UIST 2026 Paper Plan

Extract-then-Assemble: An Evidence-First Paradigm

v6.7 — 溯源 Pipeline 技术细节澄清版

2026 年 2 月 12 日

v6.7 更新概要

[v6.7 定位] 澄清溯源 Pipeline Step 3、4 的技术实现。核心信息：structured generation 不只是实现选择，它是溯源引擎区别于通用 RAG 系统的关键设计决策——显式标注优先、语义检索兜底。

v6.6 → v6.7 三个修补点：

- **#16** Step 3 展开为 Constrained Petition Generation：Structured JSON 输出，每句携带 snippet_ids，从源头限制幻觉 + 提供确定性溯源锚点
- **#17** Step 4 重新定义为 Hybrid Retrieval：显式标注为主，语义检索为 fallback（处理部分标注 + 手动编辑后标注失效）
- **#18** 新增设计 rationale：deterministic anchor + probabilistic fallback，直接支撑 high-stakes writing 叙事

一、论文定位（v6.3）

1.1 核心定位

本文挑战 AI-assisted writing 中 “generate-then-verify” 的主流协作模式，提出 “extract-then-assemble” 替代范式。核心洞察：在 evidence-based professional writing 中，让人类构建 evidence-to-argument 映射能同时实现更高准确性和更强信任感——因为构建过程本身就是验证过程（mapping as verification）。

1.2 核心故事线（三拍结构）

第一拍 — 领域假设：AI writing 普遍采用 generate-then-verify。隐含假设：人能有效审查 AI 输出。

第二拍 — 假设失效：Formative study 发现律师无法有效审查 AI 映射。 **[v6.4] Probe task 提供直接证据。**

第三拍 — 替代范式：Extract-then-assemble。人从 verifier 变为 assembler。Mapping as verification。

1.3 标题候选

- ★ *"Extract-then-Assemble: An Evidence-First Paradigm for Human-AI Collaborative Legal Writing"*
- ★ *"Beyond Generate-then-Verify: Evidence-First Authoring for High-Stakes Professional Writing"*

1.4 研究问题

- **RQ1:** Extract-then-assemble 如何影响律师对证据-论点关系的理解和信任?
- **RQ2:** 跨粒度溯源是否帮助发现事实性错误?
- **RQ3:** Human-assembled mapping vs read-only AI mapping 对 petition 质量的影响?
- **RQ4:** Argument 中间层是否帮助组织多证据聚合论证?

二、贡献排序（v6.3）

#	贡献	内容	类型
C1	**Extract-then-assemble 范式**	通过 formative study 识别 generate-then-verify 失效，提出替代范式。核心洞察 “mapping as verification”。	**Empirical + Conceptual**
C2	**[SystemName]**	实例化 C1。核心技术组件为 cross-granularity provenance engine，通过语义检索将生成文本回溯到源文档 BBox 级别，实现三层溯源 (Sentence→Snippet→BBox)。系统还包含三层映射界面和空间化 Writing Canvas。	**Artifact**
C3	**实证验证**	N 名律师 within-subjects，human-assembled vs read-only。人控映射提高错误检测率和信任感。	**Empirical**

三、Introduction 段落级结构（~800 words）

P1 — AI Writing 的主流协作模式 (~150w)

要点：AI writing 工具普遍采用 generate-then-verify。创意写作中有效，但在 evidence-based professional writing 中，每个断言必须有可追溯证据。

开头句：*"AI-powered writing tools increasingly adopt a common collaboration pattern: AI generates content or structure, and humans verify the output. While this 'generate-then-verify' paradigm works well for creative and general-purpose writing, we argue it is fundamentally mismatched for evidence-based professional writing, where every claim must be traceable to source documents and errors carry high stakes."*

P2 — 失效模式 1：信息不透明 (~120w)

要点：End-to-end 工具产出完整文本，不暴露推理过程。

关键句：*"The first failure mode of generate-then-verify in evidence-based writing is opacity: current tools produce text without exposing which evidence supports each claim, depriving the human verifier of the information needed to perform meaningful review."*

[v6.4 #2] P3 — 失效模式 2：审查失效 (~120w)

~~旧版：...lawyers tend to accept plausible-looking AI mappings without deep verification — a form of automation bias...~~

[v6.4 #2] 新版要点：即使暴露了 AI 映射，verifier 仍面临困境：AI 映射本身可能错误，而审查的认知负荷极高。我们的 formative study 发现律师对"看起来合理"的 AI 映射审查不足（insufficient scrutiny），这一现象与人机交互文献中的 automation bias 一致（Parasuraman & Riley, 1997）。

关键句：*"The second failure mode is more insidious: even when AI exposes its evidence mappings, these mappings are unreliable, and our formative study (Section 2) reveals that lawyers exhibit insufficient scrutiny of plausible-looking AI mappings — a pattern consistent with the automation bias phenomenon documented in human-automation interaction research (Parasuraman & Riley, 1997)."*

为什么这样改：① 不直接称之为 automation bias，而说 "consistent with"，更安全。② 引用经典文献作为理论锚点。③ 引用 "Section 2"（因为 Formative Study 现在在 Related Work 之前）。

P4 — Extract-then-Assemble (~180w)

要点：AI 低层提取，人高层构建。三层结构。Mapping as verification。EB-1A 实例化。

关键句：*"We propose an alternative paradigm — extract-then-assemble — that reverses the human-AI division of labor... mapping is verification."*

P5 — Contributions (~150w)

1. *An extract-then-assemble collaboration paradigm* for evidence-based professional writing, grounded in formative interviews with immigration lawyers. We identify the failure modes of generate-then-verify and articulate a design insight — *mapping as verification* (Section 2).

2. *[SystemName]*, a system that instantiates this paradigm with a three-layer mapping interface, a spatial Writing Canvas, and a cross-granularity provenance engine (Section 3).

3. *Empirical evidence* from a within-subjects study with N immigration lawyers (Section 4).

[v6.4 #3] 注意 Section 编号已更新：Formative Study = Sec 2, System = Sec 3, Related Work = Sec 4, Evaluation = Sec 5

P6 — 结构导读 (~80w)

[v6.4 #3] *"We first report our formative study that yielded four design principles and identified the failure modes of generate-then-verify (Section 2), review related work (Section 3), describe [SystemName]'s design (Section 4), present our user study (Section 5), and discuss implications and limitations (Section 6)."*

[v6.4 #3] 四、Formative Study（论文中为 Section 2，在 Related Work 之前）

[v6.4 #3 章节顺序调整] Formative Study 从 Section 3 提前到 Section 2。原因：

- Introduction P3 引用了访谈数据，读者希望立即看到证据
- Related Work 中用 extract-then-assemble 框架定位所有工具，这个框架本身是 Formative Study 的产出——先读 Formative Study 再读 Related Work 逻辑更顺

论文中的新章节顺序：

旧编号	旧标题	新编号	新标题	变更
Sec 1	Introduction	Sec 1	Introduction	不变
Sec 3	Formative Study	**Sec 2**	**Formative Study**	**提前**
Sec 2	Related Work	**Sec 3**	**Related Work**	**后移**
Sec 4	System Design	Sec 4	System Design	不变
Sec 5	User Study	Sec 5	User Study	不变
Sec 6	Discussion	Sec 6	Discussion	不变

核心原则：所有设计决策基于自主访谈数据。Swoopes et al. (2025) 作为 corroborating evidence。

[v6.4 #1 新增职责] Formative Study 现在承担双重任务：① 产出设计原则 DP1-DP4；② 提供 generate-then-verify 失效的实证证据。

4.1 参与者

来源：律所合作伙伴 | **人数：**3–5 名 | **标准：**≥1 年 petition 经验 + EB-1A/NIW

[v6.4 #1] 4.2 访谈协议（含 Probe Task）

时长：60–75min 半结构化访谈，四个模块：

Part A — 当前工作流 (15min)

- 你现在怎么写 petition？证据收集和组织流程是什么？
- 哪个环节最耗时/最容易出错？

Part B — AI 工具态度 (10min)

- 用过哪些 AI 辅助工具？信任度如何？
- 在什么情况下你不会信任 AI 的判断？

[v6.4 #1 全新] Part B+ — Probe Task: 审查 AI 映射 (20min)

目的：直接测量律师能否有效审查 AI 的 evidence-to-standard 映射，产出 generate-then-verify 失效

的实证证据。

材料：一份纸质或屏幕上的 AI 生成 evidence-to-standard 映射结果（基于真实 EB-1A 案例），预埋 5 个错误。

预埋错误类型：

#	错误类型	具体例子	为什么难发现
1	错误标准	引用数据映射到 “Original Contributions” 而非 “Scholarly Articles”	两个标准都和学术成就相关，表面看都合理
2	错误标准	推荐信中的领导力评价映射到 “Judging” 而非 “Leading Role”	“评判”和“领导”语义重叠
3	遗漏	专利文件中的关键技术描述未被提取	需要领域知识才能判断其重要性
4	遗漏	媒体报道中的行业影响力描述未被提取	淹没在长文本中，容易被忽略
5	冗余	无关的地址信息被映射到 “Scholarly Articles”	在大量映射中容易被忽视

流程：

- “这是 AI 为一位 ML 研究者生成的 EB-1A 证据映射。请您审查这些映射，标记您认为有问题的地方。”
- 记录：(a) 发现了几个错误（/5）；(b) 花了多长时间；(c) 哪些错误没发现及原因
- 审查完成后揭示错误，追问：“为什么您没有注意到这个？”“如果您能自己拖拽调整这些映射，体验会不同吗？”

预期产出：

- **定量：**“5 个错误只发现 2 个” — generate-then-verify 失效的直接证据
- **质性：**律师描述审查困难的原因 — 认知负荷、缺乏审查策略、过度信任
- **洞察：**“如果能自己拖”的回答 — extract-then-assemble 的需求验证

为什么比回忆性问题强：Probe task 是 in situ 的行为数据，不依赖律师的自我报告。“5 个错误只发现 2 个”比“你有没有被 AI 误导过”有说服力得多。

[v6.5 #6 新增] 方法论框定（必须写入论文）：

"The probe task is not intended as a controlled experiment but as a structured interview probe designed to elicit concrete examples of verification difficulties and ground participants' subsequent reflections in observed behavior rather than self-report. We report probe task outcomes descriptively to characterize the verification challenge, not to establish statistical significance."

为什么必须框定：Probe task 在 3-5 人的 formative study 里做，如果不框定，reviewer 会用实验标准衡量（样本量太小、无对照组等）。定义为 interview technique 就只需要满足访谈的标准。

Part C — 概念验证 (15min)

- 展示拖拽映射的低保真原型/线框图，收集反馈

4.3 分析方法

Reflexive thematic analysis (Braun & Clarke, 2006)。两名研究者独立编码，讨论合并主题。

4.4 预期设计原则

DP1 — 律师需要控制映射过程： → 拖拽映射

DP2 — 源文档必须可追溯： → 跨粒度溯源

DP3 — 多证据聚合是常见模式： → Argument 中间层

DP4 — 视觉化显性化证据缺口： → Focus 模式

4.5 与 Swoopes et al. 的关系

Swoopes et al. 是 corroborating evidence。两个独立来源指向同一方向 = 更强说服力。

[v6.5 #9] 4.6 小结桥接段（必须写入论文）

"In summary, our formative study yields two types of findings: (1) four design principles (DP1–DP4) that guide the system design in Section 4, and (2) direct evidence that the generate-then-verify paradigm fails in evidence-based legal writing — lawyers detected only X of 5 planted errors in the probe task, and reported that plausible-looking AI mappings received insufficient scrutiny. These findings motivate the extract-then-assemble paradigm we describe next, and position our work within the broader landscape of AI-assisted writing tools (Section 3)."

为什么需要这段：Formative Study 现在承担双重任务（DP + 失效证据），需要一个小结把两个产出捆在一起，同时为接下来的 Related Work（Sec 3）和 System Design（Sec 4）做预告。

五、系统设计（论文中为 Section 4，[v6.6] ~2500w）

Running Example: Attorney Maria + Dr. Lin Chen。

4.1 两阶段工作流

Phase 1: 上传 → OCR → 粗分类 → 律师确认 → AI 提取 snippet + 初始映射建议

Phase 2: 审查 AI 建议 → 拖拽修正 → 框选新 snippet → 创建 Argument → 生成 petition → 溯源验证

[v6.6 #12] *Implementation detail: "We use [model] for OCR, which provides both bounding box coordinates and semantic text extraction in a single pass, enabling the dual indexing required by our provenance engine."* — 不作为技术贡献推。

4.2 三区布局 (Evidence Mapping View)

左区 Document Viewer | 中区 Evidence Card Pool | 右区 Standards Panel。（与 v6.2 相同）

4.3 Writing Canvas

Snippet/Argument/Standard 三层节点 + 分屏双向联动。（与 v6.2 相同）

[v6.6 #13] ~~删除"snippet 删除后 LLM 重新生成子论点"的描述。~~ 未实现的设想不应出现在 System

Design 里。降级为 Future Work 一句话。

[v6.6 #10 核心新增] 4.4 Cross-Granularity Provenance Engine (~400w)

[v6.6 #10] 这是本版最关键的新增内容。溯源引擎从一句话展开为完整 subsection。

问题定义：

When [SystemName] generates a petition sentence, how does the system automatically trace it back to the physical location in the source document? The core difficulty is that a single generated sentence may synthesize information from multiple snippets across different documents, and conversely, a single snippet may support multiple generated sentences. The provenance engine must maintain these many-to-many relationships while providing instant, fine-grained navigation.

三层溯源图数据模型：

层级	节点定义	边的语义	示例
Sentence Layer	生成的 petition 中每个句子	sentence → snippet: “grounded-in”	“Dr. Chen’s work has been cited 3200 times” → citation_snippet
Snippet Layer	从源文档提取的证据片段	snippet → bbox: “located-at”	citation_snippet → Google Scholar PDF p.1 (120,340,580,380)
BBox Layer	源文档中的物理位置 (page, x, y, w, h)	bbox → document: “part-of”	(120,340,580,380) ∈ google_scholar.pdf

Pipeline（需在论文中配架构图 Figure X）：

- **Step 1 — Dual Indexing：**OCR 对每份源文档产出 text + bounding box 坐标。每个文本区块同时拥有语义表示（embedding）和物理位置。
- **Step 2 — Snippet Extraction：**LLM 识别文档中的证据片段，每个 snippet 继承其源 text block 的 bbox 坐标。如果 snippet 跨越多个 text block，取 bounding box 并集。
- **Step 3 — Constrained Petition Generation [v6.7 #16 重写]：**系统将律师已映射的 snippet 集合作为唯一输入 context，要求 LLM 以 structured JSON 格式输出，每个句子对象包含 text 和 snippet_ids 数组。两个作用：(1) 只将已映射 snippet 放入 context window，从源头限制幻觉；(2) structured output 产出显式的 snippet-sentence 链接，为溯源提供 deterministic anchor。

Structured Output 示例：

{ "sentences": [{ "text": "Dr. Chen's research has been cited over 3,200 times...", "snippet_ids": ["snip_014", "snip_027"] }, { "text": "His foundational work on...", "snippet_ids": ["snip_014"] }] }

- **Step 4 — Hybrid Retrieval [v6.7 #17 重写]：**溯源优先使用 Step 3 产出的显式 snippet_id 标注（precision 高）。语义 embedding 检索作为 fallback，处理两种情况：(a) LLM 融合多个 snippet 但只标注了部分；(b) 律师手动编辑生成文本后原始标注失效。系统对两个信号加权合并：显式标注权重高于语义检索。
- **Step 5 — BBox Highlight：**检索到 snippet 后，系统查找其 bbox 坐标，在 Document Viewer 中高亮对应区域。整个回溯在 <200ms 内完成。

一对多处理策略：

- **一句话→多个 snippet：**按 (a) 显式引用优先 (b) 语义相似度排序。默认展示 top-3，可展开查看全部。每个 snippet 用不同颜色的 bbox 高亮。
- **一个 snippet→多句话：**反向查看：用户点击 Document Viewer 中的 snippet，右侧 petition 中所有引用该 snippet 的句子高亮。

Figure 计划：需要一张 pipeline 架构图，展示：

Source Document → [OCR: text + bbox] → [Snippet Extraction] → [Lawyer Drag Mapping] → [Petition Generation with snippet ID annotation] → [User clicks sentence] → [Semantic Retrieval] → [BBox Highlight in Document Viewer]

这张图会成为 System Design 里最重要的 figure 之一。

[v6.7 #18 新增] 设计 Rationale（写入论文 4.4 末尾）：

"We chose structured generation over post-hoc retrieval as the primary provenance mechanism because, in high-stakes legal writing, traceability cannot rely on probabilistic semantic matching alone — lawyers need deterministic evidence links. The structured JSON output provides a deterministic anchor: each generated sentence explicitly declares its source snippets at generation time. Semantic retrieval serves as a graceful degradation mechanism for cases where explicit annotations are incomplete (e.g., when the LLM synthesizes multiple snippets but annotates only a subset) or invalidated (e.g., when the lawyer manually edits the generated text). This 'deterministic anchor + probabilistic fallback' architecture reflects our broader design philosophy that high-stakes professional writing demands different reliability guarantees than general-purpose AI writing tools."

为什么这段很重要：① 直接支撑“high-stakes writing 需要不同设计”的整体叙事。② 把溯源引擎与通用 RAG 系统区分开——RAG 完全依赖语义检索，我们用 structured generation 作为主要手段。③ “deterministic anchor + probabilistic fallback” 可以作为一个可推广的技术设计原则。

4.5 多视图模式

Line View | Embedded View | Matrix View | Sankey View。（与 v6.2 相同）

[v6.6 #14 新增] 4.6 Technical Evaluation: 溯源准确率 (~200w)

[v6.6 #14] 低成本但重要的技术验证。可放在 System Design 末尾或 User Study 开头。

目的：验证溯源引擎的检索准确性，给 C2 技术贡献提供定量支撑。

方法：

- 从生成的 petition 中随机抽取 50 句文本
- 两名研究者独立标注每句话的正确 snippet 来源（gold standard）
- 计算溯源引擎的 Precision@3、Recall、MRR（Mean Reciprocal Rank）
- 记录 BBox 定位的 IoU（Intersection over Union）

预期报告格式：

Precision@3	Recall	MRR	BBox IoU
XX%	XX%	0.XX	0.XX

成本估算：一人 2-3 小时即可完成标注。不需要律师参与，研究者自己或研究助理即可。

为什么这很重要：溯源引擎是技术贡献的核心，但如果没有任何定量数据证明它 work，reviewer 会觉得你只是“声称”而非“验证”。Precision/Recall + IoU 是最直接的证据。

六、Related Work（论文中为 Section 3）

[v6.4 #3] Related Work 现在在 Formative Study 之后。读者已经看到了 probe task 数据和 DP1-DP4，再读 Related Work 时对 extract-then-assemble 框架已有认知。

3.1 AI-Assisted Writing with Provenance：InkSync | HaLLMark | ABScribe | CorpusStudio。所有都是 generate-then-verify。

3.2 Argument Construction & Sensemaking：VISAR | Sensecape | Graphologue | GLITTER。VISAR 定位重写为问题结构差异。

[v6.5 #7 新增] VISAR 对比中引用 formative findings（利用前置红利）：

"While VISAR's generation-oriented workflow is well-suited for open-ended essays, our formative study (Section 2) reveals that in evidence-based writing, the core bottleneck is not argument invention but evidence assembly under fixed standards — a fundamentally different problem structure that calls for a different collaboration paradigm."

为什么这样改：Formative Study 已在 Section 2 呈现，读者已经看到 probe task 数据。现在在 Related Work 中引用自己的数据来论证“为什么现有工具不够用”，比纯文献对比更有说服力。这是 Formative Study 前置的核心红利。

[v6.5 #7] 同样的交叉引用可以用在其他关键对比处：

- **vs InkSync/HaLLMark：***"Our formative study shows that in evidence-based writing, the verification bottleneck is not tracking AI's changes (InkSync) or AI's contribution history (HaLLMark), but tracing generated claims to specific source documents — a form of source provenance that existing tools do not provide."*
- **vs Clearbrief：***"While Clearbrief addresses post-writing citation verification, our formative interviews reveal that errors introduced during the evidence-to-standard mapping phase — before writing begins — are harder to detect and more consequential."*

3.3 Legal Writing & Domain-Specific AI：Swoopes | Clearbrief | QuickFiling | Visalaw。

3.4 差异化矩阵：含"协作范式"维度。

（完整内容见 v6.1/v6.3 文档）

[v6.4 #4] 七、评估设计（论文中为 Section 5）

类型：Within-subjects | **参与者：**6-8 名 | **核心对比：**人控映射 vs 只读映射（溯源两组都有）

条件 A：完整系统（拖拽 + Argument 层 + Canvas + 溯源）

条件 B：AI 映射只读 + 溯源可用 + 无拖拽/Argument/Canvas

材料：2 套（律所提供），各预埋 5 错误，拉丁方平衡

[v6.4 #4 新增] 归因逻辑显式化（必须写入论文）：

"A critical feature of our experimental design is that both conditions present identical information: the same AI-generated mappings are visible in both the interactive (Condition A) and read-only (Condition B) conditions, and both include the provenance engine for tracing claims to source documents. The only difference is whether lawyers can actively modify the mappings through drag-and-drop. Therefore, any observed differences in error detection or confidence cannot be attributed to differences in information availability, but rather to the cognitive engagement demanded by the construction task — providing direct empirical support for the mapping-as-verification principle."

为什么这段很重要：Reviewer 可能质疑"实验组胜出是因为有更多功能（Argument 层、Canvas），不是因为 mapping as verification"。这段话把逻辑链说清楚：信息量相同 → 差异在于交互方式 → 支持 mapping as verification。配合访谈 ablation 块进一步拆分各组件。

（其余评估设计内容与 v6.2 相同：任务设计、测量指标、访谈协议含 ablation 块、审稿风险应对）

[v6.4 #5] 八、Discussion（论文中为 Section 6）

8.1 Mapping as Verification (~400w)

（与 v6.3 相同：主流范式批判 + 构建即验证 + 其他领域推广 + automation level 讨论）

[v6.4 #5 全新] 8.2 Why Not Just Better AI? (~200w)

[v6.4 #5] 预防性论证：即使 AI 映射准确率显著提高，extract-then-assemble 仍有独立价值。

A natural question is whether the extract-then-assemble paradigm will become unnecessary as AI mapping accuracy improves. We argue that this paradigm retains independent value even with near-perfect AI, for three reasons:

(1) Accountability. *In legal practice, the attorney — not the AI — bears professional responsibility for every claim in a petition. Regulatory and ethical obligations require lawyers to exercise independent judgment on each evidence-to-standard mapping, regardless of AI accuracy. A paradigm that embeds this judgment into the workflow, rather than relegating it to a perfunctory review step, better serves the professional requirement of accountability.*

(2) Case comprehension. *The act of constructing evidence-to-argument mappings helps lawyers form a holistic understanding of the case — an understanding that is essential in downstream tasks such as responding to Requests for Evidence (RFEs), preparing for interviews, and crafting supplementary arguments. A fully automated mapping, even if accurate, would deprive lawyers of this constructive engagement with the case materials.*

(3) Trust as a product of engagement, not accuracy. *Our experimental results suggest that lawyers' confidence was higher in the interactive condition not merely because the mappings were more accurate (both conditions showed identical AI mappings), but because the construction process itself engendered a sense of ownership and control. This trust benefit persists regardless of AI accuracy — it derives from the human's active role in the mapping process, not from the AI's error rate.*

为什么必须加这段："等 GPT-6 出来你的 motivation 就没了"是审稿人最容易想到的质疑。三条论证形成递进：法律责任（制度约束）→ 案件理解（认知价值）→ 信任来自参与而非准确率（实证证据）。第三条直接引用实验数据，最有说服力。

8.3 Generalizability (~250w)

（与 v6.3 相同：专利、医疗保险、合规审计、科研基金四个推广场景）

8.4 Limitations (~250w)

- **1. 样本量：**N=6-8。三角验证增强。
- **2. 未做 ablation：**访谈 ablation 块 + log。
- **3. 预设材料：**未来 deployment study。
- **4. AI 组件成熟度：**可能 Wizard-of-Oz。
- **5. 领域特异性：**DP1-DP4 基于移民法。

8.5 Future Work (~150w)

- Longitudinal deployment | Ablation study | 跨领域验证 | 协作功能 | AI 自适应
- **[v6.6 #13]** 智能重组：snippet 删除/移动后 LLM 自动重新生成子论点（从 System Design 降级到此处）

九、论文结构（约 **8500** 字）

[v6.6 #15] System Design +300w（溯源展开），Related Work -200w（矩阵表格化），User Study +200w（技术评估）

Sec	标题	字数	核心内容
1	**Introduction**	~800	gen-then-verify 假设 → 两个失效 → extract-then-assemble + mapping as verification
2	**Formative Study**	~800	Probe task + DP1-DP4 + gen-then-verify 失效证据 + 桥接段
3	**Related Work**	**~1100**	**压缩: 矩阵表格化节省空间; 交叉引用 formative findings**
4	**System Design**	**~2500**	**扩充: 溯源引擎完整展开(+300w) + 技术评估 + 架构图**
5	**User Study**	~2000	归因逻辑显式化 + 人控 vs 只读 + ablation 访谈
6	**Discussion**	~1000	Mapping as verification + Why not better AI + Generalizability + Limitations
7	**Conclusion**	~400	总结贡献 + 未来工作

[v6.4] 十、Abstract

AI-assisted writing tools typically follow a “generate-then-verify” paradigm: AI produces content or structure, and humans review the output. While effective for creative writing, we argue this paradigm is fundamentally mismatched for evidence-based professional writing, where every claim must be traceable to source documents and verification requires deep engagement with the underlying materials.

Through formative interviews with immigration lawyers — including a probe task in which lawyers reviewed AI-generated evidence mappings with planted errors — we identify two failure modes of generate-then-verify: (1) end-to-end generation obscures the evidence behind each claim, and (2) even when AI mappings are exposed, lawyers exhibit insufficient scrutiny of plausible-looking mappings, a pattern consistent with automation bias. We propose an alternative paradigm — extract-then-assemble — where AI extracts evidence snippets from source documents and lawyers construct the evidence-to-argument mapping through drag-and-drop interactions. A key design insight is that the act of mapping inherently subsumes verification: deciding where to place evidence requires the same judgment that a separate verification step would demand.

We instantiate this paradigm in [SystemName] for immigration petition writing, featuring a three-layer mapping interface (Snippet → Argument → Standard) and a cross-granularity provenance engine. In a within-subjects study with N immigration lawyers, human-constructed mappings improved error detection rates (XX% vs YY%) and increased confidence — despite both conditions presenting identical information — providing direct empirical support for the mapping-as-verification principle.

十一、提交前检查清单

[v6.7 新增项以 ☆ 标记]

- ☆ Step 3 是否展开为 Constrained Petition Generation + Structured JSON 示例?
- ☆ Step 4 是否重定义为 Hybrid Retrieval（显式标注主 + 语义检索 fallback）?
- ☆ 是否有 “deterministic anchor + probabilistic fallback” 设计 rationale?
- ★★★ 溯源引擎是否展开为完整 subsection（问题定义 + 数据模型 + Pipeline + 一对多策略）?
- ★★★ 是否有 pipeline 架构图（Figure X）?
- ★★★ 是否有溯源准确率技术评估（Precision@3, Recall, MRR, BBox IoU）?
- ★★★ C2 是否以溯源引擎为核心技术组件表述?
- ★★★ OCR 是否定位为 implementation detail 而非贡献?
- ★★★ 是否删除了未实现的 “snippet 删除后 LLM 重新生成”?
- ★★ Probe Task 是否框定为 structured interview probe?
- ★★ Related Work 是否交叉引用 formative findings?
- ★★ Formative Study 是否有桥接段?
- ★ Automation bias 是否 “consistent with” + Parasuraman & Riley?
- ★ 章节顺序 Formative → Related Work?
- ★ User Study 是否写明“两组信息量相同”?
- ★ Discussion 是否含 “Why not better AI?”?
- P1 从 generate-then-verify 开头? C1=范式+洞察? Mapping as verification 在 P4?
- VISAR 基于问题结构差异? 评估人控 vs 只读? 访谈含 ablation 块?

十二、引用列表（v6.4 更新）

[v6.4 #2 新增]

1. **Parasuraman & Riley (1997).** Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2):230–253. — *automation bias 经典文献*

核心引用（与 v6.2/v6.3 相同）：

2. Laban et al. 2024. InkSync. UIST’24.
3. Hoque et al. 2024. HaLLMark. CHI’24. arXiv:2311.13057
4. Reza et al. 2024. ABScribe. CHI’24. arXiv:2310.00117
5. Dang, Swoopes, Buschek, Glassman. 2025. CorpusStudio. CHI’25. arXiv:2503.12436
6. Zhang et al. 2023. VISAR. UIST’23. arXiv:2304.07810
7. Suh et al. 2023. Sensecape. UIST’23.
8. Jiang et al. 2023. Graphologue. UIST’23.

9. Peng et al. 2025. GLITTER. UIST'25. arXiv:2504.14695
10. Swoopes et al. 2025. arXiv:2509.24854
11. Braun & Clarke. 2006. Using thematic analysis in psychology.
12. Hart & Staveland. 1988. NASA-TLX.

十三、当前实现状态

Evidence Mapping： ✅ 三区布局 | ✅ 拖拽映射 | ✅ 连线+Focus | ✅ 框选 Snippet | ✅ 多视图

Writing Canvas： ✅ 三层节点图 | ✅ 平移/缩放 | ✅ 连线编辑 | ✅ 分屏 | ✅ 双向联动

全局： ✅ i18n | ✅ 状态管理 | ✅ localStorage

待完成：LLM OCR | AI 生成 Snippet | AI 生成 Petition | 溯源引擎 | 导出

— v6.7 溯源 Pipeline 澄清版：Structured Generation + Hybrid Retrieval + Deterministic Anchor 设计原则 —