

公文智能核查实践

——基于 doc-fact-check 工具的某校规划两版本核查案例

核心议题：公文事实性表述与依据材料的一致性核对

01 项目背景与挑战

核心问题：公文事实性表述与依据材料的一致性

在党政机关、企事业单位的日常工作中，各类公文（汇报材料、政策文件、报告、规划文本等）通常包含大量事实性表述——
数据指标 · 成果统计 · 案例引用 · 政策表述

这些表述必须严格以支撑材料（原始数据、调研报告、规划文件等）为依据。一旦两者出现偏差，可能引发材料失实、引述不严谨等公文质量问题，
进而影响决策参考、削弱文件公信力、产生合规风险。

常见需要核对的场景

					
汇报材料	规划文件	讲话稿	申报材料	制度文件	决策报告
数据指标、成果统计	目标设定、任务部署	政策表述、案例引用	项目数据、佐证依据	条款引用、历史沿革	调研结论、对比数据

本案例

某校教育事业发展规划 · 校内版 vs 校外版 · 同源材料的不同呈现版本

02 本案例 · 某校规划双版本

同源材料的不同呈现版本 · 面向不同受众需要

校内版本

完整参考材料 / 内部使用

- 144 段 · 完整版
- 48 个重点任务完整展开
- 发展基础段多领域详细数据
- 实施保障段全覆盖
- 作为 " 依据材料 "



校外版本

压缩呈现版 / 外部使用

- 124 段 · 压缩版
- 48 个重点任务保留框架
- 发展基础段压缩为综合表述
- 实施保障段内容精简
- 作为 " 公文文档 "

核对的核心问题

- ① 压缩过程中是否丢失了关键内容？
- ② 数字指标、核心表述在两个版本中是否一致？
- ③ 校内版本 48 个重点任务之间是否存在内部不一致？

03 三轮复核方法论

" 自动初查 → 人工复检 → 反向验证 " 三轮闭环

1

第一轮
自动核对

关键词全文检索 + 子命题分解 + 实体覆盖 + 共现验证

全流程自动运行，生成 Excel 核对清单

2

第二轮
人工复检

对 " 未找到 " 项做分段搜索、近义词替换、全量浏览

从大量 " 未找到 " 中找回真实匹配

3

第三轮
反向验证

对 " 已确认 " 项做反向检查，揪出自动误判

上下文语义、专有名词、数值精确性验证

05 关于 Skill · 工具的封装与复用

Tool Encapsulation & Reusability

什么是 Skill ?

Skill (技能) 是一组封装好的、可重用的指令集合 , 专门用于完成某一类特定任务。

一个 skill 通常包含三部分 :

 **SKILL.md** · 使用说明与方法论

 **scripts/** · 可执行脚本代码

 **其他资源** · 模板、配置、参考材料

优势 :

- 可跨会话、跨设备复用
- 可分享、可版本化管理
- 使用时只需说 " 使用某某 skill "

如何加载 doc-fact-check ?

步骤一 · 克隆仓库

```
git clone https://github.com/hongyuatcufe/doc-fact-check
```

步骤二 · 安装依赖

```
pip install openpyxl # 依赖: pandoc
```

步骤三 · 加载并使用

1. 将项目放到 skill 目录
`~/.agents/skills/doc-fact-check/`
2. 在对话中说 " 使用 doc-fact-check "
3. 指定目标文档与参考文档,
工具将自动输出 Excel 核对清单

o6 doc-fact-check 工具 · 原理

基于关键词匹配的全文检索 · 子命题分解 · 实体覆盖验证

核心机制

① 文本提取

使用 pandoc 将 .docx/.doc 转换为纯文本，
兼容 macOS textutil 自动回退。

② 表述分解

复合表述按分隔符拆解为独立子命题，
"A 入选 X，B 获批 Y" → 分别验证。

③ 关键词检索

在参考文档 txt 中全文搜索关键词、
数字、实体名称，匹配原文片段。

④ 实体覆盖检查

提取引号 / 书名号内专有名词，
检查 300 字窗口内的实体 - 语境共现。

v3 核心特性

01

复合表述分解

长句拆分为原子子命题

02

实体覆盖检查

专有名词 / 项目名独立验证

03

共现语境验证

实体与关键词 300 字窗口

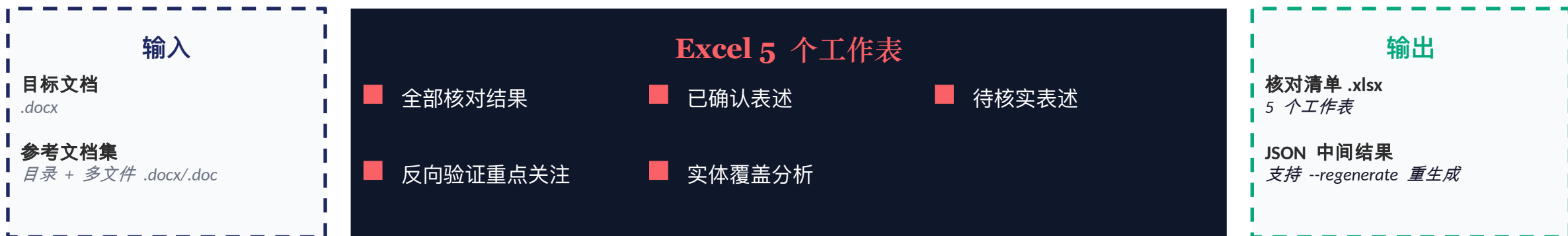
04

子命题独立评分

任一不匹配则整体降级

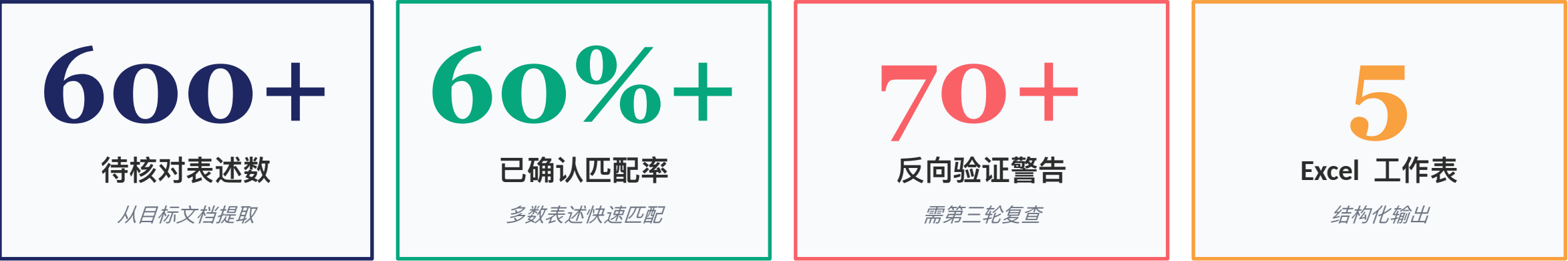
07 doc-fact-check 工具 · 架构

五阶段流水线架构 · 输入到输出完整闭环

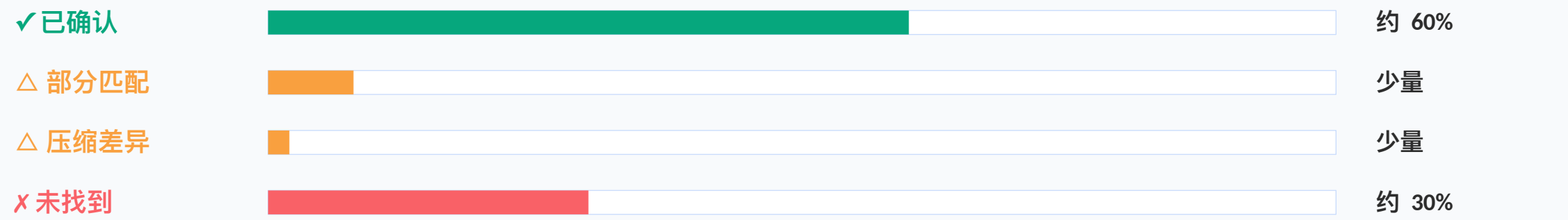


o8 doc-fact-check 工具 · 效果

本案例实测数据 · 工具能力直观展示



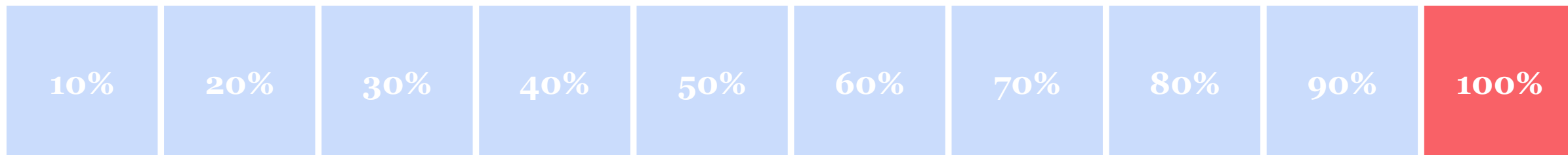
首轮自动核对结果分布



09 第一轮 · 自动核对结果

工具在数秒内完成全量比对 · 生成 Excel 核对清单

全文检索进度（实时）



关键类别指标全部一致 ✓

- 留学生培养规模指标
- 图书文献资源指标
- 对外合作办学项目数
- 学生升学质量指标
- 重大科研项目立项数
- 国际学术论坛举办数

(具体数据详见参考文档)

需要第二轮人工复核

少量 "数据不一致" — 需逐条核验是否为压缩导致

少量 "部分匹配" — 关键词被分散匹配

数十条 "反向验证警告" — 需第三轮复查

10 第二轮 · 人工复核发现

从 " 未找到 " 项中找回真实匹配 · 区分压缩差异与真实偏差

人工复核策略

01

分段搜索

拆长句为独立关键词

02

核心名词

去除修饰词保留主体

03

数字直搜

直接搜数字部分

04

近义词替换

" 课题 " ↔ " 项目 " 、 " 平台 " ↔ " 中心 "

关键发现

✓ 所有关键类别的数字指标在依据材料中均找到一致出处

⚠ " 数据不一致 " 项经人工复核均为压缩差异，非真实冲突

⚠ 个别任务段存在细微差异（如量词缺失、技术名词写法）

11 第三轮 · 反向验证

对 " 已确认 " 项做反向检查 · 揪出自动误判

6 项反向验证检查

① 上下文语义一致性

原文片段是否真的说同一件事

② 专有名词精确性

项目 / 机构名逐字比对

③ 数值精确性

每个数字独立验证

④ 增长率独立验证

基数对 ≠ 增长率对

⑤ 复合数据完整性

"A、B、C" 每个子项独立

⑥ 口径一致性

统计范围 / 单位 / 分级口径

实际验证结果

✓ 目标文档独有内容经实体提取确认在依据材料中存在

△ 发现部分增长率虽匹配，但上下文需确认是否属于同一语境

△ 学科提质相关关键词在依据材料存在但分散，需综合判定

✓ 所有反向验证警告经人工复核后确认不影响最终结论

12 48 个任务的内部一致性盘查

从文档本身出发 · 数字 / 术语 / 逻辑三维验证



10 个章节 · 48 个任务的章节分布



数字：该章节包含的任务数量

🔍 文档中 7 处表述问题：文件引用不统一（1）+"经济学"概念层次不清（1）+"领军人才"重复（1）+ 任务衔接（2）+ 目标对应（2）

13 修订实施 · 16 个任务修订

补回压缩中丢失的关键内容 · 严格遵循原文风格

● 必须修订

3 个

部分核心任务段

● 建议修订

4 个

学科建设与国际合作相关任务

● 细节优化

9 个

党建、基础设施等任务

重点修订 · 人才培养模式改革相关任务

补回若干关键内容：

- ① 人才培养创新工程相关表述
- ② 产教深度融合的培养模式
- ③ 激发研究生科研创新活力的相关表述
- ④ 海外攻博与联合培养等招生专项相关表述
- ⑤ 招生专项种类与覆盖面相关表述
- ⑥ 育人模式优化相关表述

14 最终效果 · 工具赋能核查工作

" 机器高效匹配 + 人工智慧判断 " 的协同模式



效率提升

- 数百条表述核对
- 在数秒内完成
- 人工核对需 1 人 / 天



准确保障

- 三轮复核闭环
- 复合表述分解
- 实体覆盖验证



可追溯性

- 原文片段保留
- JSON 中间结果
- Excel 5 工作表

最终结论

- ✓ 目标文档与依据材料在所有核心内容上保持一致
- ✓ 未发现重大数据不一致或事实偏差 · 可放心使用
- ✓ 16 个任务修订完成

15 经验启示

从本次实践中提炼的方法论

01

工具是手段，不是目的

doc-fact-check 提供数秒内完成大量核对的能力，但反向验证警告必须由人工判断。机器匹配 + 人工复核，二者缺一不可。

02

"压缩差异" ≠ "事实错误"

当参考文档为压缩版时，"未找到"项是预期的。区分"压缩差异"与"事实错误"是核查工作的核心难点。

03

格式是细节，但不可忽视

修订时可能无意破坏原文档格式。建议同步检查文档结构，确保与原文档风格一致。

项目地址：<https://github.com/hongyuatcufe/doc-fact-check>