Week 7｜课时 4｜Evidence Anchor：让每个 chunk 可以回指原文位置

Table of contents

Citation 的事实来源必须在数据阶段生成

这一讲把 Week7 的证据链收紧：

每个 chunk 至少要有一个 evidence anchor，告诉 Week8 它来自哪个文件、哪个版本、哪一页、哪个 section、哪个 bbox。

进入课时 5 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

很多系统把 citation 放到生成阶段，让 LLM 根据上下文“拼一个来源”。这会留下两个根本问题：

- 引用可能不是检索结果真实来源

- 出问题时无法回到原始文档定位

Week7 的做法是：

citation 只能消费 evidence anchor；LLM 不允许发明证据。

参考学习时间

45–55 分钟

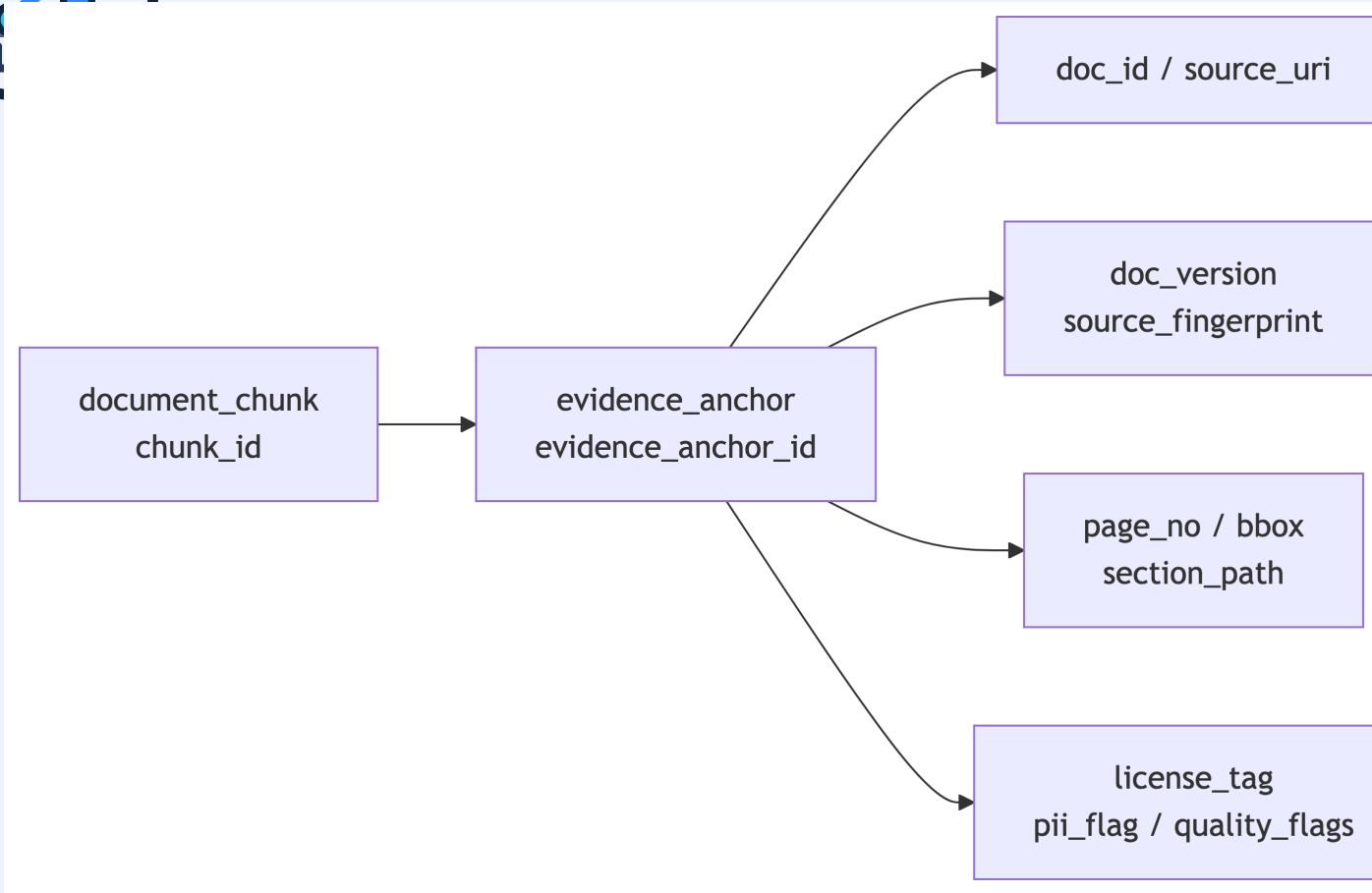
学完这一讲，你应该能做到什么

1. 区分 chunk、section 和 evidence anchor。
2. 设计 evidence anchor 的最小字段。
3. 解释为什么 anchor 需要 source_fingerprint 和 doc_version。
4. 判断 page/bbox 缺失时应该 fail、warn 还是 not applicable。

本课产出

- docs/blueprints/week07/evidence-anchor-contract.md
- contracts/data/evidence_anchor.schema.json

evidence anchor 映射图



1. Anchor 不是装饰字段

Evidence anchor 至少要支撑四类动作：

动作	Anchor 提供什么
citation	文件、页码、section、bbox
bad case replay	原始对象、版本、指纹、策略版本
quality review	bbox 缺失原因、质量 flags
policy filter	license、visibility、PII 状态

没有 anchor，Week8 检索出来的 chunk 就只是“像证据的文本”。

2. Anchor contract 最小字段

```
{
  "evidence_anchor_id": "stable_sha256_prefix",
  "chunk_id": "document_chunk_id",
  "anchor_type": "text",
  "doc_id": "workspace_help_center_v1",
  "source_uri": "s3://omni-raw-documents/workspace-help/source.pdf",
  "raw_object_uri": "s3://omni-raw-documents/workspace-help/source.pdf",
  "parsed_object_uri": "s3://omni-parsed/week07/parsed_doc.json",
  "doc_version": "v1",
  "source_fingerprint": "sha256...",
  "page_no": 4,
  "bbox": {"x0": 0.13, "y0": 0.22, "x1": 0.84, "y1": 0.31},
  "bbox_missing_reason": null,
  "section_path": ["Setup", "SSO"],
  "license_tag": "internal_training",
  "quality_flags": []
}
```

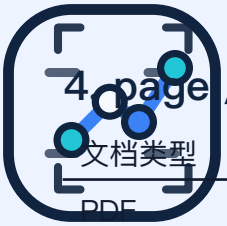
3. 稳定 anchor ID

推荐公式：

```
evidence_anchor_id = sha256(
  chunk_id
  + anchor_type
  + source_fingerprint
  + doc_version
  + page_no
  + bbox
  + section_path
)[:32]
```

这样能保证：

- 同一 chunk 重跑后 anchor ID 不漂移
- bbox 或 section_path 变化时，anchor ID 能反映事实变化
- Week8 citation 可以稳定引用同一证据对象



4. page / bbox 缺失怎么处理

文档类型	page_no	bbox	Week7 gate
PDF	必须有	缺失 warning，并要求 bbox_missing_reason	page 缺失 hard fail
HTML / Markdown	可为空	可为空	需要 section_path 和 source_uri
DOCX	视 parser capability	视 parser capability	hierarchy 优先，bbox 作为 warning
scanned PDF	取决于 OCR	取决于 OCR	Student Core 不强制 OCR，报告中标明限制

本课最重要判断

Evidence anchor 是 Week8 citation、Week11 eval、Week12 tracing 和 Week14 governance 的根。它必须在数据阶段生成。

自检清单

- 我能说明 citation 为什么不能由 LLM 临时发明。
- 我知道 evidence anchor 最少要保留哪些字段。
- 我知道 PDF、HTML、DOCX 的 page/bbox gate 不一样。
- 我能解释 source_fingerprint 对 bad case replay 的作用。

最小行动命令

```
python -m pipelines.parse_normalize.evidence \
  --sections artifacts/week07/sections.json \
  --chunks artifacts/week07/chunks.json \
  --out artifacts/week07/evidence_anchors.json
```