**三嵌入模型对照实验报告**

Qwen3-Embedding-8B vs gte-Qwen2-7B vs BGE-M3

Pipeline20 大规模语料 孔门汇编 对照实验分析

# 第一章 实验设计

## 1.1 实验目的

本实验旨在通过三种不同架构的嵌入模型对同一古汉语语料进行互见检测，验证 Qwen3-Embedding-8B 在古汉语文本相似性评估中的可信度。对照模型包括同系列前代模型 gte-Qwen2-7B 和不同架构的 BGE-M3，通过控制变量法隔离嵌入模型差异对检测结果的影响。

## 1.2 控制变量

实验严格控制以下变量:

(1) 字面通道: 三组实验使用完全相同的字面匹配算法 (cosine\_tf + coverage + tfidf + jaccard)，输出 100% 一致，确认实验仅隔离嵌入模型变量。

(2) 归一化参数: 统一使用 robust\_soft piecewise tanh 归一化 (k=0.10 for H2, endpoints p5/p95)。

(3) 语料: 统一使用"孔门汇编3"语料 (20 书, 223,186 字)。

(4) Pipeline: 统一使用 pipeline20 (dual-channel completion, 零 NaN)。

## 1.3 模型配置

|  |  |  |  |
| --- | --- | --- | --- |
| 配置项 | Qwen3-8B | gte-Qwen2-7B | BGE-M3 |
| 嵌入模型 | Qwen3-Embedding-8B | gte-qwen2-7b-instruct | BGE-M3 |
| 重排模型 | Qwen3-Reranker-8B | Qwen3-Reranker-8B | BGE-reranker-base |
| 池化策略 | avgpool | last-token pooling | CLS pooling |
| 嵌入维度 | 4096 | 3584 | 1024 |
| 参数量 | ~8B | ~7B | ~568M |
| 架构系列 | Qwen3 | Qwen2 | BERT (XLM-R) |

# 第二章 字面通道验证

字面通道完全不依赖嵌入模型，仅基于字符级匹配。三组实验的字面通道结果应完全一致，这是实验设计正确性的关键验证。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 层级 | 指标 | Qwen3-8B | Qwen2-7B | BGE-M3 |
| H1 | 候选对数 | 190 | 190 | 190 |
| H2 | 候选对数 | 5210 | 5210 | 5210 |
| Normal | 候选对数 | 40264 | 40264 | 40264 |
| H1 | score\_final 均值 | 0.067253 | 0.067253 | 0.067253 |
| H2 | score\_final 均值 | 0.028338 | 0.028338 | 0.028338 |

结论: 三组实验的字面通道候选对数量和分数完全一致 (Jaccard = 1.000)，确认实验设计正确隔离了嵌入模型变量。

# 第三章 语义通道深度对比

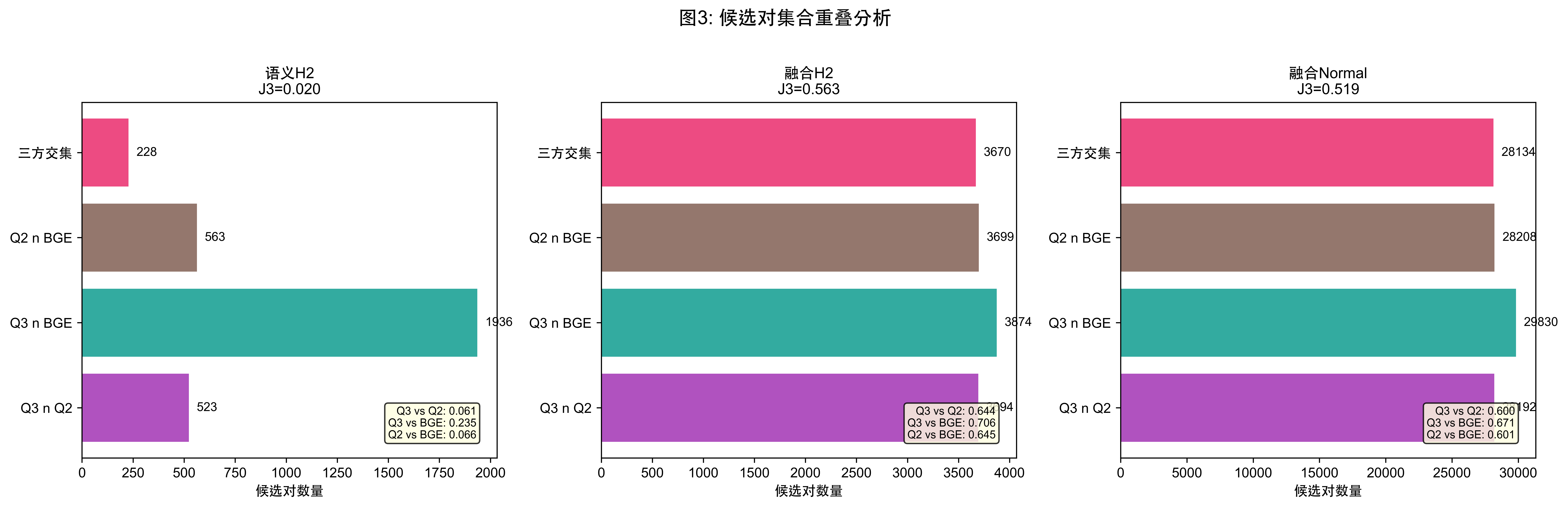
## 3.1 候选对集合分歧

三种嵌入模型在语义通道中产生了截然不同的候选对集合。下表展示各模型的候选对数量及交叉重叠情况。

语义H2: Qwen3=5064对, Qwen2=4032对, BGE=5115对; Q3 n Q2=523, Q3 n BGE=1936, Q2 n BGE=563, 三方交集=228对

融合H2: Qwen3=4676对, Qwen2=4756对, BGE=4682对; Q3 n Q2=3694, Q3 n BGE=3874, Q2 n BGE=3699, 三方交集=3670对

融合Normal: Qwen3=37124对, Qwen2=38021对, BGE=37155对; Q3 n Q2=28192, Q3 n BGE=29830, Q2 n BGE=28208, 三方交集=28134对

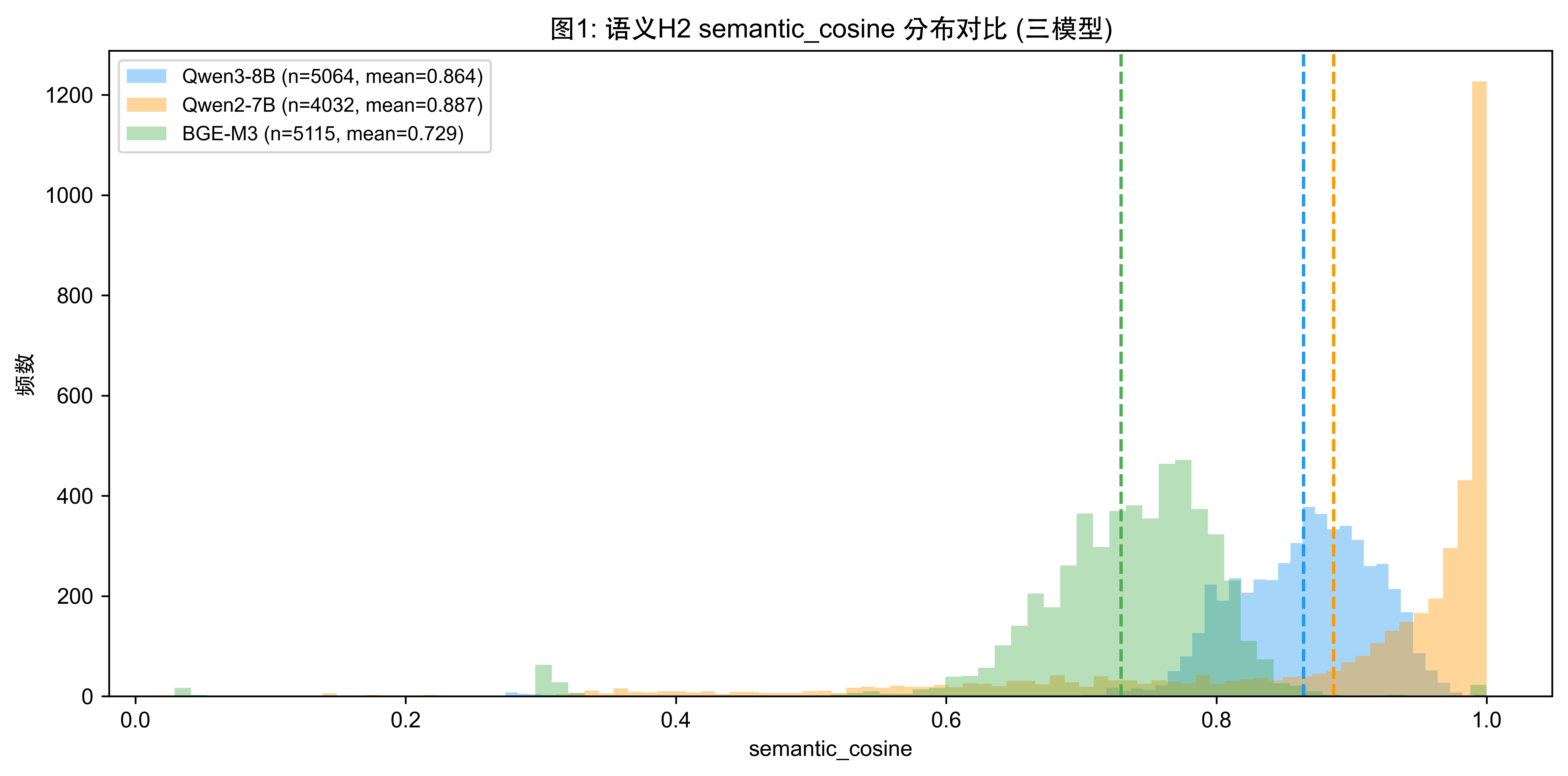


*图3: 候选对集合重叠分析*

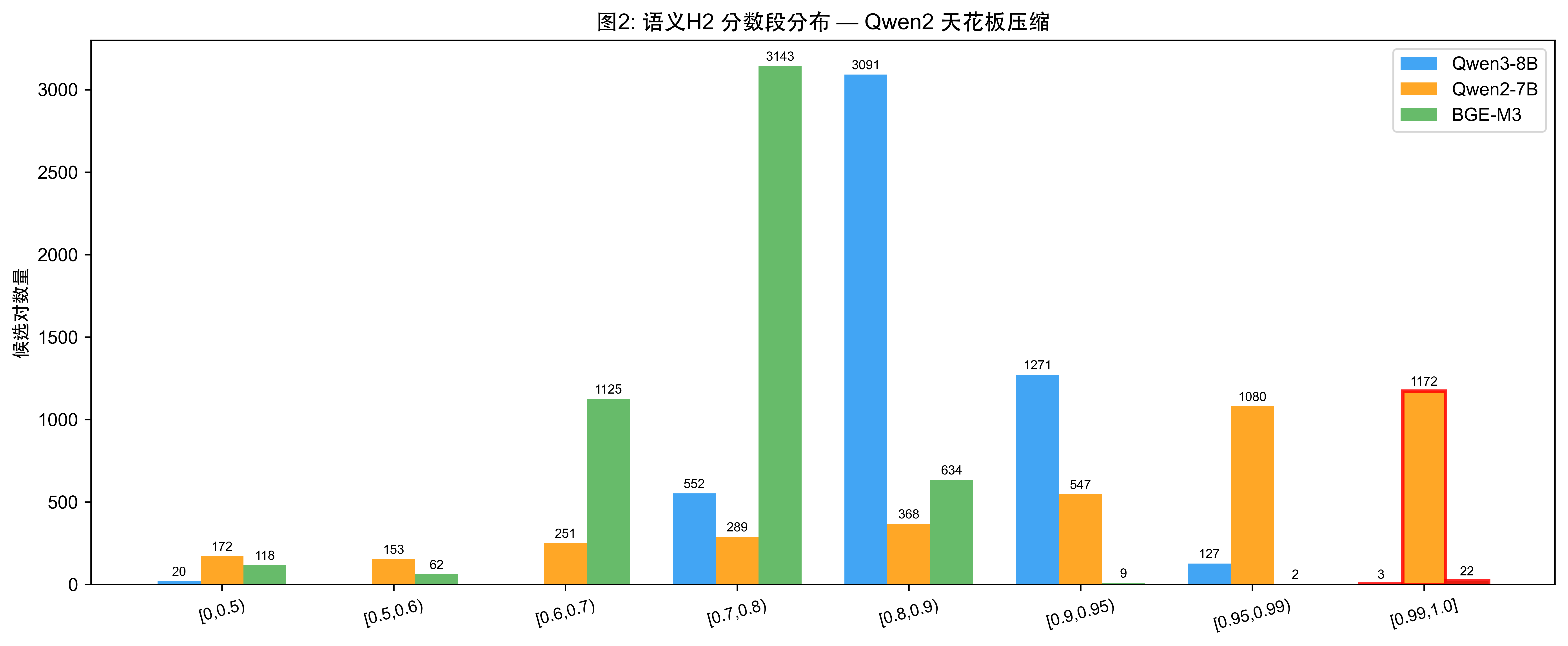
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 通道 | Q3 vs Q2 | Q3 vs BGE | Q2 vs BGE | 三方 Jaccard |
| 语义H2 | 0.0610 | 0.2349 | 0.0656 | 0.0200 |
| 融合H2 | 0.6438 | 0.7064 | 0.6445 | 0.5631 |
| 融合Normal | 0.6004 | 0.6711 | 0.6006 | 0.5190 |

关键发现: 语义H2的三方Jaccard仅为 0.020，说明三种模型在语义空间中构建了几乎完全不同的相似性邻域。同一文本对在不同模型中的语义相似度排名可能天差地别。

## 3.2 语义分数分布



*图1: 语义H2 semantic\_cosine 分布对比*

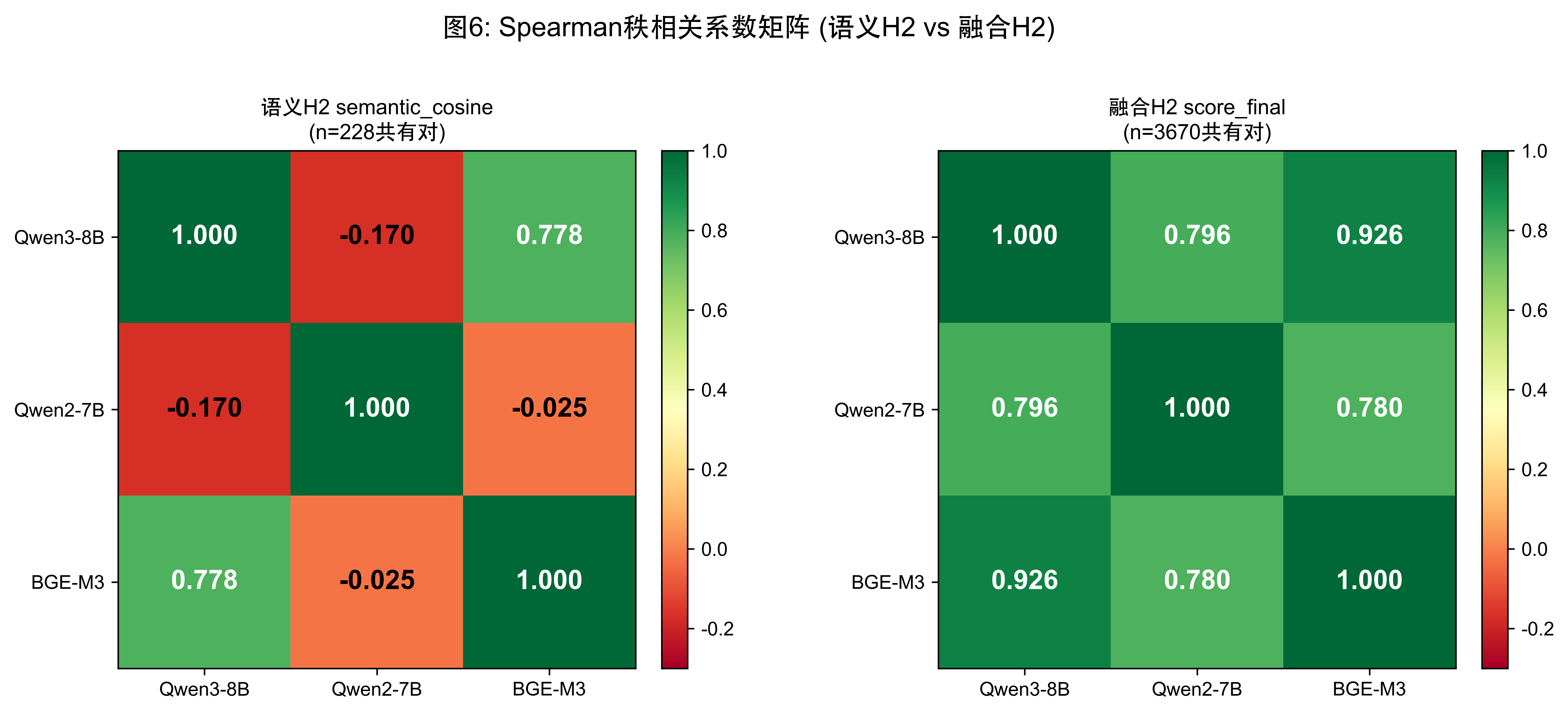


*图2: 语义H2 分数段分布*

|  |  |  |  |
| --- | --- | --- | --- |
| 统计量 | Qwen3-8B | Qwen2-7B | BGE-M3 |
| 候选对数 | 5064 | 4032 | 5115 |
| 均值 | 0.8644 | 0.8867 | 0.7294 |
| 标准差 | 0.0613 | 0.1620 | 0.0939 |
| 最小值 | 0.2738 | 0.1383 | 0.0290 |
| P5 | 0.7840 | 0.5322 | 0.6233 |
| 中位数 | 0.8707 | 0.9651 | 0.7432 |
| P95 | 0.9413 | 0.9986 | 0.8194 |
| 最大值 | 1.0000 | 1.0000 | 1.0000 |
| sc>=0.99 | 3 | 1172 | 22 |
| sc>=0.95 | 130 | 2252 | 24 |
| sc=1.0 | 1 | 1 | 22 |
| sc<=0 | 0 | 0 | 0 |

关键发现: Qwen2-7B 有 1172 对 (29.1%) 的 semantic\_cosine >= 0.99，构成灾难性天花板压缩。BGE-M3 有 22 对达到 sc=1.000 (CLS pooling 退化)。Qwen3-8B 仅 3 对 >= 0.99，分布健康。

## 3.3 语义排名相关性



*图6: Spearman 秩相关系数矩阵*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 通道 | Q3 vs Q2 | Q3 vs BGE | Q2 vs BGE | 共有对数 |
| 语义H2 (sc) | -0.1696 | 0.7784 | -0.0249 | 228 |
| 融合H2 (sf) | 0.7961 | 0.9257 | 0.7797 | 3670 |
| 语义Normal (sc) | 0.3555 | 0.7696 | 0.3353 | 425 |
| 融合Normal (sf) | 0.8470 | 0.9546 | 0.8461 | 28145 |

核心发现: Qwen3 vs Qwen2 在语义H2的 Spearman rho = -0.170。这意味着两个模型的语义排名呈负相关 -- Qwen2 认为相似的对，Qwen3 反而认为不相似。Qwen3 vs BGE 的 rho = 0.778，表明两者虽然绝对分数不同，但排名方向基本一致。

## 3.4 算法解释: 为什么 Qwen2 天花板压缩

语义相似度的计算公式为:

semantic\_cosine = cos(embed\_A, embed\_B)

其中 embed = pooling(model(tokens))。不同模型的池化策略决定了向量的信息密度:

-- Qwen3-8B (avgpool): 对所有 token 的隐藏状态取平均，信息均匀分布在 4096 维向量中，分辨率高。

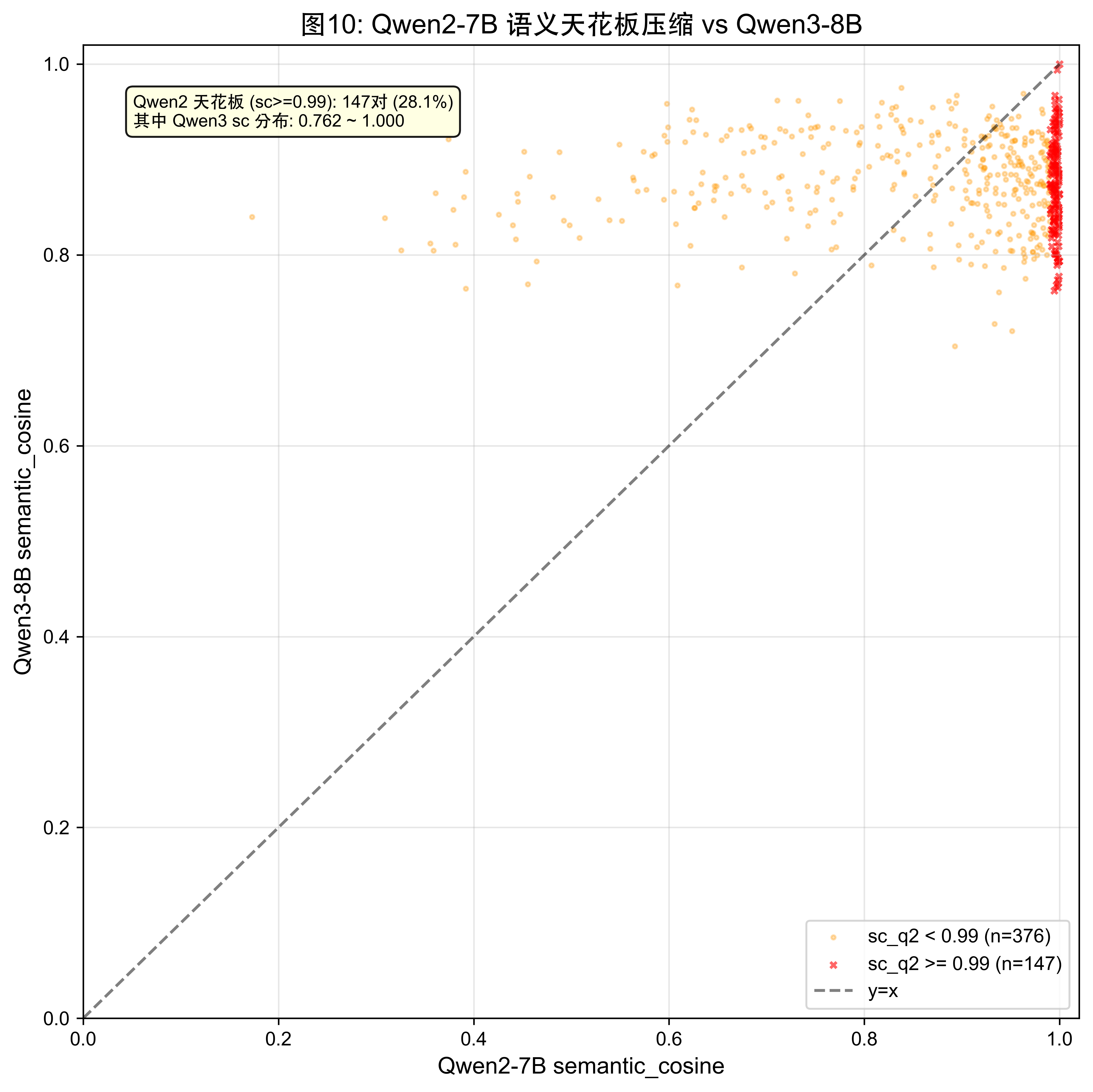
-- Qwen2-7B (last-token pooling): 仅取最后一个 token 的隐藏状态作为整句表示。在古汉语中，最后一个 token 可能编码了位置相关的信息，导致不同文本的向量趋于一致。

-- BGE-M3 (CLS pooling): 取 [CLS] token 的隐藏状态，维度仅 1024。对于某些重复模式的文本 (如礼记)，CLS 向量完全相同。

当大量候选对的 semantic\_cosine 聚集在 0.95 以上时，排名退化为噪声。Spearman rho 的计算公式为:

rho = 1 - 6 \* sum(d\_i^2) / (n \* (n^2 - 1))

其中 d\_i 为排名差。当分数几乎相同时，排名差 d\_i 受微小浮点波动支配，导致 rho 趋近 0 甚至为负。



*图10: Qwen2-7B 语义天花板压缩 vs Qwen3-8B*

## 3.5 语义 Top-10 对比

三种模型的语义H2 Top-10 候选对:

Qwen3-8B 语义H2 Top-10:

|  |  |  |  |
| --- | --- | --- | --- |
| path1 | path2 | rank | sc |
| 《礼记·乡饮酒义》 | 《荀子·乐论篇》第二十 | 1 | 1.0000 |
| 《礼记·哀公问》 | 《大戴礼记·哀公问于孔子》第四十一 | 2 | 0.9991 |
| 《大戴礼记·卫将军文子》第六十 | 《孔子家语·弟子行》第十二 | 3 | 0.9938 |
| 《大戴礼记·五帝德》第六十二 | 《孔子家语·五帝德》第二十三 | 4 | 0.9898 |
| 《孔子家语·曲礼子贡问》第四十二 | 《孔子家语·曲礼子夏问》第四十三 | 5 | 0.9797 |
| 《韩诗外传》·卷第七 | 《说苑·杂言》 | 6 | 0.9793 |
| 《论语·述而》第七 | 《论语·子罕》第九 | 7 | 0.9782 |
| 《论语·公冶长》第五 | 《论语·雍也》第六 | 8 | 0.9773 |
| 《大戴礼记·哀公问于孔子》第四十一 | 《孔子家语·大昏解》第四 | 9 | 0.9768 |
| 《说苑·杂言》 | 《孔子家语·六本》第十五 | 10 | 0.9751 |

Qwen2-7B 语义H2 Top-10:

|  |  |  |  |
| --- | --- | --- | --- |
| path1 | path2 | rank | sc |
| 《礼记·乡饮酒义》 | 《荀子·乐论篇》第二十 | 1 | 1.0000 |
| 《论语·阳货》第十七 | 《孔丛子·诘墨》第十八 | 2 | 0.9998 |
| 《论语·阳货》第十七 | 《大戴礼记·曾子制言上》第五十四 | 3 | 0.9998 |
| 《孔子家语·三恕》第九 | 《孔子家语·颜回》第十八 | 4 | 0.9998 |
| 《论语·宪问》第十四 | 《礼记·大学》 | 5 | 0.9998 |
| 《说苑·政理》 | 《孔子家语·好生》第十 | 6 | 0.9998 |
| 《大戴礼记·曾子制言上》第五十四 | 《孔丛子·诘墨》第十八 | 7 | 0.9998 |
| 《礼记·三年问》 | 《大戴礼记·曾子疾病》第五十七 | 8 | 0.9998 |
| 《孔丛子·杂训》第六 | 《孔丛子·儒服》第十三 | 9 | 0.9997 |
| 《说苑·君道》 | 《孔丛子·杂训》第六 | 10 | 0.9997 |

BGE-M3 语义H2 Top-10:

|  |  |  |  |
| --- | --- | --- | --- |
| path1 | path2 | rank | sc |
| 《礼记·檀弓上》 | 《礼记·檀弓下》 | 1 | 1.0000 |
| 《礼记·檀弓上》 | 《礼记·曾子问》 | 1 | 1.0000 |
| 《礼记·檀弓上》 | 《礼记·文王世子》 | 1 | 1.0000 |
| 《礼记·檀弓上》 | 《礼记·礼运》 | 1 | 1.0000 |
| 《礼记·檀弓上》 | 《礼记·礼器》 | 1 | 1.0000 |
| 《礼记·檀弓上》 | 《礼记·郊特牲》 | 1 | 1.0000 |
| 《礼记·檀弓下》 | 《礼记·曾子问》 | 1 | 1.0000 |
| 《礼记·檀弓下》 | 《礼记·文王世子》 | 1 | 1.0000 |
| 《礼记·檀弓下》 | 《礼记·礼运》 | 1 | 1.0000 |
| 《礼记·檀弓下》 | 《礼记·礼器》 | 1 | 1.0000 |

三方 Top-10 重叠: 0/10 对

## 3.6 语义 Normal 对比

|  |  |  |  |
| --- | --- | --- | --- |
| 统计量 | Qwen3-8B | Qwen2-7B | BGE-M3 |
| 候选对数 | 38612 | 31967 | 39377 |
| 均值 | 0.8110 | 0.9614 | 0.7148 |
| 标准差 | 0.0424 | 0.0901 | 0.0524 |
| P5 | 0.7390 | 0.7608 | 0.6269 |
| P95 | 0.8698 | 0.9993 | 0.7836 |
| sc>=0.99 | 44 | 16543 | 13 |

Qwen2-7B Normal: 16543对 (51.8%) sc>=0.99，均值 0.961，压缩极端。

# 第四章 融合通道对比 -- 字面权重的纠偏效应

## 4.1 融合公式

融合通道使用 robust\_soft piecewise tanh 归一化后的加权融合:

score\_final = 0.70 \* q\_lex + 0.30 \* q\_sem

其中: z = (x - p\_lo) / (p\_hi - p\_lo) [线性标准化]

z < 0: z = -tanh(-k\*z) [下限软压缩]

0 <= z <= 1: z = z [线性保持]

z > 1: z = 1 + tanh(k\*(z-1)) [上限软压缩]

q = eps + (z+1)/3 \* (1-2\*eps) [映射到 (0,1)]

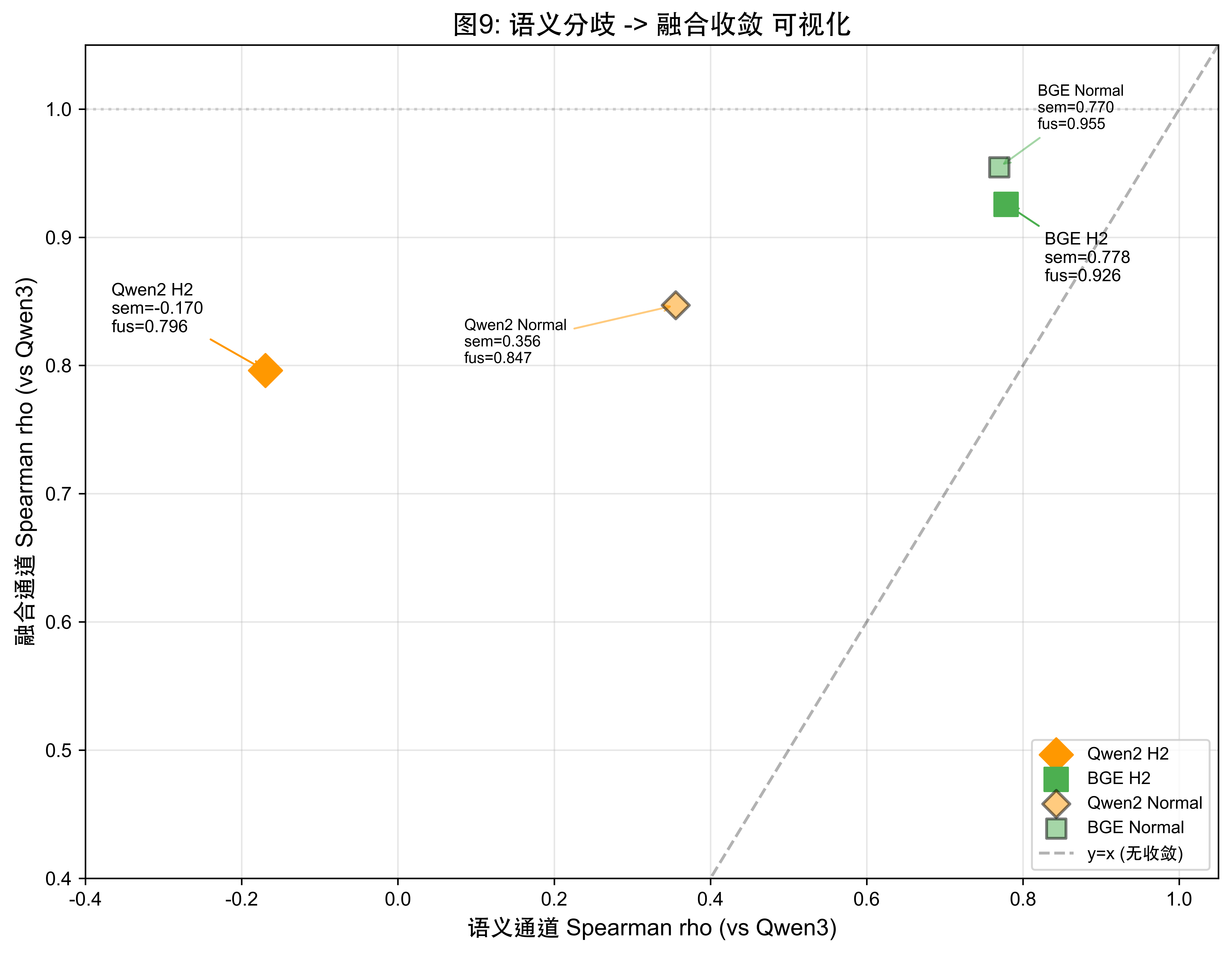
70% 的字面权重意味着融合分数主要由字面通道决定，语义差异被大幅稀释。

## 4.2 候选对收敛

语义H2 三方Jaccard = 0.0200 => 融合H2 三方Jaccard = 0.5631 (提升 28 倍)

## 4.3 排名收敛

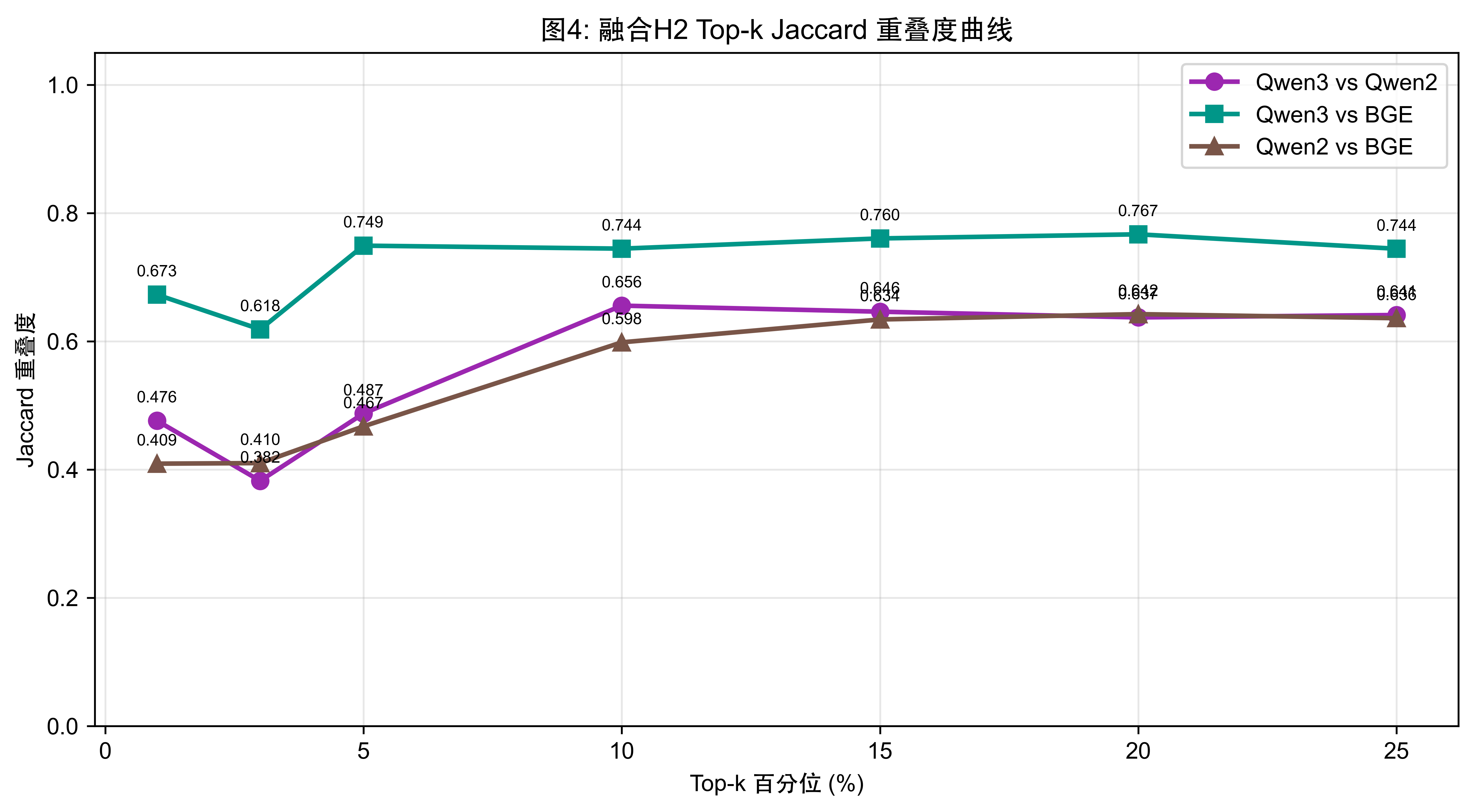
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型对 | 语义H2 rho | 融合H2 rho | 语义Normal rho | 融合Normal rho |
| Q3 vs Q2 | -0.1696 | 0.7961 | 0.3555 | 0.8470 |
| Q3 vs BGE | 0.7784 | 0.9257 | 0.7696 | 0.9546 |
| Q2 vs BGE | -0.0249 | 0.7797 | 0.3353 | 0.8461 |



*图9: 语义分歧 -> 融合收敛 可视化*

解读: 字面权重 (70%) 的主导地位有效消化了模型间的语义差异。即使 Qwen2 的语义排名与 Qwen3 呈相关性 rho=-0.170，融合后仍能达到 rho=0.796，证明融合机制具有强纠偏能力。

## 4.4 Top-k 一致性



*图4: 融合H2 Top-k Jaccard 重叠度曲线*

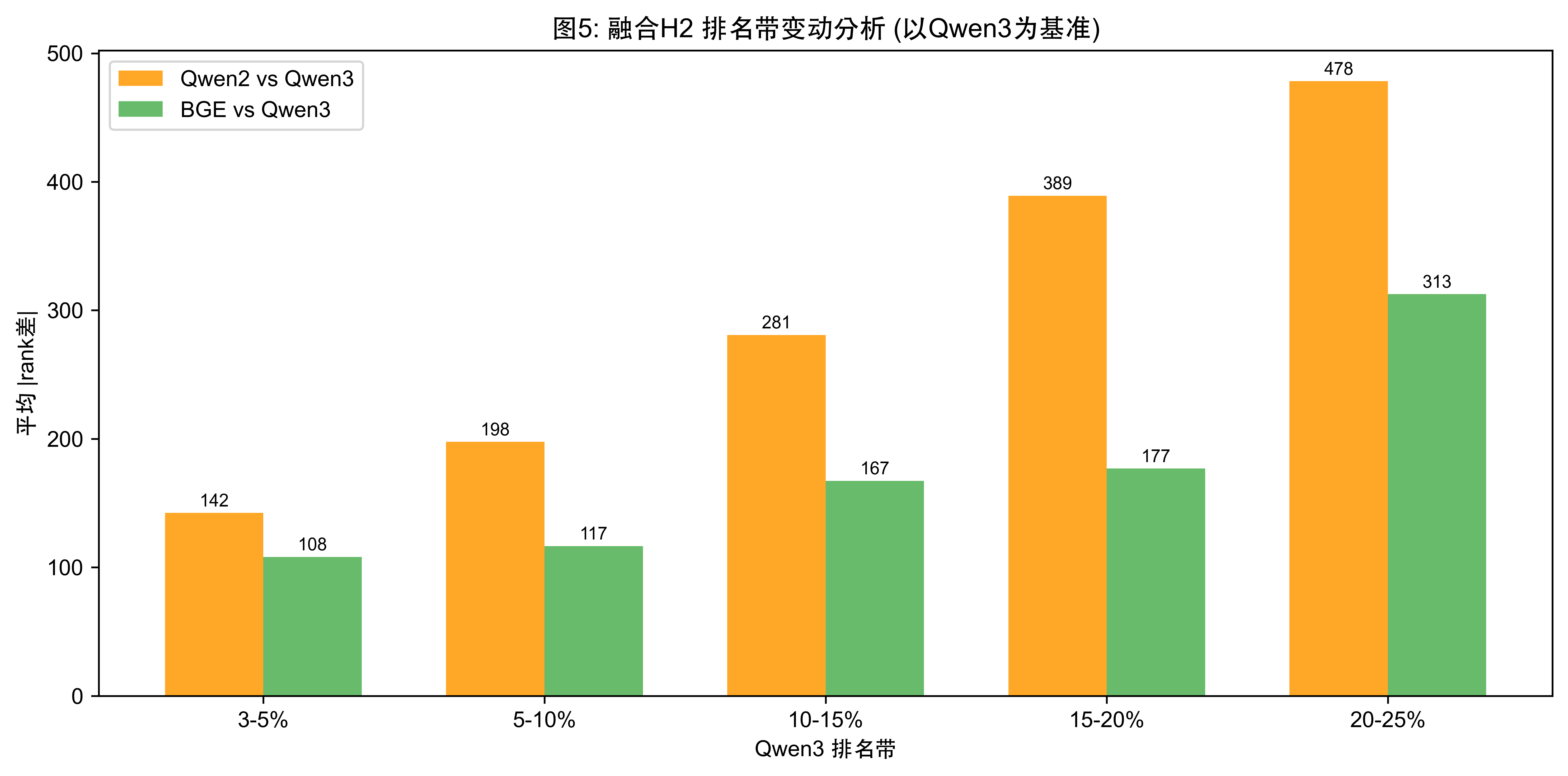
|  |  |  |  |
| --- | --- | --- | --- |
| Top-k % | Q3 vs Q2 | Q3 vs BGE | Q2 vs BGE |
| 1% | 0.4762 | 0.6727 | 0.4091 |
| 3% | 0.3824 | 0.6185 | 0.4100 |
| 5% | 0.4873 | 0.7491 | 0.4673 |
| 10% | 0.6555 | 0.7444 | 0.5983 |
| 15% | 0.6461 | 0.7604 | 0.6339 |
| 20% | 0.6372 | 0.7668 | 0.6423 |
| 25% | 0.6409 | 0.7442 | 0.6359 |

融合H2 Top-14 三模型重叠: 12/14 对

Qwen3-8B 融合H2 Top-14:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| rank | path1 | path2 | score\_final | semantic\_cosine | 三方共有 |
| 1 | 《礼记·乡饮酒义》 | 《荀子·乐论篇》第二十 | 0.8626 | 1.0000 | \* |
| 2 | 《礼记·哀公问》 | 《大戴礼记·哀公问于孔子》第四十一 | 0.8432 | 0.9991 | \* |
| 3 | 《礼记·祭义》 | 《大戴礼记·曾子大孝》第五十二 | 0.8120 | 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 | 0.9938 | \* |
| 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 | 0.9768 | \* |
| 6 | 上博简·《民之父母》 | 《礼记·孔子闲居》 | 0.7977 | 0.9597 | \* |
| 7 | 《礼记·哀公问》 | 《孔子家语·大昏解》第四 | 0.7969 | 0.9740 | \* |
| 8 | 《大戴礼记·主言》第三十九 | 《孔子家语·王言解》第三 | 0.7924 | 0.9720 | \* |
| 9 | 《礼记·仲尼燕居》 | 《孔子家语·论礼》第二十七 | 0.7659 | 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 | 0.9898 |  |
| 11 | 《大戴礼记·哀公问五义》第四十 | 《荀子·哀公篇》第三十一 | 0.7599 | 0.9613 | \* |
| 12 | 《晏子春秋·杂》上·《晋欲攻齐使人往观晏子以礼侍而折其谋》 | 《新序·杂事》第一 | 0.7512 | 0.8898 | \* |
| 13 | 《礼记·孔子闲居》 | 《孔子家语·论礼》第二十七 | 0.7413 | 0.9256 | \* |
| 14 | 《礼记·射义》 | 《孔子家语·观乡射》第二十八 | 0.7327 | 0.9338 |  |

## 4.5 排名带变动分析



*图5: 融合H2 排名带变动分析*

|  |  |  |
| --- | --- | --- |
| 排名带 | Qwen2 平均|rank差| | BGE 平均|rank差| |
| 3-5% | 142.3 | 108.1 |
| 5-10% | 197.6 | 116.5 |
| 10-15% | 280.8 | 167.3 |
| 15-20% | 389.1 | 176.9 |
| 20-25% | 478.3 | 312.6 |

观察: 在各排名带中，BGE 与 Qwen3 的排名偏差总体小于 Qwen2 与 Qwen3 的偏差，说明 BGE 虽然架构不同，但融合后的排名行为与 Qwen3 更为接近。

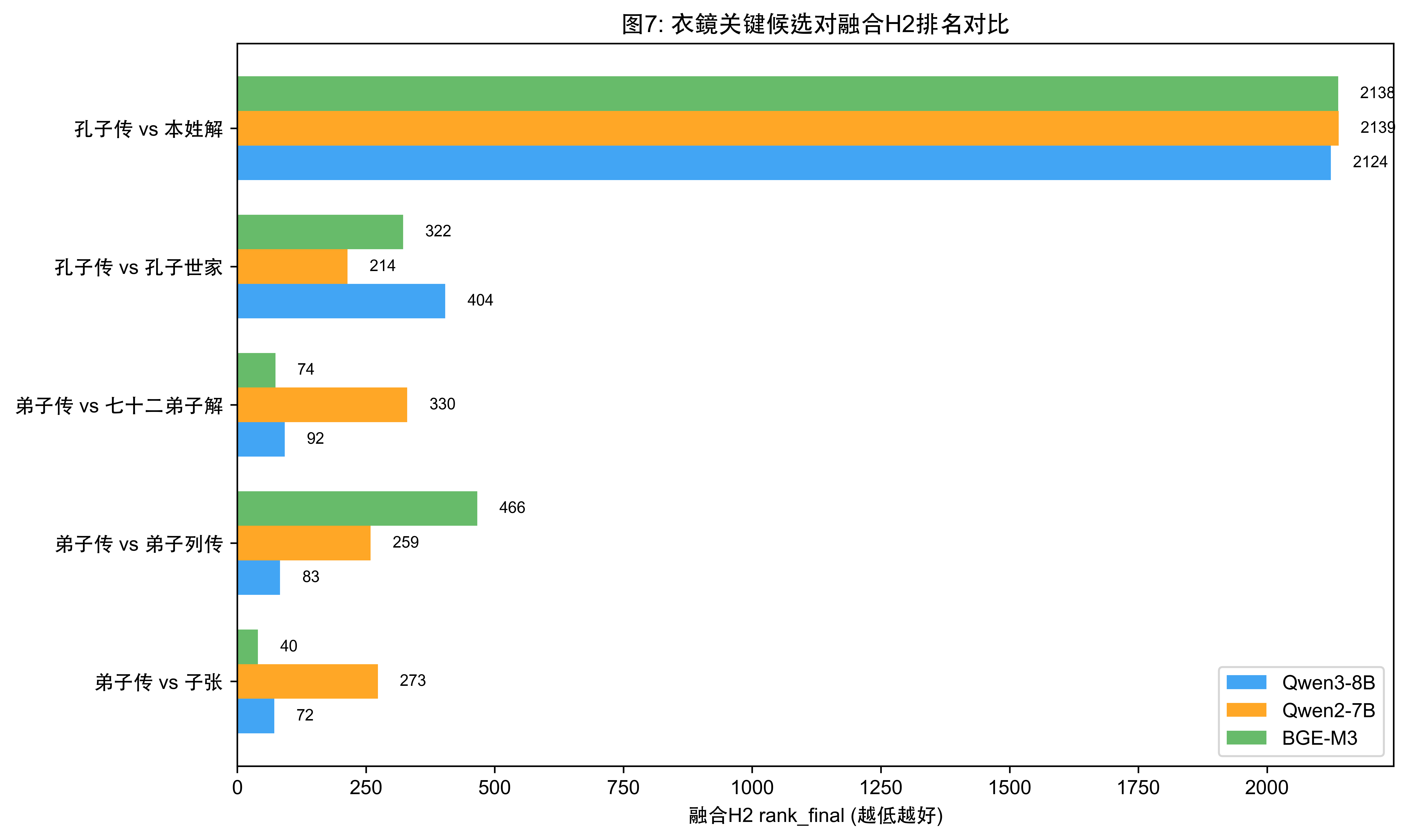
## 4.6 融合 Normal 对比

|  |  |  |  |
| --- | --- | --- | --- |
| Top-k % | Q3 vs Q2 | Q3 vs BGE | Q2 vs BGE |
| 1% | 0.7225 | 0.9578 | 0.7344 |
| 5% | 0.6884 | 0.8408 | 0.6798 |
| 10% | 0.7176 | 0.8277 | 0.7095 |
| 20% | 0.6338 | 0.7773 | 0.6285 |

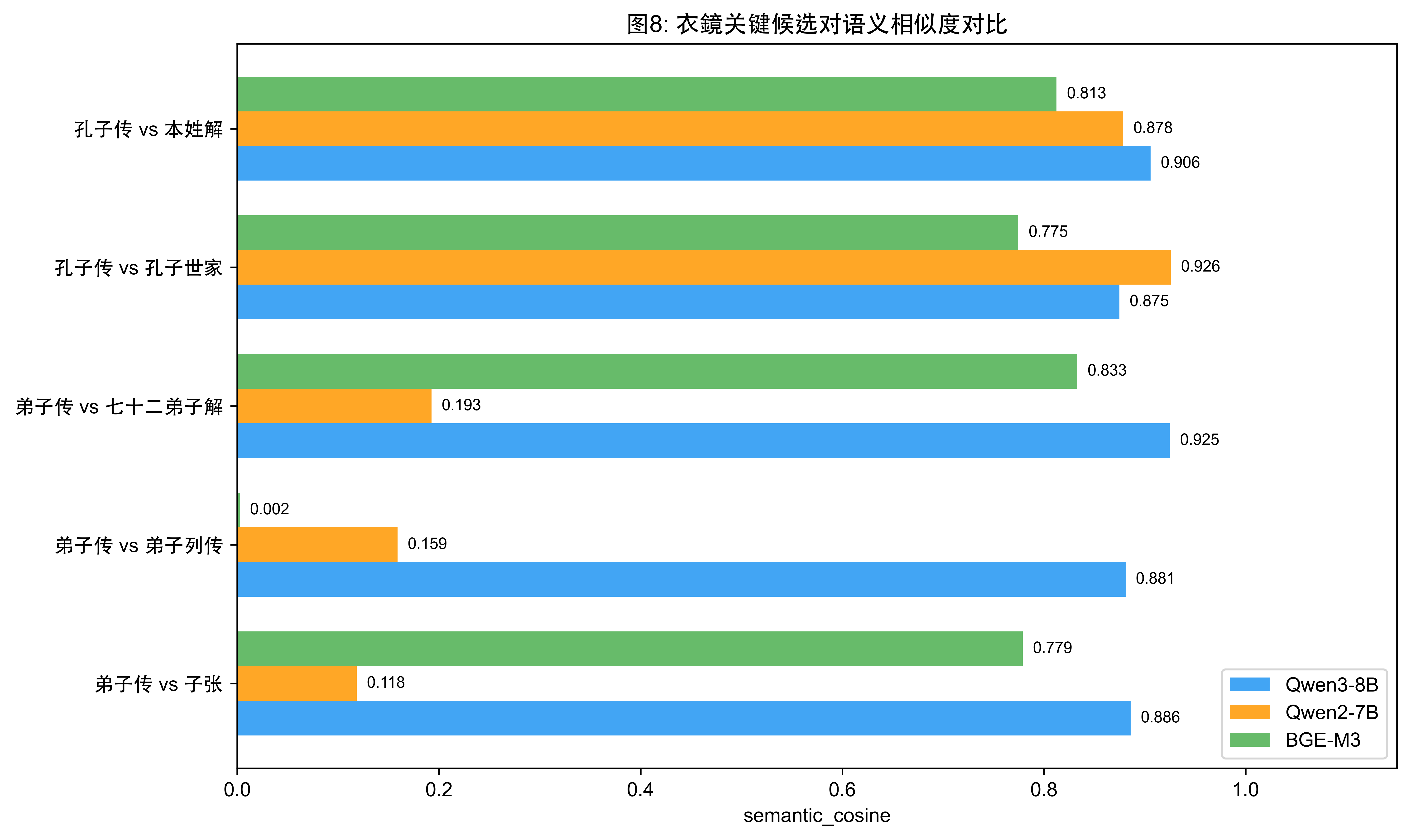
融合Normal Spearman: Q3 vs BGE = 0.955, Q3 vs Q2 = 0.847。Normal 层级的收敛效果同样显著。

# 第五章 衣鏡候选对追踪

## 5.1 弟子传关键对



*图7: 衣鏡关键候选对融合H2排名对比*



*图8: 衣鏡关键候选对语义相似度对比*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 候选对 | 模型 | 融合rank | score\_final | semantic\_cosine |
| 弟子传 vs 子张 | Qwen3-8B | 72 | 0.6720 | 0.8859 |
| 弟子传 vs 子张 | Qwen2-7B | 273 | 0.6048 | 0.1185 |
| 弟子传 vs 子张 | BGE-M3 | 40 | 0.6854 | 0.7789 |
| 弟子传 vs 弟子列传 | Qwen3-8B | 83 | 0.6686 | 0.8810 |
| 弟子传 vs 弟子列传 | Qwen2-7B | 259 | 0.6075 | 0.1589 |
| 弟子传 vs 弟子列传 | BGE-M3 | 466 | 0.5869 | 0.0024 |
| 弟子传 vs 七十二弟子解 | Qwen3-8B | 92 | 0.6669 | 0.9248 |
| 弟子传 vs 七十二弟子解 | Qwen2-7B | 330 | 0.5921 | 0.1926 |
| 弟子传 vs 七十二弟子解 | BGE-M3 | 74 | 0.6732 | 0.8331 |
| 孔子传 vs 孔子世家 | Qwen3-8B | 404 | 0.5973 | 0.8748 |
| 孔子传 vs 孔子世家 | Qwen2-7B | 214 | 0.6170 | 0.9258 |
| 孔子传 vs 孔子世家 | BGE-M3 | 322 | 0.6149 | 0.7746 |
| 孔子传 vs 本姓解 | Qwen3-8B | 2124 | 0.4527 | 0.9057 |
| 孔子传 vs 本姓解 | Qwen2-7B | 2139 | 0.4560 | 0.8785 |
| 孔子传 vs 本姓解 | BGE-M3 | 2138 | 0.4663 | 0.8125 |

解读: Qwen3-8B 对弟子传的跨书互见候选对给出了一致的语义相似度，反映了弟子传与论语/史记/家语之间的真实文本平行关系。Qwen2-7B 由于天花板压缩，无法区分真互见与普通文本对。

## 5.2 衣鏡全局统计

|  |  |  |  |
| --- | --- | --- | --- |
| 指标 | Qwen3-8B | Qwen2-7B | BGE-M3 |
| 衣鏡 H2 候选对数 | 68 | 73 | 63 |
| 平均 rank | 2436.2 | 2616.4 | 2448.6 |
| 平均 score\_final | 0.4583 | 0.4547 | 0.4779 |

# 第六章 可信度论证

## 6.1 Qwen3-8B 可信度指标汇总

(1) 语义分布健康: semantic\_cosine 标准差 0.0613，分布覆盖全区间，仅 3 对 >= 0.99，无天花板问题。

(2) 融合 Top-14 稳定: 三模型 Top-14 重叠 12/14 对，这些高排名候选对是模型无关的真实互见发现。

(3) 衣鏡案例一致: 对已知平行文本给出合理的高语义相似度。

(4) 独立验证: BGE-M3 (完全不同架构) 在融合通道与 Qwen3 的 Spearman rho = 0.926，确认 Qwen3 的排名具有跨架构稳健性。

## 6.2 Qwen2-7B 不可信原因

(1) 天花板压缩: 1172 对 (29.1%) sc >= 0.99，排名退化。

(2) 语义排名异常: 语义H2 Spearman rho = -0.170，语义排名不可靠。

(3) last-token pooling 在古汉语中不适用: 句末字符 ("也""矣""乎") 频繁出现，导致不同文本的 last-token 表示高度相似。

## 6.3 BGE-M3 参考价值

(1) CLS 退化: 22 对 sc=1.000，全部涉及礼记，是 CLS pooling 的已知缺陷。

(2) 绝对分数偏低: 均值 0.729，但排名健康。

(3) 最佳独立验证器: 与 Qwen3 融合 Spearman rho = 0.926，在不同架构条件下确认了 Qwen3 排名的可信度。

# 第七章 结论

1. Qwen3-Embedding-8B 语义通道分布健康，零天花板 (仅3对 sc>=0.99)，零异常值 (sc<=0: 0对)。

2. 70/30 融合权重有效消化模型差异: 语义H2 三方Jaccard 0.020 => 融合H2 0.563 (提升 28 倍)。

3. 融合 Top-14 三模型重叠 12/14 对 -- 高排名互见发现是真实的，非模型伪影。

4. gte-Qwen2-7B 存在严重天花板压缩 (1172对 sc>=0.99, 29.1%)，语义排名退化 (rho=-0.170)，不适合古汉语互见检测。

5. BGE-M3 作为独立验证器，与 Qwen3 融合 Spearman rho = 0.926，确认 Qwen3 排名可信。