议

基于本次全面的自动化评测，我们可以得出以下最终结论：

**最优策略**: **v6\_recursive\_bge-large-zh-v1.5** 版本所采用的 **“大尺寸递归切分 (Chunk Size=512) + 高性能中文向量模型 (bge-large-zh-v1.5)”** 组合，是本次所有实验中的最优方案，其命中率达到了**86.67%**。