Week 7｜课时 5｜解析质量与切片质检：把 Week8 索引准入做成工程门禁

Table of contents

Week8 不能消费“看起来差不多”的 chunks

这一讲收口 Week7：

解析质量、chunk 质量和 evidence coverage 必须变成门禁，而不是靠人工感觉放行。

进入实验 返回 Week7 总览

下载讲义

提供适合离线阅读的 PDF 版和适合批注整理的 Word 版。

PDF 版 · 打印 / 离线阅读 Word 版 · 批注 / 二次整理

这节课解决什么问题

Week7 产出 chunks 之后，不能直接进入 Week8 索引。

必须先回答：

- metadata 是否完整
- 每个 chunk 是否有 anchor

- page_no / bbox 覆盖率是否符合文档类型
- 空 chunk、孤儿 chunk、重复 overlap 是否过高
- 是否有明显 PII leakage risk
- 样本数不足时如何说明

参考学习时间

45–55 分钟

学完这一讲，你应该能做到什么

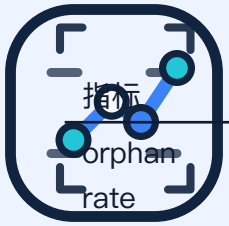
1. 设计 chunk_quality_report.md 的核心指标。
2. 区分 hard gate、warning、not applicable。
3. 按 content type 判断 page/bbox coverage。
4. 写出 week8_ready_gate.json 的准入结论。

本课产出

- docs/blueprints/week07/quality-gate-spec.md
- reports/week07/chunk_quality_report.md
- reports/week07/week8_ready_gate.json

Quality Gate 仪表板

指标	计算方式	Student Core 阈值	Gate
metadata completeness	必需字段完整 chunk 占比	100%	hard
anchor coverage	有 evidence anchor 的 chunk 占比	100%	hard
PDF page reference rate	PDF chunk 有 page_no 的占比	100%	hard
PDF bbox coverage	PDF chunk 有 bbox 或 missing reason 的占比	100%	warn / hard
empty chunk rate	空内容 chunk 占比	0%	hard



指标	计算方式	Student Core 阈值	Gate
orphan chunk rate	无 section / doc 关联 chunk 占比	0%	hard
overlap duplicate rate	overlap 重复噪声比例	需说明	warn
PII leakage risk	heuristic 命中比例	需复核	warn / quarantine

1. content-type aware gate

不能用同一条 bbox 规则压所有文档。

content type	page_no	bbox	gate 解释
PDF	必须有	缺失要写原因	citation 基本条件
HTML	可为空	可为空	需要 URL / section_path
Markdown	可为空	可为空	需要 source_uri / heading path
DOCX	视 parser capability	视 parser capability	hierarchy 优先

报告必须区分：

- overall_bbox_coverage
- pdf_bbox_coverage
- non_paged_bbox_not_applicable_count

否则会把 HTML / Markdown 误判成解析失败。

2. 样本不足也要报告

作业要求至少 50 份文档的抽样质检流程。如果本地样本不足，不是直接跳过，而是写清：

- 当前样本数
- 缺口数量
- 抽样规则是否已实现
- 扩展到 50+ 文档时要怎么跑

这对应 sample_shortfall。

3. Week8 ready gate 最小输出

```
{
  "status": "WARN",
  "parse_strategy_version": "docling_v1_no_ocr",
  "chunk_strategy_version": "section_aware_v1",
  "total_chunks": 128,
  "ready_chunks": 124,
  "blocked_chunks": 4,
  "hard_failures": [],
  "warnings": ["sample_shortfall", "pdf_bbox_coverage_below_target"],
  "week8_consumption_rule": "consume pass/warn chunks with evidence_anchor only"
}
```

4. PII leakage risk 只是 heuristic

Week7 可以做明显模式检查：

- email
- phone
- token-like string
- secret-like key

但不能宣称已经完成完整合规扫描。Week12 / Week14 才会进入更完整 tracing 和 governance。

本课最重要判断

质量报告不是附录，而是 Week8 的准入门禁。没有 quality gate，就不应该建索引。

自检清单

- 我能设计 content-type aware page/bbox gate。
- 我能解释 sample_shortfall 不是失败，而是可交接限制。
- 我知道 PII risk 在 Week7 只是 heuristic。
- 我能写出 Week8 ready gate 的消费规则。

最小行动命令

```
python -m pipelines.parse_normalize.quality \
  --input-dir artifacts/week07 \
  --sample-size 50 \
  --out reports/week07/chunk_quality_report.md \
  --gate-out reports/week07/week8_ready_gate.json
```