# 数据集更新说明报告

对比版本:

* **v1**: cleaned\_medical\_data.json (修复前)
* **v2**: cleaned\_medical\_data\_v2.json (修复后)

**核心议题:** 本报告旨在说明 v2 版本数据集相较于 v1 版本的核心改进。v1 版本在数据质量分析中暴露出内容不完整的问题。

### 1. 问题识别：v1 版本的数据缺陷

在对 v1 数据集进行质量分析时，分析显示存在大量\*\*“结构单一”\*\*的词条。

这些词条，尤其是“药品”类别，尽管在源网站上拥有丰富的内容分区（如“功效作用”、“用法用量”、“禁忌”、“注意事项”等），但在数据集中却只显示有一个“概述”分区，且内容极短，通常只包含了药品的成分列表。

这种现象表明，数据集存在严重的**内容缺失**问题，属于不合理的数据状态。这些不完整的词条作为RAG系统的知识源，信息量过低，无法为用户查询提供有价值的回答，构成了“数据噪音”。

### 2. 最终成果：v2 版本的核心价值

通过修复，最终生成了 cleaned\_medical\_data\_v2.json。

与 v1 版本相比，v2 版本的核心提升在于**数据的完整性和丰富度**。之前那些仅有寥寥数语成分介绍的药品词条，现已更新为包含用法用量、功效作用、注意事项等完整信息的、高质量的知识片段。

**结论：** v2 版本通过修复初代爬虫的逻辑缺陷并对不完整数据进行精准重抓，极大地提升了知识库的整体质量，为构建一个高性能、高准确率的RAG系统奠定了更坚实的基础