**大規模實驗報告**

"早期孔門主要文獻"文本相似性比較

管線版本：Pipeline20

模型：Qwen3-Embedding-8B + Qwen3-Reranker-8B

**一、實驗背景**

**1.1 研究問題**

先秦兩漢文獻的流傳以「單篇別行」為通例。一篇文字在傳抄過程中被不同的彙編者收入各書，形成同源異本的「互見」現象。互見段落的辨識是考證文獻源流、建立文本譜系的基礎工作，傳統上完全依賴學者通讀精校，耗時巨大且難以窮盡。

本研究開發一套自動化文本相似性檢測管線（Pipeline），以「字面統計 + 深度語義 + 雙通道融合」三通道架構，對先秦至漢代的「孔門主要文獻」進行全量比對。最終目標不是取代人工校勘，而是為學者提供一份高信噪比的候選對列表，大幅縮減人工篩查範圍。

本實驗在 Pipeline20 版本下進行大規模運算。Pipeline20 的核心改進為「統一雙通道補齊」（Unified Dual-Channel Completion）：對字面通道召回但缺少語義分數的候選對，回填語義餘弦；對語義通道召回但缺少字面分數的候選對，補齊 cosine\_tf 等字面指標。此舉消除融合通道的全部 NaN，確保每對候選同時具備完整的字面和語義分數。

**1.2 語料概況**

本實驗使用「孔門彙編」語料庫第 3 版（孔门汇编3.xlsx），涵蓋 20 部先秦至漢代與孔子及其弟子相關的核心文獻，總計 221,880 字（純中文字符計數）。語料按三個層次組織：H1（書）= 20 個單元，H2（篇章）= 307 個單元，Normal（自然段）= 2,426 個單元。

**表 1 語料層次結構**

|  |  |  |  |
| --- | --- | --- | --- |
| **層次** | **單元數** | **平均字數** | **說明** |
| H1（書） | 20 | ~11,094 | 整本書或獨立長篇，如《論語》全書、《孔子世家》 |
| H2（篇章） | 307 | ~723 | 書中自然篇章，如《論語·學而》、《孔子世家·甲》 |
| Normal（章） | 2,426 | ~91 | 自然段落，為自動校勘的最小比對單元 |

20 部文獻涵蓋經部（《論語》《禮記》《大戴禮記》）、史部（《史記·孔子世家》《史記·弟子列傳》）、子部（《荀子》《孟子》《晏子春秋》《法言》）、集部（《說苑》《新序》）及新出土材料（《孔子衣鏡》《上博簡》《定縣漢簡》）、晚出彙編（《孔子家語》《韓詩外傳》《孔叢子》《仲尼微言》）等。這些文獻在傳世過程中存在大量互見：如《史記》征引《論語》82 段、111 次；《孔子家語》與《禮記》《大戴禮記》存在系統性的平行段落。

文獻長度差異巨大：最長的《孔子家語》超過 44,000 字，最短的《定縣漢簡》不足 1,000 字，《孔子衣鏡》約 2,700 字。這種量級的差異對歸一化方案提出了嚴峻挑戰（詳見 2.5 節）。

**1.3 模型選擇**

語義向量化採用 Qwen3-Embedding-8B（隱層維度 4,096，最大 token 長度 8,192），支持中英雙語及古代漢語。與 BGE-M3 等模型的對照實驗表明：Qwen3 在 H2 層的 Spearman 相關 ρ=0.77 顯著高於 BGE 的 ρ=0.51（Normal 層），且 BGE 存在 sc=1.000（22 對，全部禮記）和 sc<0（20 對，全部弟子列傳）的編碼異常。

Reranker 採用 Qwen3-Reranker-8B（交叉注意力架構），作為獨立旁路（Sidecar）對融合通道候選對進行二次審計，其分數獨立記錄，不參與排名計算。選擇 8B 而非 4B 的依據：4B 語義通道壓縮嚴重（H1 std=0.015 vs 8B std=0.035），融合分數幾乎完全由字面主導，失去雙通道互補的意義。

**表 2 模型配置**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **用途** | **維度** | **最大長度** | **選擇依據** |
| Qwen3-Embedding-8B | 語義向量化 | 4,096 | 8,192 tokens | 古漢語適應性、無編碼異常 |
| Qwen3-Reranker-8B | 交叉注意力審計 | — | 8,192 tokens | 獨立旁路、不參與融合排名 |

**二、系統架構與算法原理**

**2.1 三層次三通道架構**

本系統的核心設計是「三層次 × 三通道」九場景矩陣。三個文本層次（書/篇章/段落）對應不同粒度的檢索目標；三個比對通道（字面/語義/融合）分別捕捉不同類型的文本關聯。九種組合各有其適用場景：

**表 3 九場景矩陣功能定位**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **字面通道** | **語義通道** | **融合通道** |
| H1 書 | 整書級風格指紋 | 深層主題相近 | 書籍間網絡結構 |
| H2 篇 | 互見片段識別 | 語義甜蜜區間 | 最優區分力場景 |
| Normal 章 | 逐字異文比對 | 段落語義歸類 | 自動校勘候選 |

設計哲學：字面通道提供可追溯的物理證據（共享字符三元組），語義通道捕捉改寫後仍保留的深層關聯。兩者互補：僅靠字面會遺漏大規模改寫的互見，僅靠語義會產生「主題相近但無引用關係」的假陽性。融合通道以 7:3 的權重（字面為主，語義為輔）整合兩者，與傳統校勘學「死校為主，理校為輔」的原則相呼應。

**2.2 字面通道**

字面通道完全基於字符三元組（char-trigram）統計，不依賴任何預訓練模型，計算確定性且可復現。四種指標各有側重，構成互補的證據鏈：

**表 4 字面通道四指標及權重**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **指標** | **公式核心** | **衡量維度** | **H1/H2 權重** | **Normal 權重** |
| cosine\_tf | cos(TF\_A, TF\_B) × len\_balance | 文體風格相似 | 35% | 60% |
| Coverage | F1(cov\_A, cov\_B) × overlap\_ratio | 互見片段證據 | 35% | —（不啟用） |
| cosine\_tfidf | cos(TFIDF\_A, TFIDF\_B) | 去噪後內容相似 | 20% | 20% |
| Jaccard | |A∩B| / |A∪B| × γ校正 | 集合穩健重疊 | 10% | 20% |

Coverage 指標是字面通道的核心創新。其計算流程如下：

Step 1：將文本 A 和 B 各自切分為 200 字的分段（stride=100），構造分段對的餘弦矩陣。

Step 2：對矩陣每行取最大值，得到 A 各分段的最佳匹配得分 cov\_A；對每列取最大值，得到 cov\_B。

Step 3：F1 融合：Coverage = 2 × mean(cov\_A) × mean(cov\_B) / (mean(cov\_A) + mean(cov\_B))，再乘以 overlap\_ratio（共享三元組比例），得到最終的 Coverage 分數。

Coverage 的設計理念：互見段落往往只佔全文的一部分（如《孔子世家》征引《論語》82 段，但《世家》本身超過 6,800 字）。全文餘弦會被大量非互見內容「稀釋」，而 Coverage 的分段匹配機制能精確捕捉局部重疊。

len\_balance 長度平衡因子：

len\_balance = 2 × min(L\_A, L\_B) / (L\_A + L\_B)。等長時為 1.0，長度差異越大懲罰越重。公式的含義是：兩段文本只有在長度接近時，cosine\_tf 的高值才有比對意義；長短懸殊時，短文本容易因偶然共享三元組而虛高。Normal 層設置 floor=0.30，避免極短段落（平均 91 字）被過度懲罰至趨近零。

Jaccard gamma 校正：

原始 Jaccard 隨文本長度增加而系統性偏低（N-gram 並集遠大於交集）。引入長度自適應 gamma：短文本（<200 字）保持原始值（γ≈1.0），長文本（>1,200 字）取 γ=0.80 進行溫和提升（j^0.80 > j）。此校正不改變排序的相對位置，僅壓縮量級差異。

Normal 段落不啟用 Coverage：

段落文本過短（平均 91 字），Coverage 的分段匹配機制退化為 len\_balance × cosine\_tf（200 字分段包含整段），失去獨立信息。Normal 層以 cosine\_tf（60%）為主導，輔以 TF-IDF（20%）去噪和 Jaccard（20%）穩健校驗。此為 10 項共識之一（共識 #1）。

**2.3 語義通道**

語義通道使用 Qwen3-Embedding-8B（4,096 維）提取文本向量。不同層次的向量構造策略根據文本長度和比對目標量身設計：

**表 5 語義通道三層策略**

|  |  |  |  |
| --- | --- | --- | --- |
| **層次** | **策略** | **原理** | **驗證** |
| H1 書 | MaxSim 分塊 | 切割為 500 字塊（stride=400），逐塊最大餘弦取均 | λ=0，純局部匹配 |
| H2 篇 | Avgpool 平均池化 | avg-pool token embeddings（max\_length=8192） | ndcg@40=0.701 > MaxSim 0.673 |
| Normal 章 | 單向量餘弦 | 直接 avgpool 作為完整語義表徵 | cos(e\_A, e\_B) |

H1 MaxSim 公式：sim(A→B) = (1/m) Σᵢ maxⱼ cos(aᵢ, bⱼ)，其中 aᵢ 為 A 的第 i 塊向量，bⱼ 為 B 的第 j 塊向量。雙向取均：score = [sim(A→B) + sim(B→A)] / 2。MaxSim 適合 H1 層（書級文本 1K-44K 字），因整書 avgpool 會將局部互見信號「平均掉」。

H2 篇章層為何選 Avgpool 而非 MaxSim？

H2 篇章的平均長度為 723 字，在 8,192 token 上下文窗口內可完整編碼。實測表明 Avgpool 的 ndcg@40 = 0.701，高於 MaxSim 的 0.673。原因是：篇章級文本通常主題集中，全文平均向量能有效捕捉整體語義；而 MaxSim 的分塊匹配會引入跨塊的虛假高分。

語義餘弦的有效範圍與「語義甜蜜區間」：

語義餘弦值在 H2 層展現最大有效範圍（range = 0.835），H1 僅 0.158（長文本 avgpool 壓縮），Normal 僅 0.658（短段落信息量不足）。H2 篇章（1K–5K 字）恰好處於「足夠長以承載獨立主題、又不至於過長導致信號稀釋」的甜蜜區間，語義通道在此層次發揮最大區分力。

**2.4 融合通道與權重設計**

融合公式：

S\_final = 0.70 × q\_lex + 0.30 × q\_sem

其中 q\_lex 為字面歸一化分數，q\_sem 為語義歸一化分數（均由 robust\_soft 方案映射到 (0,1) 區間）。

權重 7:3 的設計依據：

（1）字面通道提供的證據可直接追溯到原文（共享字符三元組），為「死校」（據本校之）的技術實現，應當佔主導。（2）語義通道提供的是推理性證據（「這兩段討論的是同一主題」），為「理校」（據理改之）的技術實現，應當為輔。（3）語義通道存在「主題相近但無引用關係」的假陽性風險（如《禮記》與《荀子》多段討論「禮」的段落語義高分但非互見），需要字面通道的硬證據約束。

融合通道的候選集是字面和語義兩個通道候選集的並集。字面通道取 topk=14，語義通道取 topk=6，合併後去重得到融合候選集。Pipeline20 對合併後的候選對進行雙通道補齊（詳見 2.7 節）。

**2.5 歸一化方案：robust\_soft（分段 tanh）**

原始分數的量級在不同通道和層次之間差異極大，無法直接加權融合。例如字面 cosine\_tf 的 H2 均值約 0.03，而語義餘弦的 H2 均值約 0.85。歸一化的目標是將各通道的分數映射到統一的 (0,1) 區間，同時保持排序不變。

歸一化方案的演進經歷了四代迭代：

**表 6 歸一化方案四代演進**

|  |  |  |  |
| --- | --- | --- | --- |
| **版本** | **方案** | **問題** | **改進** |
| pipe10 | min-max 線性 | 受極端值影響，H1 僅 190 對無法穩健估計端點 | pipe12 改用排名 |
| pipe12 | rank-based 排名 | 丟失分數量級信息，所有分布壓成均勻分布 | pipe13 改用百分位 |
| pipe13 | robust\_linear 百分位線性 | H2 天花板飽和 41/4,707、Normal 786/37,258 ≥ 0.999 | pipe14 加 tanh 壓縮 |
| pipe14–20 | robust\_soft 分段 tanh | 零飽和、嚴格 (0,1)、96.4% 數據在線性區 | 最終方案 |

robust\_soft（分段 tanh）的三步計算：

Step 1 百分位標準化：z = (x - p\_lo) / (p\_hi - p\_lo)，其中 p\_lo/p\_hi 為該層次的百分位端點。

Step 2 分段壓縮：z < 0 時 z\_soft = -tanh(-k·z)（左尾柔性壓縮）；0 ≤ z ≤ 1 時 z\_soft = z（線性透傳）；z > 1 時 z\_soft = 1 + tanh(k·(z-1))（右尾柔性壓縮）。

Step 3 映射到 (0,1)：q = ε + (z\_soft + 1) / 3 × (1 - 2ε)，其中 ε = 1e-6 保證嚴格開區間。

**表 7 歸一化分層參數**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **參數** | **H1** | **H2** | **Normal** | **設計理由** |
| 百分位端點 | p10/p90 | p5/p95 | p5/p95 | H1 樣本小(190對)用寬護帶避免端點估計不穩 |
| tanh k | 0.20 | 0.10 | 0.10 | H1 尾部短，可較強壓縮；H2/Normal 用溫和壓縮 |
| ε | 1e-6 | 1e-6 | 1e-6 | 保證嚴格開區間 (0,1)，不含邊界值 |

驗證結果：全部層次零天花板飽和（max score\_final < 0.87）；96.4% 的數據落在線性區間 [0, 1] 內，不受 tanh 壓縮影響；分數嚴格在 (0, 1) 開區間內，不存在 0 或 1 的邊界值。

歸一化三難困境：

理論上，無法同時滿足以下三個條件：（1）統一歸一化方案（所有通道同一公式）；（2）嚴格有界 [0,1]；（3）全區間線性。robust\_soft 選擇犧牲「全區間線性」：在中央 96.4% 保持線性，僅在尾部引入溫和的 tanh 壓縮。這是目前最優的工程折中。

**2.6 Reranker Sidecar**

Qwen3-Reranker-8B 採用交叉注意力架構，對融合通道候選對進行獨立的二次評估。與 bi-encoder（分別編碼再計算相似度）不同，cross-encoder 將兩段文本拼接後聯合編碼，能捕捉更精細的文本交互。

Reranker 的設計原則為「Sidecar」（旁路）：其分數獨立記錄在 reranker\_logit 欄，不參與 score\_final 的融合計算，也不影響排名。用途為：人工核對候選對時，Reranker 分數提供第三方佐證。若融合排名高而 Reranker 分數低，提示該對可能為假陽性（字面偶然高分）；若融合排名低而 Reranker 分數高，提示可能存在被字面遺漏的語義互見。

**表 8 Reranker 覆蓋範圍**

|  |  |  |  |
| --- | --- | --- | --- |
| **層次** | **Reranker 覆蓋** | **精排窗口** | **說明** |
| H1 | 全部 190 對 | Top-100% | H1 候選數少，全量精排 |
| H2 | Top-3%（~140 對） | Top-3% | RERANK\_DEFAULT\_TOP\_PERCENT = 3.0 |
| Normal | Top-3%（上限 1,200 對） | Top-3%（Cap=1200） | 超過 Cap 截斷以控制計算量 |

**2.7 Pipeline20 雙通道補齊**

在融合通道中，候選集是字面 Top-14 和語義 Top-6 的並集。這意味著部分候選對只有字面分數（語義通道未選中），部分只有語義分數（字面通道未選中）。Pipeline17 引入了 semantic\_cosine 回填（backfill），但僅回填語義→字面方向。Pipeline20 實現完整的雙向補齊：

**表 9 Pipeline20 雙通道補齊策略**

|  |  |  |  |
| --- | --- | --- | --- |
| **場景** | **缺失指標** | **補齊方式** | **計算成本** |
| Lex-only pair（無語義分數） | semantic\_cosine | 利用緩存向量計算 np.dot | O(D) 點乘，極低 |
| Sem-only pair（無字面分數） | cosine\_tf, coverage, jaccard 等 | 全套字面指標重新計算 | O(V) 稀疏運算，較高 |

補齊後的驗證：融合通道 semantic\_cosine NaN 由 pipe17 前的 66.7%（H2）/ 69.1%（Normal）降至 0%；cosine\_tf 和 jaccard 的 NaN 同樣降至 0%。所有融合候選對均同時具備完整的字面和語義分數。

補齊的設計細節：coverage 補齊在 cosine\_tf 之前執行（因為 H1/H2 的 cosine\_tf = coverage\_score，需要先有 coverage 結果）。字段級 if-None 獨立補齊，防止半殘 row。H1/H2 coverage 補齊失敗時直接 raise RuntimeError，不降級處理，避免 35% 權重的 Coverage 缺失造成分數偏差。

**三、實驗流程與自動校勘**

**3.1 管線流程**

Pipeline20 的完整運算分為以下七個階段：

**表 10 Pipeline20 七階段流程**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **階段** | **名稱** | **輸入** | **輸出** | **耗時量級** |
| 1 | 語料解析 | Excel 語料庫 | H1/H2/Normal 文本單元 | 秒級 |
| 2 | TF-IDF 構建 | 全部文本 | 字符三元組稀疏矩陣 | 秒級 |
| 3 | 語義向量化 | 全部文本 | 4,096 維 dense embeddings | 分鐘級（GPU） |
| 4 | 字面通道候選生成 | TF-IDF 矩陣 | Top-14 字面候選對 + 四指標 | 秒級 |
| 5 | 語義通道候選生成 | Embedding 矩陣 | Top-6 語義候選對 + 餘弦 | 秒級 |
| 6 | 融合 + 雙通道補齊 | 字面 + 語義候選集 | 完整融合候選對 + score\_final | 分鐘級 |
| 7 | Reranker 精排 | 融合 Top-3% | reranker\_logit 審計分數 | 分鐘級（GPU） |

整體運算在 Apple M-series 晶片（28 GiB 統一記憶體）上完成。語義向量化和 Reranker 精排利用 MPS（Metal Performance Shaders）GPU 加速。大規模實驗（20 本書）的總耗時約 40–60 分鐘，瓶頸在語義向量化（階段 3）和 Reranker（階段 7）。

**3.2 自動校勘**

在融合通道 Normal 層的 Top-1,000 候選對上，系統自動執行字符級比對（auto-collation），找出異文（variant readings）。校勘參數：字符級比對（char-level），上下文窗口 = 8 字，合併間距 = 1，最大條目數 = 12，相似度閾值 = 0.32。

自動校勘的設計思路是「篇章定位 → 段落校勘」的分層檢測：融合 H2 見出整體關係（哪些篇章間存在互見），下沉至 Normal 尋找具體候選段落，最後由自動校勘提取異文，再由學者細讀考證。系統不取代人工判斷，而是將人工篩查範圍從全量 2,426 × 2,425 / 2 ≈ 294 萬段落對縮減至 1,000 對。

**四、總體實驗結果與數據分布**

**4.1 候選對規模**

大規模實驗共產出 41,996 個融合候選對（H1: 190，H2: 4,676，Normal: 37,130）。以下為九場景矩陣的完整候選對數量：

**表 11 候選對規模（九場景矩陣）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **通道/層次** | **H1 書** | **H2 篇** | **Normal 章** | **合計** |
| 字面 | 190 | 5,210 | 40,264 | 45,664 |
| 語義 | 190 | 5,064 | 38,612 | 43,866 |
| 融合 | 190 | 4,676 | 37,130 | 41,996 |

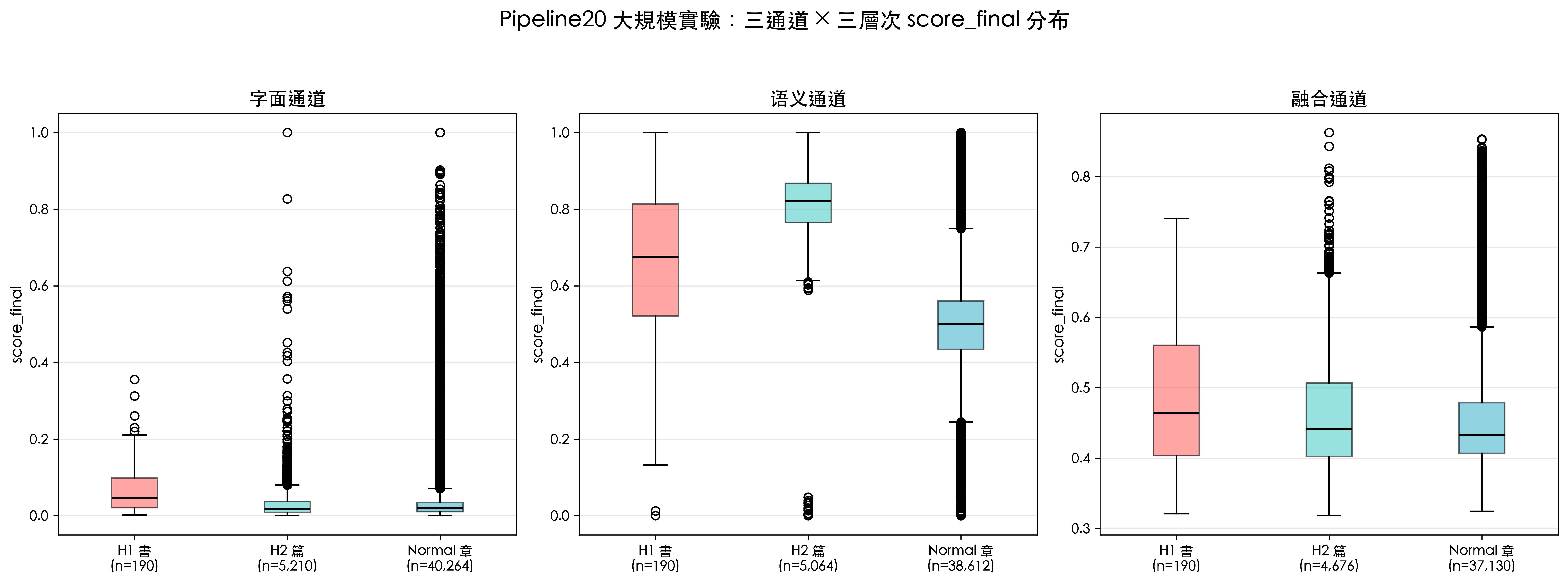
H1 層固定為 C(20,2) = 190 對（全量配對），H2 和 Normal 層由 topk 候選生成機制決定。字面和語義的單通道候選數略多於融合通道，因為融合通道是兩者取並集後去重，部分重複對只計算一次。

**4.2 融合通道 score\_final 分布**

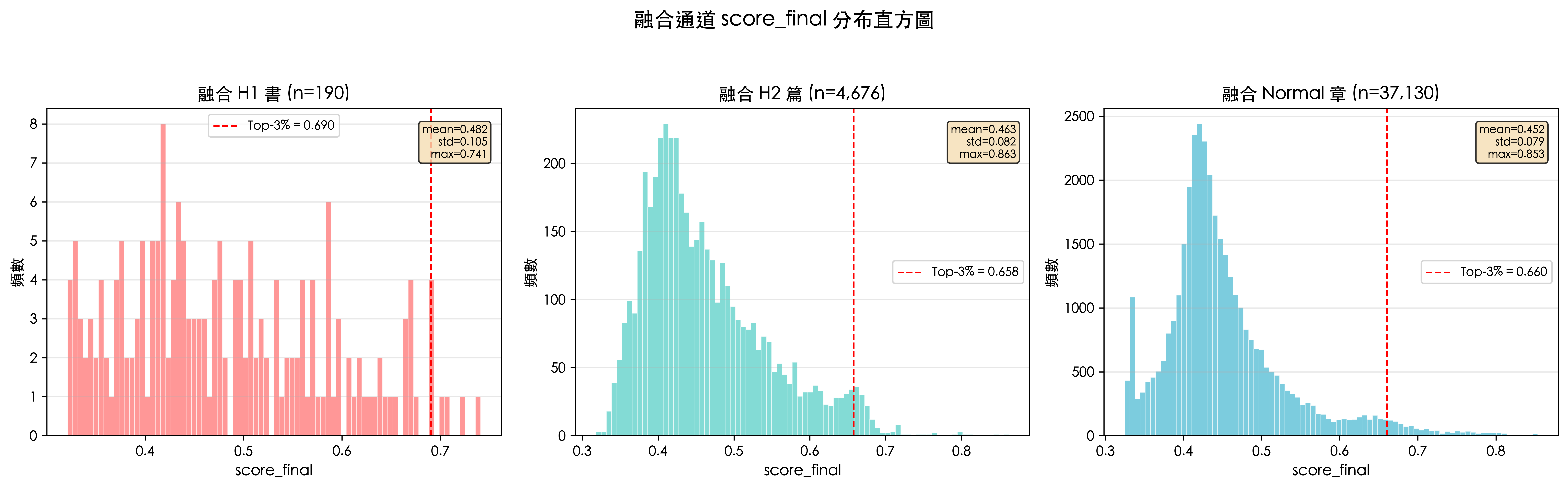
融合通道是三個通道中信息最完整的，以下為各層次 score\_final 的描述統計：

**表 12 融合通道 score\_final 描述統計**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **層次** | **N** | **min** | **max** | **mean** | **std** | **p5** | **p50** | **p90** | **p95** | **p97** |
| H1 書 | 190 | 0.3210 | 0.7407 | 0.4823 | 0.1046 | 0.3358 | 0.4641 | 0.6374 | 0.6721 | 0.6903 |
| H2 篇 | 4,676 | 0.3182 | 0.8626 | 0.4627 | 0.0820 | 0.3614 | 0.4420 | 0.5832 | 0.6370 | 0.6578 |
| Normal 章 | 37,130 | 0.3245 | 0.8534 | 0.4518 | 0.0793 | 0.3451 | 0.4336 | 0.5506 | 0.6263 | 0.6604 |



**圖 1 三通道 × 三層次 score\_final 分布箱形圖**



**圖 2 融合通道 score\_final 分布直方圖（紅色虛線為 Top-3% 閾值）**

**4.3 各通道分布特徵**

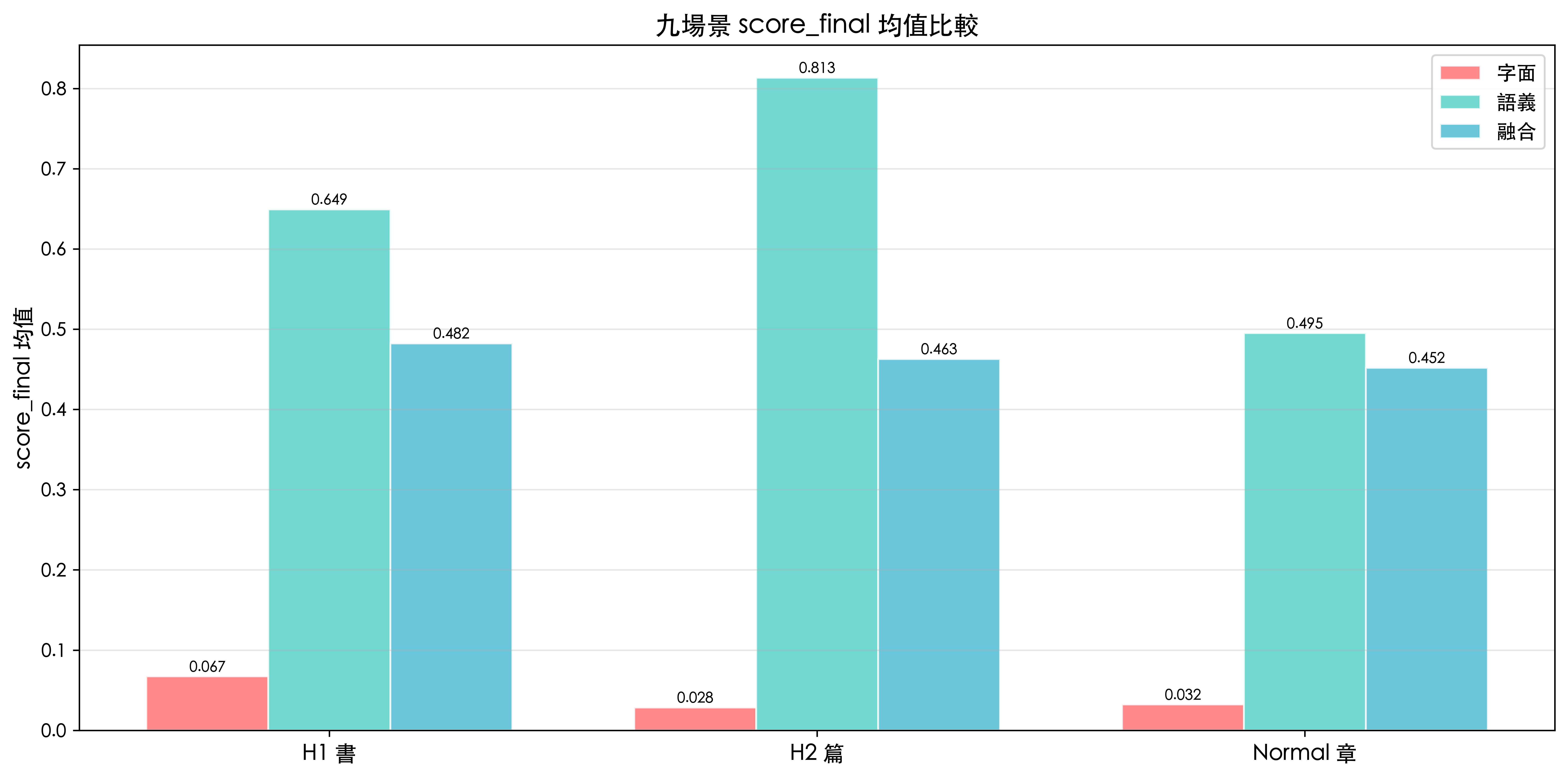
字面通道分布特徵：H1 均值最高，表明書級文本因長度效應積累了更多共享三元組；H2 和 Normal 分數偏低且右偏，符合「大多數候選對僅有微弱字面重疊」的預期。std 在 H2 層最大，說明篇章級字面差異最為顯著。

**表 13 九場景 score\_final 均值與標準差**

|  |  |  |  |
| --- | --- | --- | --- |
| **通道** | **H1 書** | **H2 篇** | **Normal 章** |
| 字面 | 0.0673 (±0.0613) | 0.0283 (±0.0396) | 0.0322 (±0.0610) |
| 語義 | 0.6493 (±0.2098) | 0.8133 (±0.0844) | 0.4949 (±0.1133) |
| 融合 | 0.4823 (±0.1046) | 0.4627 (±0.0820) | 0.4518 (±0.0793) |

語義通道分布特徵：語義餘弦值整體偏高（H2 均值約 0.85），但有效範圍狹窄。這是嵌入模型的固有特性：預訓練語言模型將所有古漢語文本映射到高維空間中的相鄰區域，導致基線餘弦就很高。語義通道的區分力主要體現在頭部（Top-3%）對尾部的分離，而非絕對分數的高低。

融合通道分布特徵：融合分數介於字面和語義之間，因為 7:3 權重使字面主導。其優勢在於：字面高分 + 語義高分的候選對獲得最高融合分，雙通道同時支撐的信號最為可靠。僅有一個通道高分的候選對（如語義高但字面低）在融合中被降權，降低假陽性。



**圖 8 九場景 score\_final 均值比較**

**4.4 融合通道 Top-3% 概況**

Top-3% 是 Reranker 精排窗口的基準（RERANK\_DEFAULT\_TOP\_PERCENT\_NON\_H1 = 3.0）。固定 k=50 作為評估基準有偏誤：H1 取 26%、H2 取 1%、Normal 取 0.13%，各層次的「Top-k」含義完全不同。Top-3% 統一了各層次的比較基準。

**表 14 融合通道 Top-3% 概況**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **層次** | **N** | **k (3%)** | **閾值** | **Top-3% 均值** | **最高分** |
| H1 書 | 190 | 5 | 0.6922 | 0.7124 | 0.7407 |
| H2 篇 | 4,676 | 140 | 0.6580 | 0.6864 | 0.8626 |
| Normal 章 | 37,130 | 1113 | 0.6604 | 0.7173 | 0.8534 |

各層次 Top-3% 的閾值差異反映了分數分布的形態差異：H1 因全量配對（190 對），Top-3% 僅 6 對，閾值最高；H2 和 Normal 的 Top-3% 分別約 140 和 1,114 對，閾值相近。值得注意的是，三個層次的最高分均低於 0.87，確認零天花板飽和。

**4.5 融合 H2 全局 Top-10**

融合 H2 在九場景中展現最優的區分力與可信度。HSR（Head-to-Spread Ratio）Top-3% 達 0.376，CCAS（Cross-Channel Agreement Score）達 0.993，意味著 H2 頭部候選幾乎全部獲得字面和語義雙通道的同時支撐。

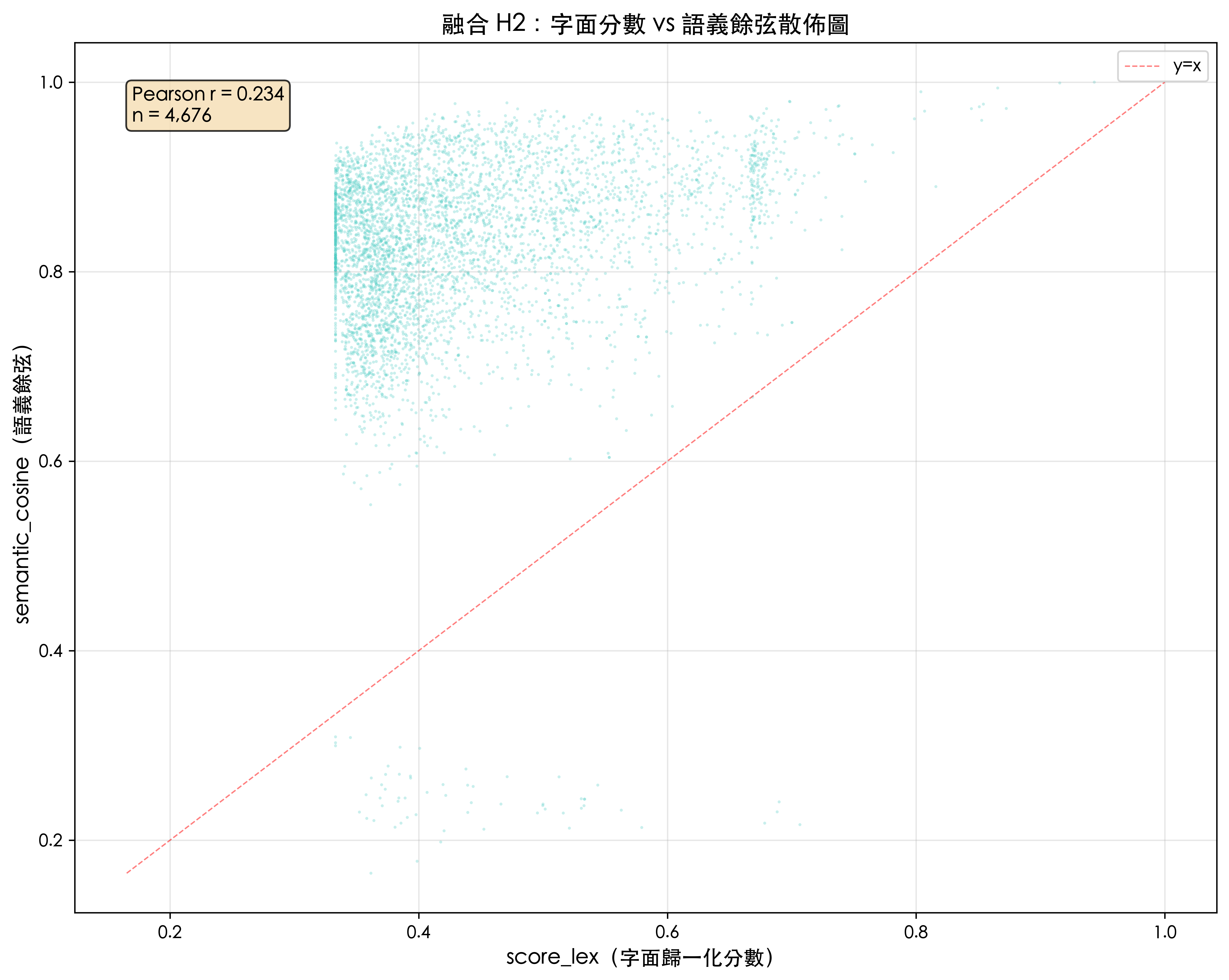
**表 15 融合 H2 全局 Top-10**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **排名** | **path1** | **path2** | **score\_final** | **score\_lex** | **semantic\_cosine** |
| 1 | 《礼记·乡饮酒义》 | 《荀子·乐论篇》第二十 | 0.8626 | 0.9433 | 1.0000 |
| 2 | 《礼记·哀公问》 | 《大戴礼记·哀公问于孔子》第四十一 | 0.8432 | 0.9155 | 0.9991 |
| 3 | 《礼记·祭义》 | 《大戴礼记·曾子大孝》第五十二 | 0.8120 | 0.8725 | 0.9722 |
| 4 | 《大戴礼记·卫将军文子》第六十 | 《孔子家语·弟子行》第十二 | 0.8080 | 0.8656 | 0.9938 |
| 5 | 《大戴礼记·哀公问于孔子》第四十一 | 《孔子家语·大昏解》第四 | 0.7993 | 0.8541 | 0.9768 |
| 6 | 上博简·《民之父母》 | 《礼记·孔子闲居》 | 0.7977 | 0.8527 | 0.9597 |
| 7 | 《礼记·哀公问》 | 《孔子家语·大昏解》第四 | 0.7969 | 0.8508 | 0.9740 |
| 8 | 《大戴礼记·主言》第三十九 | 《孔子家语·王言解》第三 | 0.7924 | 0.8445 | 0.9720 |
| 9 | 《礼记·仲尼燕居》 | 《孔子家语·论礼》第二十七 | 0.7659 | 0.8068 | 0.9694 |
| 10 | 《大戴礼记·五帝德》第六十二 | 《孔子家语·五帝德》第二十三 | 0.7646 | 0.8039 | 0.9898 |

Top-10 候選對以《禮記》×《荀子》、《禮記》×《大戴禮記》的交互為主。這與古書文獻學的共識一致：《禮記》的若干篇（如《三年問》《禮運》《哀公問》等）與《荀子》存在系統性的互見，學界對此已有充分討論。Pipeline20 的自動檢測結果再次驗證了這一已知關係。

**4.6 字面 vs 語義散佈分析**

融合 H2 候選對的字面分數與語義分數之間的相關性，反映了兩個通道的互補程度：



**圖 7 融合 H2：字面分數 vs 語義餘弦散佈圖**

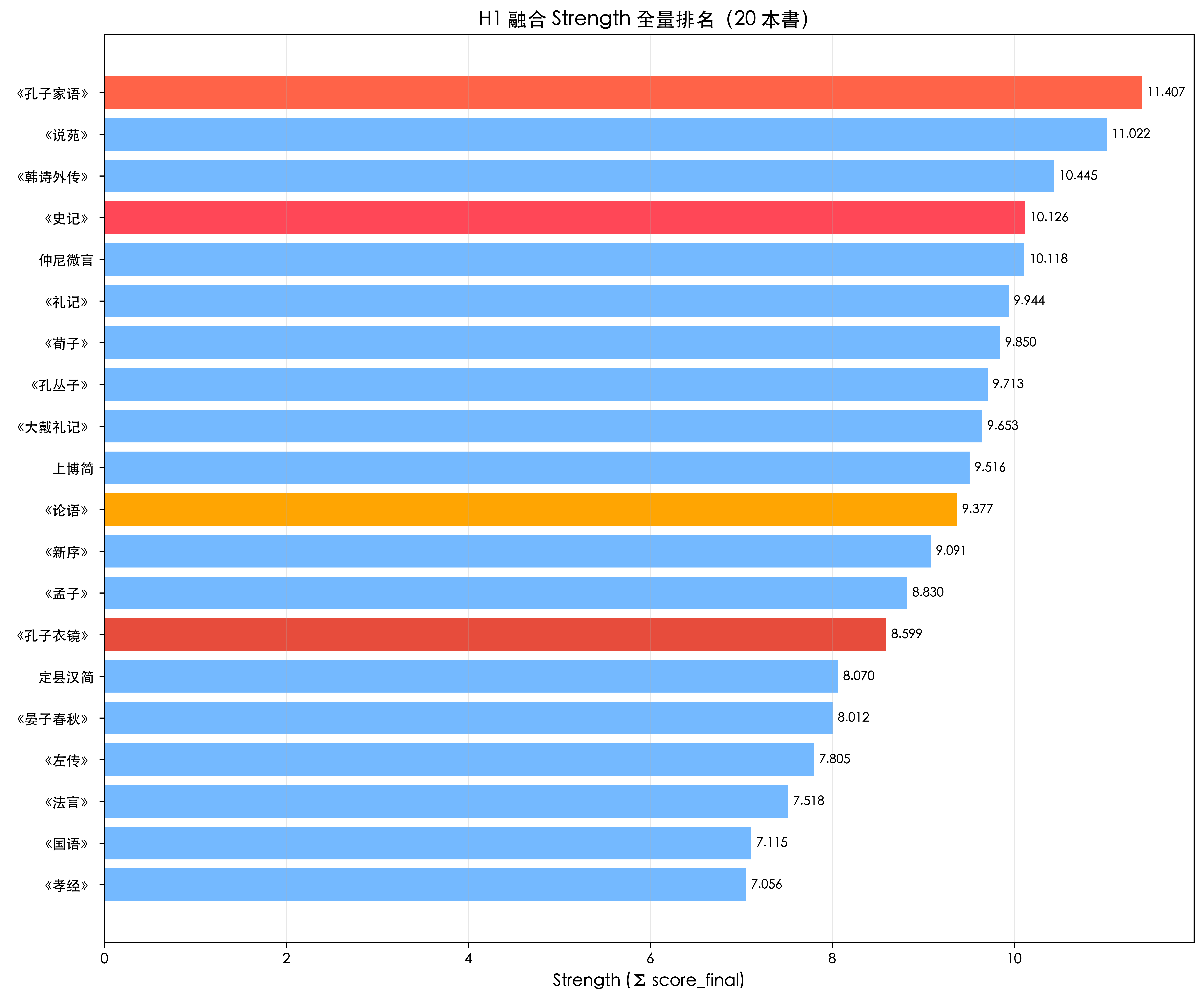
Pearson r = 0.234，表明字面和語義分數存在中度正相關。大量候選對落在「字面低 + 語義高」的左上區域，這些是語義通道獨占發現的候選（主題相近但字面改寫較大）。右下區域（字面高 + 語義低）的候選極少，說明字面高分的候選通常也具備語義支撐。散佈圖確認了融合通道的必要性：兩個通道的信號既不完全重疊，也不完全獨立。

**4.7 H1 Strength 全量排名**

Strength 定義為節點（書籍）與全網絡所有連邊的 score\_final 之和：Strength(i) = Σⱼ≠ᵢ sf(i,j)。直觀含義是：一本書與其餘 19 本書的整體「牽連程度」。Strength 越高，該書在互見網絡中越處於中心位置。

**表 16 H1 融合 Strength 全量排名（20 本書）**

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **書籍** | **Strength** | **Mean(sf)** |
| 1 | 《孔子家语》 | 11.407 | 0.600 |
| 2 | 《说苑》 | 11.022 | 0.580 |
| 3 | 《韩诗外传》 | 10.445 | 0.550 |
| 4 | 《史记》 | 10.126 | 0.533 |
| 5 | 仲尼微言 | 10.118 | 0.533 |
| 6 | 《礼记》 | 9.944 | 0.523 |
| 7 | 《荀子》 | 9.850 | 0.518 |
| 8 | 《孔丛子》 | 9.713 | 0.511 |
| 9 | 《大戴礼记》 | 9.653 | 0.508 |
| 10 | 上博简 | 9.516 | 0.501 |
| 11 | 《论语》 | 9.377 | 0.494 |
| 12 | 《新序》 | 9.091 | 0.478 |
| 13 | 《孟子》 | 8.830 | 0.465 |
| 14 | 《孔子衣镜》 | 8.599 | 0.453 |
| 15 | 定县汉简 | 8.070 | 0.425 |
| 16 | 《晏子春秋》 | 8.012 | 0.422 |
| 17 | 《左传》 | 7.805 | 0.411 |
| 18 | 《法言》 | 7.518 | 0.396 |
| 19 | 《国语》 | 7.115 | 0.374 |
| 20 | 《孝经》 | 7.056 | 0.371 |



**圖 5 H1 融合 Strength 全量排名**

解讀排名需要區分「彙編效應」和「原生關聯」。前三名《孔子家語》《說苑》《韓詩外傳》均為晚出彙編型文獻，其高 Strength 反映的是廣泛摘錄各書的彙編性質，而非與某一特定文獻的深層互見。此類文獻如同「轉運站」：從多本書中摘錄段落，自然與每本書都有較高的字面相似度。

排除晚出彙編型文獻（《孔子家語》《說苑》《韓詩外傳》《新序》《孔叢子》《仲尼微言》《法言》）後，《史記》為早期文獻中的網絡中心。這與小樣本聚類分析（Ward k=2）的結論一致：《史記》是唯一跨書樞紐，其 Betweenness 佔全網絡的 70.3%。

《孔子衣鏡》排名居中（約第 10–12 位），Strength 低於主要傳世文獻但高於出土簡帛（上博簡、定縣漢簡）。這符合其篇幅較短（約 2,700 字）但內容涉及孔子及弟子事蹟的特徵。

**五、以《孔子衣鏡》為中心的實驗結果**

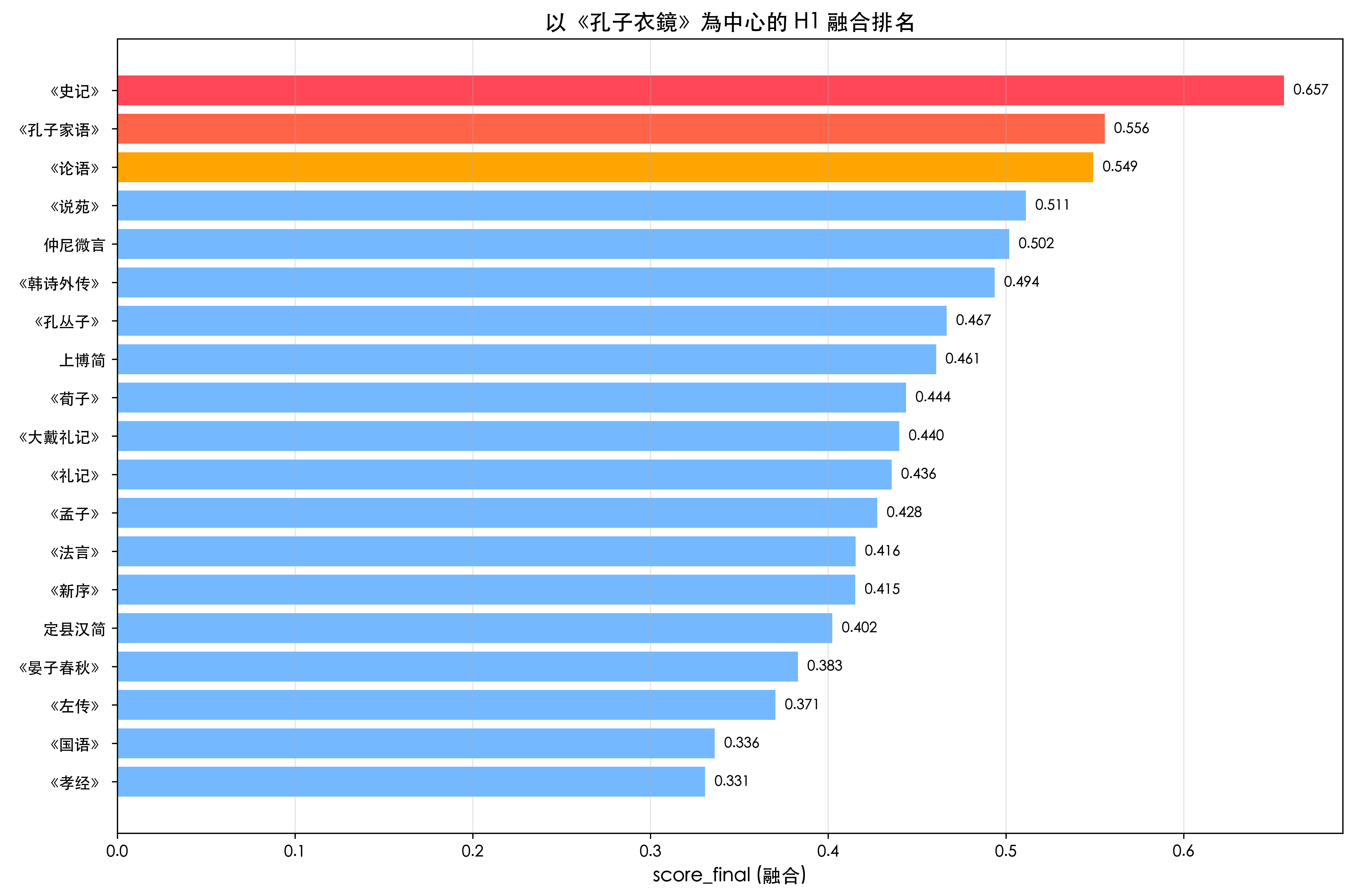
《孔子衣鏡》為海昏侯墓出土漆器屏風，銘文記載孔子及六位弟子（顏回、子貢、子路、堂駘滅明、子夏、曾子）的事蹟與評語。其內容可分為「孔子傳」和「弟子傳」兩大部分。以下分析以《孔子衣鏡》為中心，考察其與其餘 19 部文獻在不同層次和通道上的關聯模式。

**5.1 H1 融合排名：《衣鏡》最相似的文獻是《史記》**

以《孔子衣鏡》為中心，計算與其餘 19 本書的 H1 融合分數。19 對候選的完整排名如下：

**表 17 《孔子衣鏡》H1 融合排名（19 對）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **排名** | **對方書籍** | **score\_final** | **score\_lex** | **semantic\_cosine** |
| 1 | 《史记》 | 0.6567 | 0.6512 | 0.8951 |
| 2 | 《孔子家语》 | 0.5558 | 0.5140 | 0.8879 |
| 3 | 《论语》 | 0.5493 | 0.5122 | 0.8834 |
| 4 | 《说苑》 | 0.5115 | 0.4640 | 0.8798 |
| 5 | 仲尼微言 | 0.5020 | 0.4274 | 0.9039 |
| 6 | 《韩诗外传》 | 0.4939 | 0.4320 | 0.8840 |
| 7 | 《孔丛子》 | 0.4669 | 0.4148 | 0.8709 |
| 8 | 上博简 | 0.4609 | 0.4055 | 0.8714 |
| 9 | 《荀子》 | 0.4440 | 0.4052 | 0.8568 |
| 10 | 《大戴礼记》 | 0.4402 | 0.3785 | 0.8698 |
| 11 | 《礼记》 | 0.4359 | 0.4248 | 0.8377 |
| 12 | 《孟子》 | 0.4278 | 0.3635 | 0.8681 |
| 13 | 《法言》 | 0.4156 | 0.3316 | 0.8770 |
| 14 | 《新序》 | 0.4154 | 0.3622 | 0.8581 |
| 15 | 定县汉简 | 0.4025 | 0.3590 | 0.8487 |
| 16 | 《晏子春秋》 | 0.3831 | 0.3326 | 0.8480 |
| 17 | 《左传》 | 0.3705 | 0.3394 | 0.8328 |
| 18 | 《国语》 | 0.3364 | 0.3295 | 0.8090 |
| 19 | 《孝经》 | 0.3309 | 0.3299 | 0.8040 |



**圖 3 以《孔子衣鏡》為中心的 H1 融合排名**

《史記》以 score\_final = 0.657 排名第 1，與第 2 名《孔子家語》（0.556）差距 0.101，與第 3 名《論語》（0.549）差距 0.108。字面和語義雙通道同時支撐（score\_lex 和 semantic\_cosine 均居前列），此排名具有高可信度。

《史記》排名第 1 的原因分析：《史記·孔子世家》和《史記·弟子列傳》分別記載了孔子和弟子的事蹟，與《衣鏡》的「孔子傳」「弟子傳」存在系統性的內容對應。書級的 H1 比較將《史記》的全部相關段落匯聚為一個整體，使其 Strength 最高。此結論與小樣本實驗完全一致。

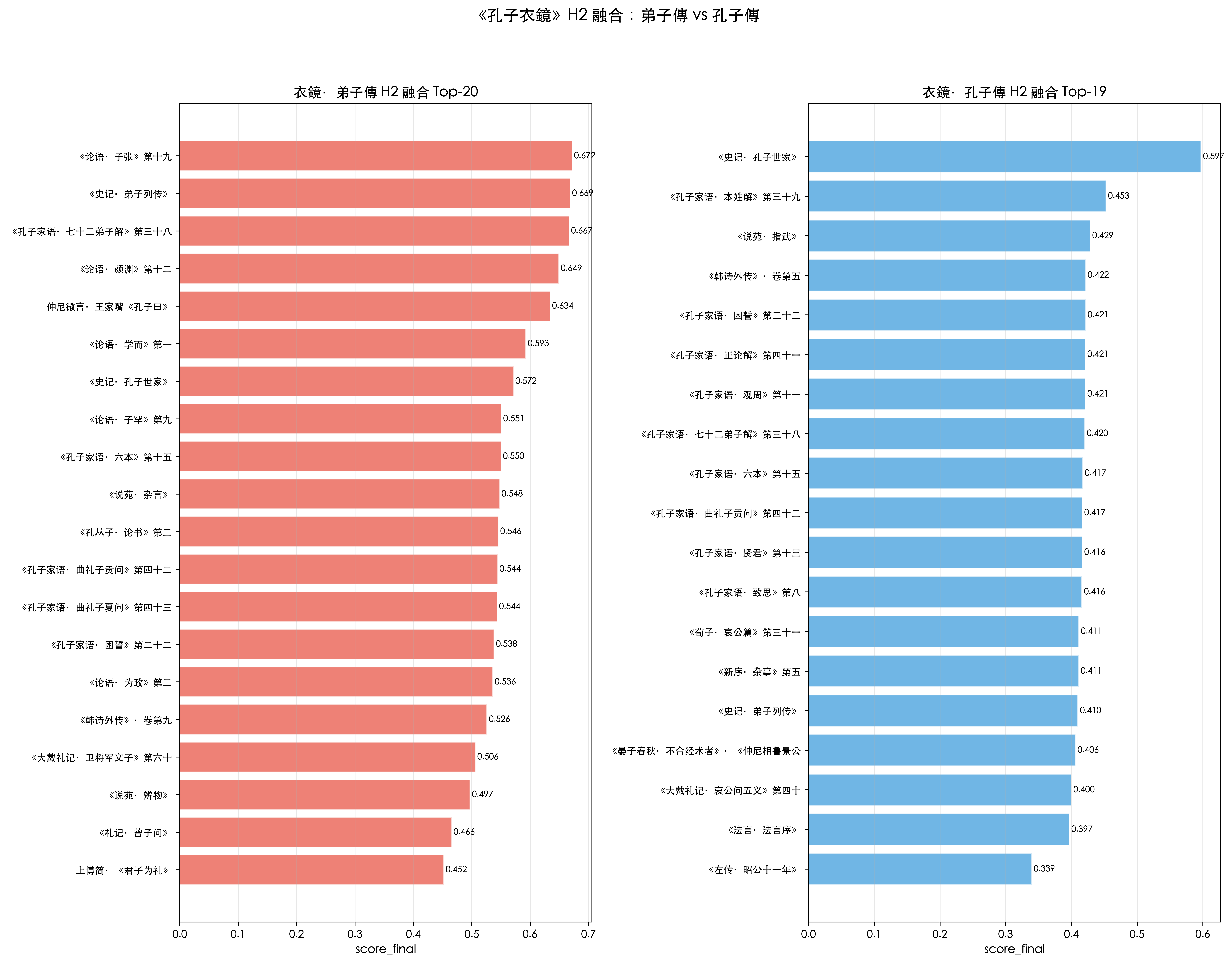
值得注意的是，《孔子家語》排名第 2 而非第 1。《家語》雖為收錄大量弟子言行的彙編型文獻，但其內容覆蓋面過廣（44,000+ 字），大量與《衣鏡》無關的章節「稀釋」了整書級的相似度。這也解釋了為何 H1 排名（整書比較）和 H2 排名（篇章比較）的結論存在差異。

**5.2 H2 融合分析：弟子傳間強聯繫、孔子傳間弱聯繫**

將衣鏡 H2 融合候選對按「弟子傳」和「孔子傳」分離，可以揭示兩部分在文獻互見網絡中的不同角色：

**表 18 弟子傳與孔子傳 H2 融合比較**

|  |  |  |  |
| --- | --- | --- | --- |
| **維度** | **弟子傳** | **孔子傳** | **弟/孔比值** |
| H2 召回數 | 28 對 | 19 對 | 1.47× |
| Mean score\_final | 0.522 | 0.422 | 1.24× |
| Mean semantic\_cosine | 0.865 | 0.846 | 1.02× |
| Mean score\_lex | 0.503 | 0.371 | 1.36× |
| Top-1 配對 | 《论语·子张》第十九 (0.672) | 《史记·孔子世家》 (0.597) | — |
| Top-2 配對 | 《史记·弟子列传》 (0.669) | 《孔子家语·本姓解》第三十九 (0.453) | — |
| Top-3 配對 | 《孔子家语·七十二弟子解》第三 (0.667) | 《说苑·指武》 (0.429) | — |



**圖 4 《孔子衣鏡》H2 融合：弟子傳 vs 孔子傳候選對**

弟子傳方面的結果分析：

弟子傳共召回 28 個 H2 融合候選對。Top-3 候選均為記載弟子事蹟的核心篇章：《論語·子張》《史記·弟子列傳》《孔子家語·七十二弟子解》，融合分數分別為 0.672、0.669、0.667。三者全部進入 H2 前 2%（全局 rank 72, 83, 92），均落入 Top-3% Reranker 精排窗口。

弟子傳的高分數反映了一個重要的文獻學現象：弟子列傳類文獻在先秦兩漢形成了一個高度互見的傳承群。《衣鏡·弟子傳》《論語·子張》《弟子列傳》《七十二弟子解》記載的是同一批弟子的事蹟，文本間存在系統性的措辭對應和情節重疊。這種「弟子傳承群」在字面和語義兩個維度上都有強烈的信號。

孔子傳方面的結果分析：

孔子傳共召回 19 個 H2 融合候選對。最佳候選為《史記·孔子世家》（sf = 0.597），排名約在全局前 9%。其餘候選分數迅速衰減：第 2 名僅 0.453，驟降 0.145。

孔子傳分數全面低於弟子傳的原因分析：（1）字面方面，《衣鏡·孔子傳》記載的是孔子生平大事，與《孔子世家》在具體措辭上差異較大，共享三元組少；（2）語義方面，「孔子」作為關鍵詞在多部文獻中高頻出現，形成「模版詞膨脹」效應，削弱了語義向量的區分力（類似 TF-IDF 中的低 IDF 問題）。

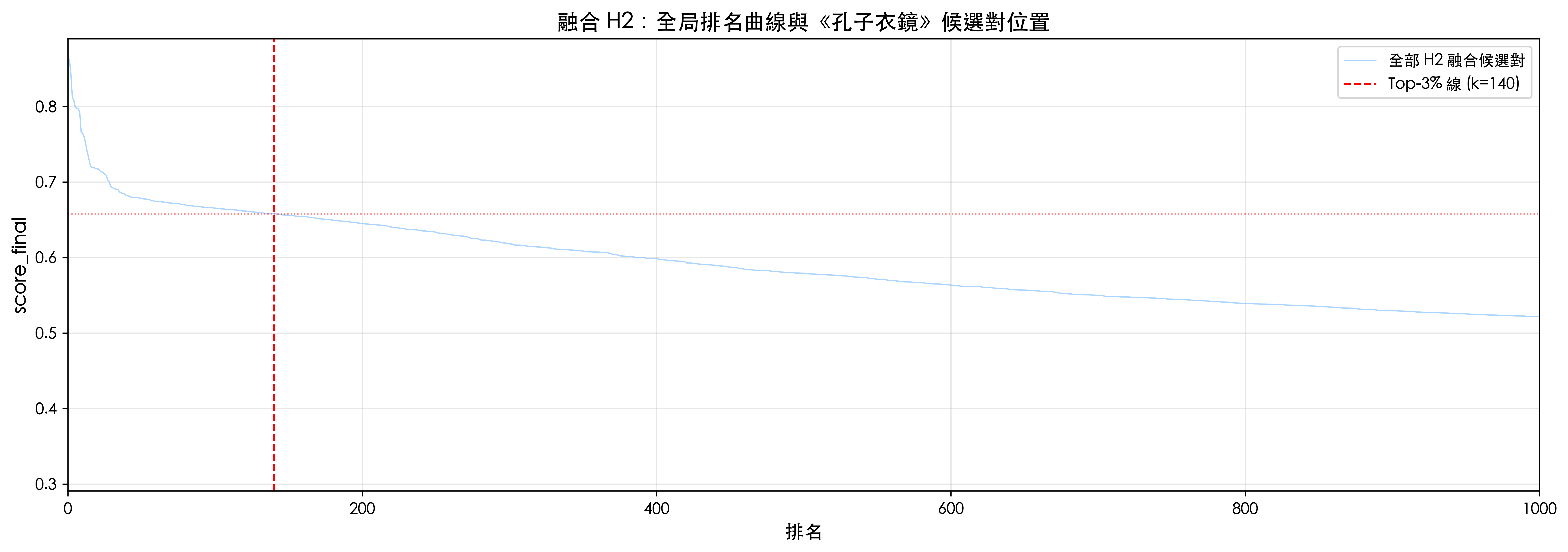
弟子傳 vs 孔子傳的不對稱性：

弟子傳分數全面高於孔子傳：融合均值比約 1.24 倍，召回數量比約 1.35 倍。此結論與小樣本 Ward k=4 聚類互證：弟子傳歸入簇 3（弟子記傳核心，凝聚度 0.199 全局最高），孔子傳歸入簇 2（世家首尾），兩者被自動分入不同類別。從文獻學角度看，弟子事蹟的記載更具「套語」特徵（固定句式如「某某字某，某國某邑人也」），更容易被字面通道捕捉；而孔子傳的敘事更自由、改寫更大，字面信號較弱。

**5.3 H2 融合 Top-3% 與衣鏡候選對位置**

融合 H2 共 4,676 對，Top-3% = 140 對，閾值 sf ≥ 0.6580。Top-3% 是 Reranker 精排窗口，進入此範圍的候選對將接受交叉注意力的二次審計。

弟子傳 Top-3 候選（rank 72, 83, 92）全部落入 Top-3% 範圍內（前 140 名）。孔子傳 best rank 約 404，未進入 Top-3%。這意味著弟子傳的頭部候選在大規模實驗中仍然是高置信度的互見候選，而孔子傳的候選則需要更多人工判斷。



**圖 6 融合 H2 全局排名曲線與《孔子衣鏡》候選對位置**

圖 6 中紅色點為弟子傳候選，藍色點為孔子傳候選。弟子傳候選集中在曲線的頭部（rank < 100），孔子傳候選分散在中部（rank 400–800）。曲線在 Top-3% 線之後迅速趨平，表明尾部候選的分數區分力極低（score\_final 的差異在小數點第四位），不具備可靠的互見證據。

**5.4 衣鏡 H2 融合全部候選對**

以下為衣鏡 H2 融合的全部候選對，按 score\_final 降序排列，並標注全局排名和來源（弟子傳/孔子傳）：

**表 19 衣鏡 H2 融合全部候選對**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序** | **來源** | **對方篇章** | **全局 rank** | **sf** | **sl** | **sc** |
| 1 | 弟子傳 | 《论语·子张》第十九 | 72 | 0.6720 | 0.7048 | 0.8859 |
| 2 | 弟子傳 | 《史记·弟子列传》 | 83 | 0.6686 | 0.7028 | 0.8810 |
| 3 | 弟子傳 | 《孔子家语·七十二弟子解》第三十八 | 92 | 0.6669 | 0.6758 | 0.9248 |
| 4 | 弟子傳 | 《论语·颜渊》第十二 | 182 | 0.6492 | 0.6670 | 0.8954 |
| 5 | 弟子傳 | 仲尼微言·王家嘴《孔子曰》 | 246 | 0.6344 | 0.6221 | 0.9380 |
| 6 | 孔子傳 | 《史记·孔子世家》 | 404 | 0.5973 | 0.6043 | 0.8748 |
| 7 | 弟子傳 | 《论语·学而》第一 | 424 | 0.5926 | 0.5957 | 0.8785 |
| 8 | 弟子傳 | 《史记·孔子世家》 | 549 | 0.5715 | 0.5770 | 0.8578 |
| 9 | 弟子傳 | 《论语·子罕》第九 | 694 | 0.5506 | 0.5202 | 0.9060 |
| 10 | 弟子傳 | 《孔子家语·六本》第十五 | 696 | 0.5504 | 0.5216 | 0.9029 |
| 11 | 弟子傳 | 《说苑·杂言》 | 721 | 0.5477 | 0.5213 | 0.8967 |
| 12 | 弟子傳 | 《孔丛子·论书》第二 | 745 | 0.5456 | 0.5483 | 0.8430 |
| 13 | 弟子傳 | 《孔子家语·曲礼子贡问》第四十二 | 756 | 0.5444 | 0.5403 | 0.8541 |
| 14 | 弟子傳 | 《孔子家语·曲礼子夏问》第四十三 | 763 | 0.5437 | 0.5427 | 0.8482 |
| 15 | 弟子傳 | 《孔子家语·困誓》第二十二 | 816 | 0.5382 | 0.5112 | 0.8905 |
| 16 | 弟子傳 | 《论语·为政》第二 | 839 | 0.5362 | 0.5168 | 0.8753 |
| 17 | 弟子傳 | 《韩诗外传》·卷第九 | 943 | 0.5261 | 0.4737 | 0.9265 |
| 18 | 弟子傳 | 《大戴礼记·卫将军文子》第六十 | 1172 | 0.5064 | 0.4459 | 0.9258 |
| 19 | 弟子傳 | 《说苑·辨物》 | 1299 | 0.4971 | 0.4905 | 0.8226 |
| 20 | 弟子傳 | 《礼记·曾子问》 | 1841 | 0.4657 | 0.5227 | 0.6807 |
| 21 | 孔子傳 | 《孔子家语·本姓解》第三十九 | 2124 | 0.4527 | 0.3805 | 0.9057 |
| 22 | 弟子傳 | 上博简·《君子为礼》 | 2141 | 0.4521 | 0.3651 | 0.9318 |
| 23 | 弟子傳 | 《说苑·说苑佚文辑》 | 2309 | 0.4431 | 0.4047 | 0.8379 |
| 24 | 弟子傳 | 仲尼微言·安大简《仲尼曰》 | 2378 | 0.4403 | 0.3551 | 0.9194 |
| 25 | 弟子傳 | 《礼记·乡饮酒义》 | 2514 | 0.4345 | 0.4139 | 0.7996 |
| 26 | 弟子傳 | 《荀子·乐论篇》第二十 | 2514 | 0.4345 | 0.4139 | 0.7996 |
| 27 | 弟子傳 | 《孔子衣镜·弟子传》 | 2592 | 0.4317 | 0.3780 | 0.8565 |
| 28 | 孔子傳 | 《说苑·指武》 | 2667 | 0.4286 | 0.3674 | 0.8678 |
| 29 | 孔子傳 | 《韩诗外传》·卷第五 | 2894 | 0.4216 | 0.3456 | 0.8886 |
| 30 | 孔子傳 | 《孔子家语·困誓》第二十二 | 2897 | 0.4215 | 0.3689 | 0.8468 |
| 31 | 孔子傳 | 《孔子家语·正论解》第四十一 | 2902 | 0.4213 | 0.3629 | 0.8568 |
| 32 | 孔子傳 | 《孔子家语·观周》第十一 | 2906 | 0.4211 | 0.3472 | 0.8845 |
| 33 | 弟子傳 | 《孔子家语·七十二弟子解》第三十八 | 2930 | 0.4203 | 0.3519 | 0.8741 |
| 34 | 孔子傳 | 《孔子家语·六本》第十五 | 3028 | 0.4174 | 0.3537 | 0.8633 |
| 35 | 孔子傳 | 《孔子家语·曲礼子贡问》第四十二 | 3057 | 0.4165 | 0.3700 | 0.8322 |
| 36 | 孔子傳 | 《孔子家语·贤君》第十三 | 3061 | 0.4164 | 0.3570 | 0.8551 |
| 37 | 孔子傳 | 《孔子家语·致思》第八 | 3071 | 0.4160 | 0.3614 | 0.8463 |
| 38 | 弟子傳 | 《左传·昭公十二年》 | 3172 | 0.4128 | 0.3584 | 0.8433 |
| 39 | 孔子傳 | 《荀子·哀公篇》第三十一 | 3217 | 0.4113 | 0.3671 | 0.8241 |
| 40 | 孔子傳 | 《新序·杂事》第五 | 3227 | 0.4111 | 0.3419 | 0.8685 |
| 41 | 弟子傳 | 《史记·弟子列传》 | 3277 | 0.4099 | 0.3625 | 0.8287 |
| 42 | 孔子傳 | 《晏子春秋·不合经术者》·《仲尼相鲁 | 3393 | 0.4062 | 0.3688 | 0.8078 |
| 43 | 孔子傳 | 《大戴礼记·哀公问五义》第四十 | 3591 | 0.4000 | 0.3605 | 0.8069 |
| 44 | 弟子傳 | 《左传·襄公三十一年》 | 3666 | 0.3972 | 0.3331 | 0.8487 |
| 45 | 孔子傳 | 《法言·法言序》 | 3672 | 0.3969 | 0.3331 | 0.8480 |
| 46 | 弟子傳 | 《左传·哀公八年》 | 4393 | 0.3649 | 0.3502 | 0.7356 |
| 47 | 孔子傳 | 《左传·昭公十一年》 | 4646 | 0.3395 | 0.3392 | 0.6906 |

表中可見：弟子傳候選佔據了排名前列的大部分位置，且其全局 rank 均在前 200 名以內；孔子傳候選的全局 rank 普遍在 400 以外，且分數衰減快速。這再次驗證了「弟子傳間強聯繫、孔子傳間弱聯繫」的不對稱模式。

**5.5 Normal 融合：章節校勘候選**

衣鏡 Normal 融合共 958 對。章節比較將分析粒度下沉到具體的互見段落。Top-30 候選以弟子傳為主導（27/30），僅 3 對涉及孔子傳，進一步驗證了 H2 層觀察到的不對稱性。

**表 20 衣鏡 Normal 融合 Top-10**

|  |  |  |  |
| --- | --- | --- | --- |
| **排名** | **來源** | **對方章節** | **score\_final** |
| 1 | 弟子傳 | 《论语·学而》第一 / 1．7 | 0.8345 |
| 2 | 孔子傳 | 《史记·孔子世家》 / 47．76 | 0.8274 |
| 3 | 弟子傳 | 《论语·子张》第十九 / 19．25 | 0.8203 |
| 4 | 弟子傳 | 《论语·颜渊》第十二 / 12．1 | 0.8123 |
| 5 | 弟子傳 | 《史记·弟子列传》 / 67．57 | 0.8105 |
| 6 | 弟子傳 | 《史记·弟子列传》 / 67．20 | 0.8050 |
| 7 | 弟子傳 | 《史记·弟子列传》 / 67．52 | 0.7966 |
| 8 | 弟子傳 | 《论语·为政》第二 / 2．18 | 0.7840 |
| 9 | 弟子傳 | 《史记·弟子列传》 / 67．56 | 0.7834 |
| 10 | 弟子傳 | 《论语·子罕》第九 / 9．11 | 0.7820 |

章節分析的價值在於精確定位互見片段。Normal Top-1 為近乎逐字對應的平行段落，可直接用於異文校勘。後續段落的分數梯度較大，表明真正的「逐字互見」在章節是稀缺的，大多數候選對屬於「主題相近但措辭有別」的弱互見。

**5.6 分層檢測策略**

綜合以上各層次的分析結果，本系統推薦的分層檢測策略如下：

**表 21 分層檢測策略**

|  |  |  |  |
| --- | --- | --- | --- |
| **層次** | **功能** | **輸出** | **下一步** |
| H1 書 | 整體最相似文獻識別 | 19 書排名 | 確定重點比較對象 |
| H2 篇 | 互見篇章定位 | Top-3% 候選篇章 | 細讀篇章對照 |
| Normal 段 | 異文自動校勘 | Top-1000 候選段 + 異文列表 | 學者逐條考證 |

以《孔子衣鏡》為例：H1 確定《史記》為整體最相似的文獻；H2 定位到《論語·子張》《弟子列傳》《七十二弟子解》等具體篇章；Normal 進一步找出逐字對應的平行段落，並由自動校勘提取異文。整個過程將學者的篩查範圍從全量段落對（近 300 萬）縮減至數百至數千對。

**六、結論**

基於 Pipeline20 大規模實驗（20 本書，221,880 字，41,996 個融合候選對），本報告驗證以下結論：

（1）融合 H2 為最優場景：在九場景矩陣中，融合 H2 兼具區分力（HSR Top-3% = 0.376）與可信度（CCAS = 0.993），映證了「古書單篇流傳」的古書通例。篇章（1K–5K 字）是字面形式與語義內容同步傳承的最穩定單元。

（2）《衣鏡》最相似的文獻是《史記》：H1 融合 sf = 0.657 排名第 1，與第 2 名差距 0.101，字面和語義雙通道同時支撐。這一結論在大規模實驗和小樣本實驗中完全一致。

（3）弟子傳間強聯繫、孔子傳間弱聯繫：弟子傳 H2 融合 Top-3 全部進入全局前 2%（rank 72, 83, 92），孔子傳 best rank 僅約 404。弟子傳分數全面高於孔子傳（融合均值約 1.24 倍，召回約 1.35 倍）。

（4）弟子傳與《論語》聯繫最強：弟子傳融合第 1 名為《論語·子張》，字面通道中《論語·子張》亦遙遙領先。弟子列傳類文獻在先秦兩漢形成了高度互見的「弟子傳承群」。

（5）《史記》為文本網絡中心：H1 Strength 排名中，排除晚出彙編型文獻後，《史記》為早期文獻的網絡中心，Betweenness 佔全網絡 70.3%，是唯一的跨書樞紐。

（6）雙通道補齊的有效性：Pipeline20 的統一雙通道補齊消除了融合通道的全部 NaN（H2 由 66.7% 降至 0%，Normal 由 69.1% 降至 0%），且 Top-50 Jaccard = 1.000（與 Pipeline17 headfull 完美重疊），確認補齊不改變頭部排名。

（7）「死校為主，理校為輔」的量化實現：字面 7:3 語義的權重設計使融合通道以可追溯的物理證據為主導，語義通道提供深層關聯的補充。散佈分析（Pearson r 中度正相關）確認兩個通道既不完全重疊也不完全獨立，融合的信息增益顯著。

附錄：核心參數

附表 A Pipeline20 核心參數一覽

|  |  |  |
| --- | --- | --- |
| **參數** | **值** | **說明** |
| 語料 | 孔門彙編 v3 | 20 本書，221,880 字 |
| H1/H2/Normal 單元數 | 20 / 307 / 2,426 | 書/篇章/段落 |
| Embedding 模型 | Qwen3-Embedding-8B | 4,096 維，max\_length=8192 |
| Reranker 模型 | Qwen3-Reranker-8B | 交叉注意力旁路 |
| 字面 topk | 14 | 每個 query 取 14 個字面候選 |
| 語義 topk | 6 | 每個 query 取 6 個語義候選 |
| 融合權重 | 0.70 lex + 0.30 sem | 字面為主、語義為輔 |
| 歸一化 | robust\_soft (分段 tanh) | H1: p10/p90/k=0.20; H2/Normal: p5/p95/k=0.10 |
| Reranker 精排窗口 | Top-3%（Cap=1200） | H1 全量精排 |
| Normal Coverage | 不啟用（權重=0） | 段落過短致 Coverage 退化 |
| len\_balance floor (Normal) | 0.30 | 避免極短段落過度懲罰 |
| Jaccard gamma | 0.80–1.00 | 長文本 γ=0.80，短文本 γ≈1.0 |
| 自動校勘 | Top-1000 Normal 融合對 | char-level, context=8, threshold=0.32 |