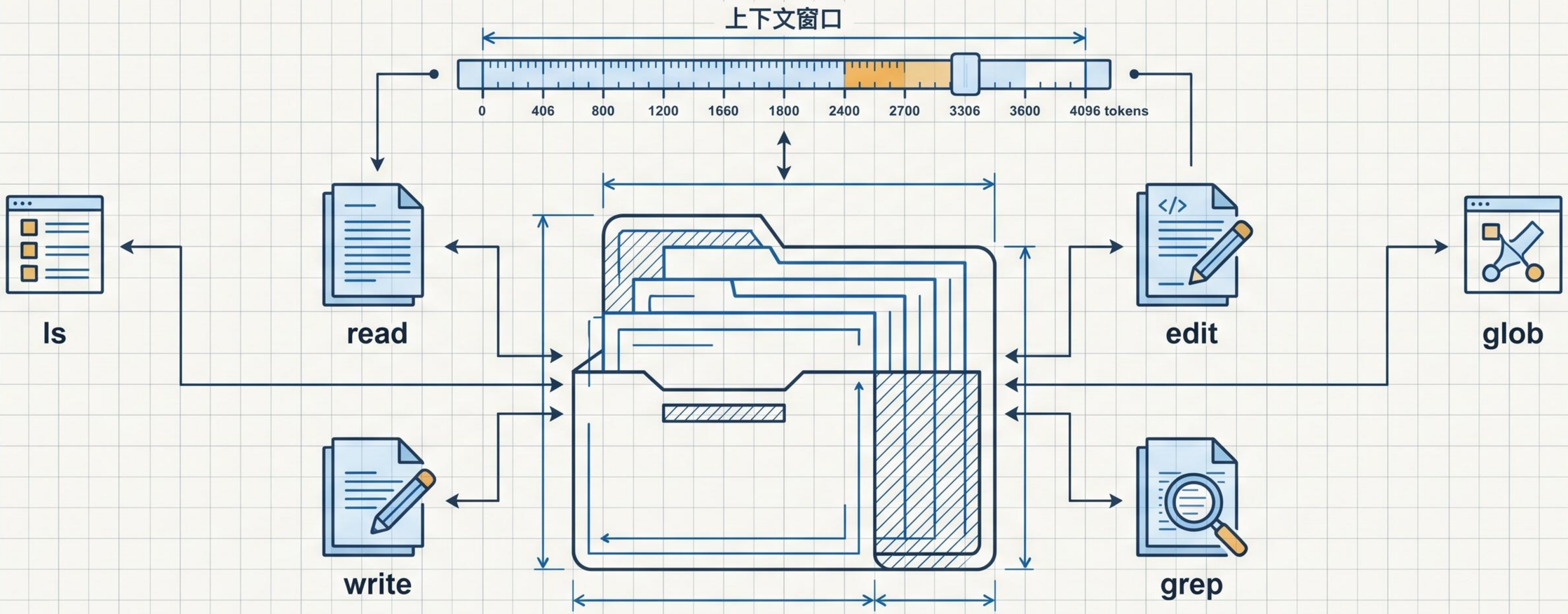
Deep Agents 实战

第 4 讲：虚拟文件系统与上下文管理

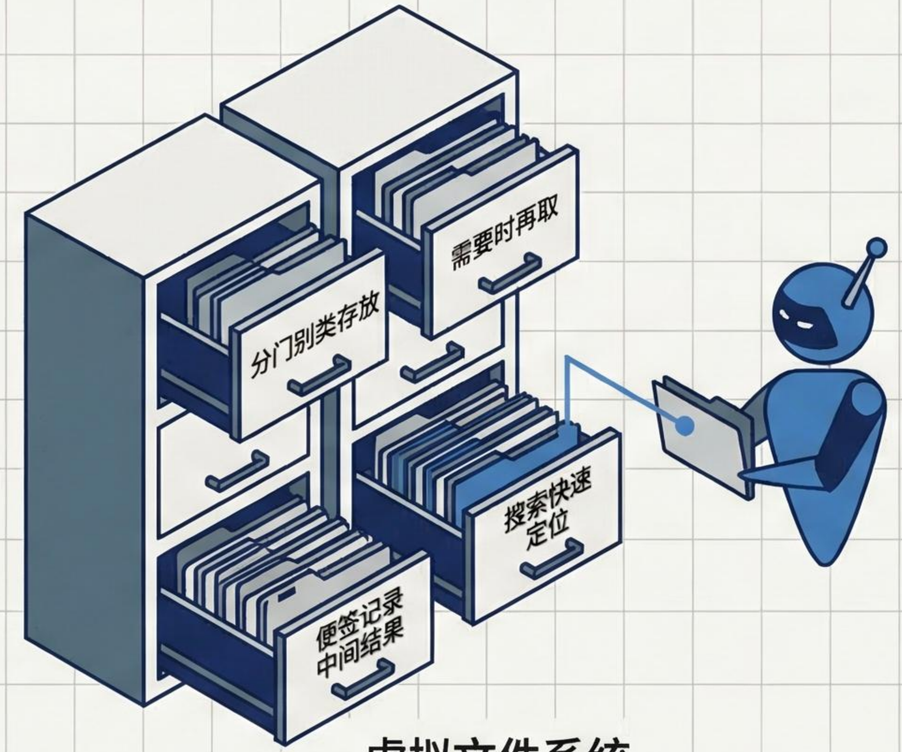


为什么用文件系统管理上下文?

从"全塞进 Prompt"到"按需读取"

传统 Prompt Stuffing

- 传统做法：所有信息直接塞进 prompt，对话历史不断膨胀



虚拟文件系统

- Deep Agents 方案：给 Agent 一个文件系统
- 像人一样工作：分门别类存放、需要时再取、搜索快速定位、便签记录中间结果

六大文件系统工具

覆盖文件操作的完整生命周期

ls

列出目录中的文件和元信息

（打开文件夹看看有什么）

read_file

读取文件内容，支持分片

（翻开某份资料阅读）

write_file

创建新文件

（写一份新的备忘录）

edit_file

精确字符串替换

（用红笔修改文档）

glob

按模式匹配查找文件

（在文件柜中按标签找）

grep

搜索文件内容，支持正则

（全文检索）

read_file 详解

分片读取 + 原生多模态——不只是读文本

Agent

分片读取

- offset + limit 参数，默认前 2000 行
- 按需读取指定范围，避免大文件撑爆上下文

- offset + limit 参数，默认前 2000 行
- 按需读取指定范围，避免大文件撑爆上下文

- 示例：read_file("report.md", offset=100, limit=50)

原生多模态支持（4 类格式）

图片	.png .jpg .jpeg .gif .webp .heic .heif
视频	.mp4 .mpeg .mov .avi .flv .webm .wmv .3gpp
音频	.wav .mp3 .aiff .aac .ogg .flac
文档	.pdf .ppt .pptx

read_file(.mp4) → 多模态内容块 → Agent 直接处理

grep 详解：三种输出模式

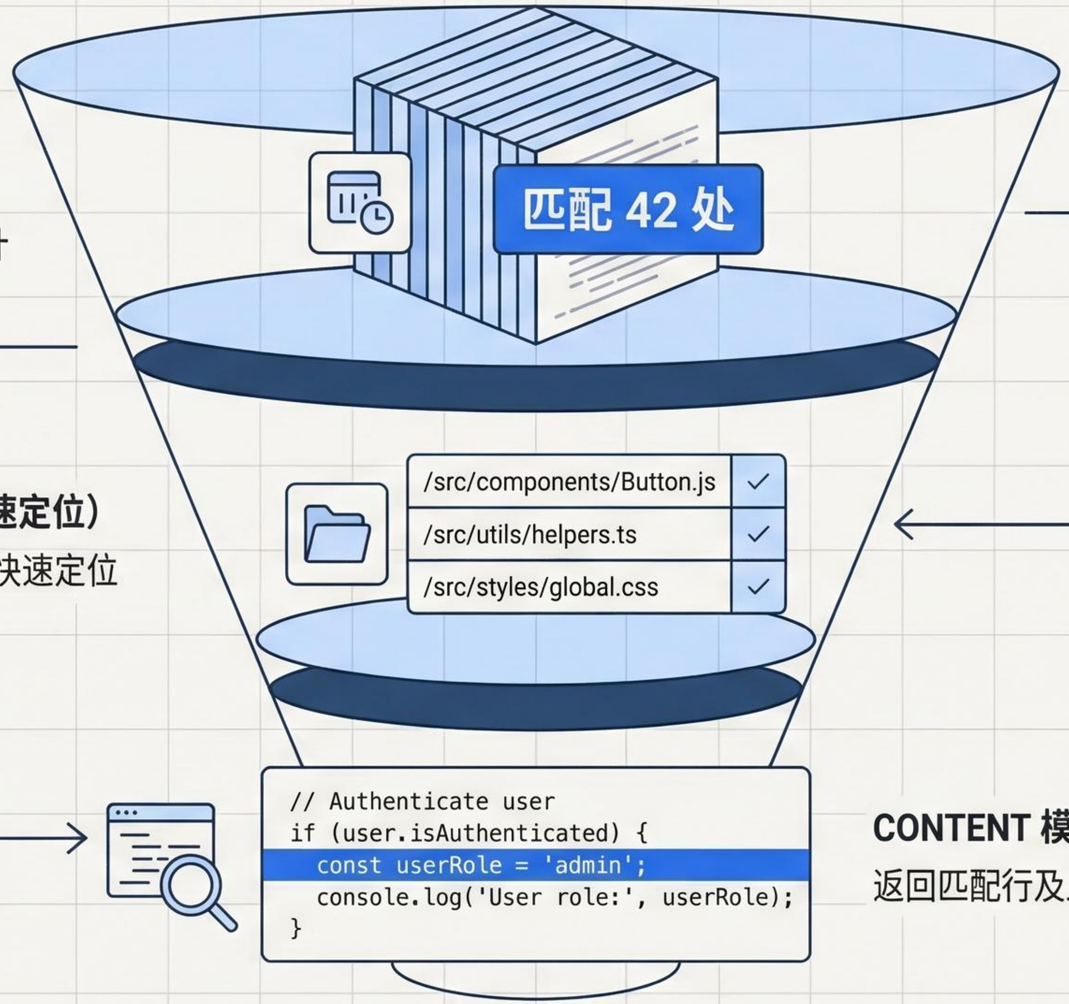
从快速定位到深入查看，逐层聚焦

COUNT 模式（概览统计）

返回匹配数量，提供概览统计

FILES_WITH_MATCHES 模式（快速定位）

只返回匹配的文件路径，实现快速定位



CONTENT 模式（深入查看）

返回匹配行及上下文（可指定context行数），进行深入查看

上下文自动管理

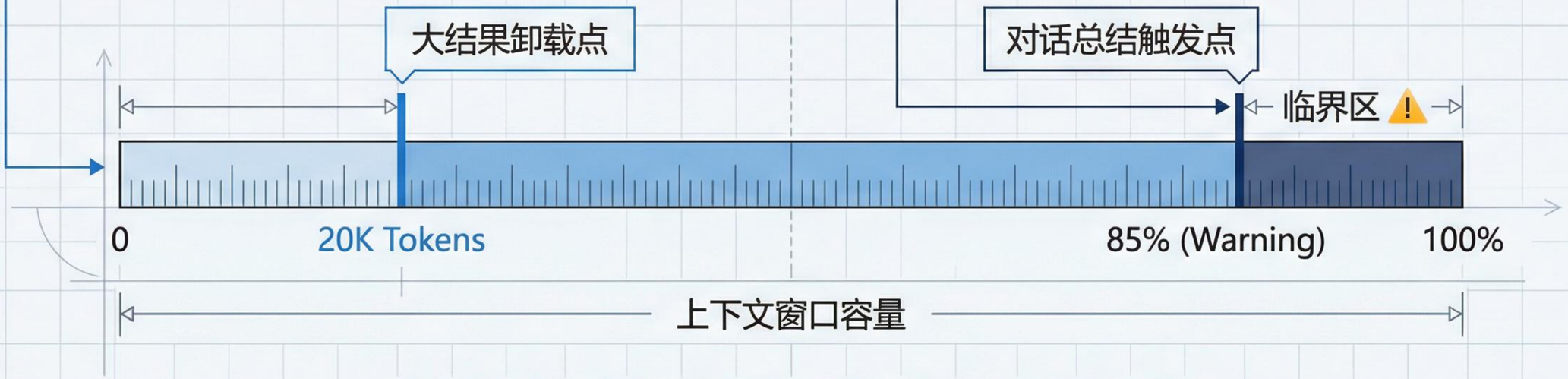
虚拟文件系统的最大价值——两道防线守护上下文窗口

第一道防线：大结果自动卸载

工具输出 > 20K tokens
自动卸载至虚拟文件系统

第二道防线：对话历史自动总结

上下文 > 85% 窗口
触发摘要，保持关键信息



核心问题：模型的上下文窗口有限，信息不断增长.
Agent 无需手动管理，一切自动完成。

第一道防线：大结果自动卸载

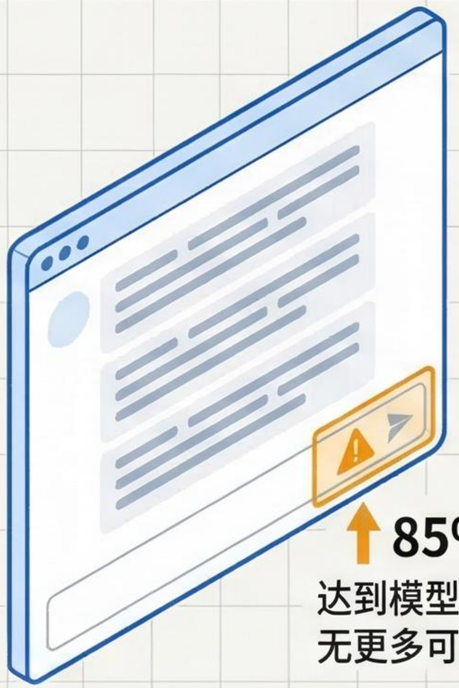
工具输出超过 20K tokens 时自动触发

- 触发条件：工具调用输入或输出 > 20,000 tokens（可配置 tool_token_limit_before_evict）
- 自动流程：
  1. 完整内容写入虚拟文件系统
  2. 对话历史中替换为文件路径引用 + 前 10 行预览
  3. Agent 需要时可通过 read_file / grep 按需读回

第二道防线：对话历史自动总结

上下文达到窗口 85% 时自动触发

触发条件



85%

达到模型窗口的 85%，
无更多可卸载内容

自动流程

1. 生成摘要（LLM）

用 LLM 生成结构化摘要
（意图、产出物、下一步）

2. 存档完整对话

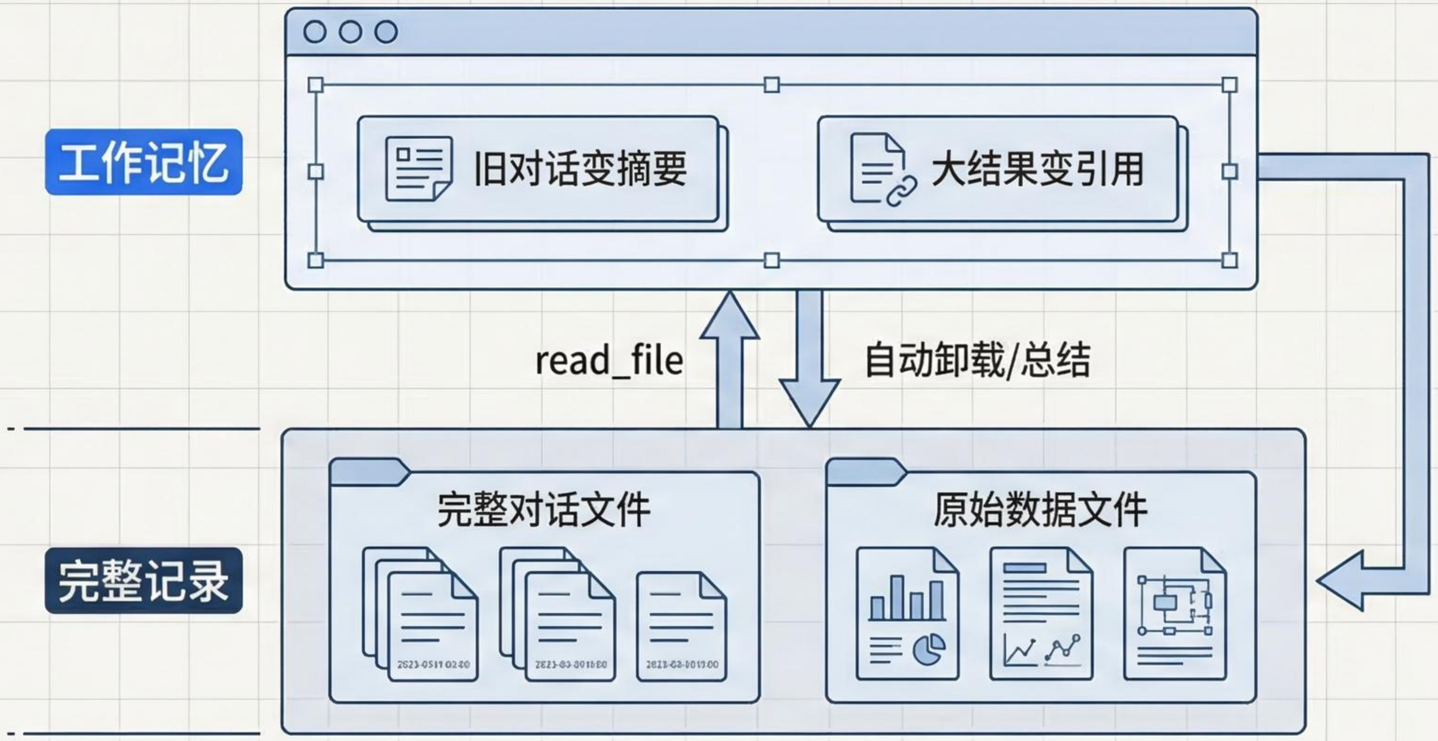
完整原始对话写入文件系统保存

结果展示

文件系统完整记录

3. 用摘要替换对话历史中的旧消息，
对话窗口恢复宽裕，摘要在最前面，
文件系统中保留完整记录

双重保障
精炼的工作记忆 + 可回溯的完整记录



- 工作记忆（对话历史）：始终保持精炼——大结果变引用，旧对话变摘要
- 完整记录（文件系统）：一切都可回溯——原始数据、完整对话随时读回
- **核心理念：Agent 既有高效的"短期记忆"，又有完整的"长期档案"**

本讲回顾

- 虚拟文件系统 = 给 Agent 一个结构化的"工作空间"
- 六大工具覆盖文件操作完整生命周期
- read_file 支持分片读取、行号定位、多模态
- grep 三种模式：从概览到细节逐层聚焦
- 两道防线守护上下文：大结果卸载 + 对话总结
- 双重保障：精炼工作记忆 + 可回溯完整记录

下一讲：可插拔的存储后端

StateBackend — FilesystemBackend — StoreBackend — 与 CompositeBackend