arXiv:2601.01751v1 [cs.IR] 5 Jan 2026

用于大语言模型相关性判断偏差分析的查询–文档稠密向量

Samaneh Mohtadi[1][0009−0003−0980−6254] 和 Gianluca Demartini[1][0000−0002−7311−3693]

澳大利亚布里斯班昆士兰大学
s.mohtadi@uq.edu.au，g.demartini@uq.edu.au

摘要。 与人工评估员相比，大语言模型（LLM）成本更低、可扩展性更强，因此已被用于信息检索（IR）评测集合构建中的相关性评估。以往研究考察了大语言模型相较于人工评估员的可靠性，而本文旨在了解大语言模型在判断相关性时是否会犯系统性错误，而不仅仅是评估其平均表现。为此，我们提出一种新颖的查询和文档表征方法，可用于分析相关性标签分布，并比较大语言模型与人工标签，从而识别分歧模式并定位系统性分歧区域。我们引入一个基于聚类的框架，将查询–文档（Q–D）对嵌入联合语义空间，并将相关性视为一种关系属性。在 TREC 深度学习 2019 和 2020 数据集上的实验表明，人工与大语言模型之间的系统性分歧集中于特定语义簇中，而非随机分布。查询级分析揭示了反复出现的失误，最常见于寻求定义、与政策相关或语境含糊的查询。在各簇间一致性差异较大的查询构成分歧热点；在这些热点中，大语言模型往往会漏掉相关内容，或纳入过多不相关材料。该框架将全局诊断与局部聚类相结合，揭示大语言模型判断中隐藏的弱点，从而实现关注偏差且更可靠的信息检索评估。

关键词： 相关性判断 · 偏差分析 · 信息检索评测 · 语义聚类 · 大语言模型。

1 引言

大语言模型（LLM）越来越多地被用作信息检索（IR）评估中的自动评判器。近期研究表明，虽然大语言模型的判断通常与人工相关性标签一致，但仍存在一定的不一致。信息检索系统的评估长期以来依赖相关性判断，即判断文档是否有助于回答给定查询。这些判断至关重要，因为它们决定了搜索系统的排名和评估方式 [25, 26]。例如，标准的 TREC 赛道需要六名专家评估员全职工作数周 [31]。除了成本高昂之外，人工判断也以

存在差异、受顺序影响且评估员之间不一致而著称 [1]。为应对这些挑战，研究人员开始探索能否将大语言模型用作自动评估员。这类模型能够根据自然语言提示生成流畅文本 [7, 18]。大语言模型已被用于预测 TREC 集合和商业搜索结果中的相关性，展现出远高于人工评审者（包括众包工作者）的处理能力，且标注成本更低 [33]。值得注意的是，多项研究报告指出，尽管基于大语言模型的判断并非完美无缺，但其与人工评估的一致性已达到通常足以进行系统级评估的水平 [6, 17, 18, 33, 34]。大语言模型似乎也较少受到上下文切换的影响，在评估条件变化时往往能做出更一致的判断 [19]。此外，大语言模型在自动化信息检索系统评估方面展现出巨大潜力，能够提供与人工标注质量非常接近的相关性标签 [22, 33, 40]。

尽管具有这些优势，基于大语言模型的判断仍可能存在偏差、可靠性缺陷，以及与细微的人工判断不一致的问题 [22, 40]。例如，Alaofi 等人 [1] 发现，当查询词出现在候选段落中时，大语言模型尤其容易产生假阳性，这表明它们过度依赖表层词汇重叠。类似地，Zhuang 等人 [40] 指出，对于仅部分相关的文档，二元“是/否”提示可能会产生嘈杂或有偏差的输出，因为此类情况缺少中间标签选项。Rahmani 等人 [27] 进一步表明，使用大语言模型构建的合成测试集合可能引入系统性偏差，并进而影响下游评估结果。正如 Faggioli 等人 [17] 强调的那样，信息检索领域的评估最终必须以人工判断为基础，才能保持可信度和可靠性；然而，关于大语言模型判断与专家标注员的一致程度，相关讨论仍在继续 [6, 16]。尽管这些研究为了解大语言模型的可靠性提供了宝贵见解，传统研究大多侧重于检测是否存在此类偏差，而非确定它们出现的位置，因此，如何将不一致映射到具体的语义语境，仍是一个悬而未决的问题。

本文旨在诊断大语言模型相关性判断中的隐性偏差，提升透明度并增强可靠性。我们将查询–文档（Q–D）对嵌入为稠密向量，映射到一个联合语义空间，并将相关性视为一种关系属性，而非分别表示查询和文档的嵌入。通过对该嵌入空间进行聚类，语义相似的 Q–D 对被归入连贯的邻域，以供分析。与全局一致性统计相比，这种基于聚类的一致性分析能够提供更丰富的洞见，使我们得以发现人工与大语言模型标注员系统性意见不一致的方面。具体而言，我们探讨以下研究问题：

RQ1 联合查询—文档嵌入在多大程度上捕捉了人类与大语言模型的相关性交互？
RQ2 在哪些语义簇中，以及为何大语言模型相关性判断会系统性地偏离人工判断？我们的贡献如下：

C1 我们提出一种基于聚类的框架，通过将查询—文档对嵌入到联合语义空间中，分析大语言模型相关性判断的一致性与偏差，从而超越全局统计，实现局部的、上下文感知的诊断。

C2 我们使用 Gwet 的 AC1 [20] 作为信息检索评估中的机会校正一致性度量。据我们所知，这是 AC1 在信息检索中的首次应用，它通过在标签不平衡的情况下提供更稳定的估计，并支持更细粒度的聚类级一致性分析，弥补了 Cohen 的 κ 等广泛使用的度量方法的局限。

C3 我们提出基于簇的一致性变化，用于衡量同一查询下大语言模型与人工之间的一致性，在不同语义邻域中的变化情况。该度量使我们能够定位系统性分歧，区分嵌入空间中稳定与不稳定的区域，并识别特别容易产生偏差的查询，从而诊断语义盲区。

2 相关工作

更广泛的研究领域日益关注大语言模型评审流程的有效性与局限性。基准测试研究表明，在精心设计提示的情况下，大语言模型能够近似反映用户偏好 [33], ，为其在信息检索评估中的潜力提供了初步证据。然而，其他研究对此提出了担忧，指出过度依赖可能带来偏差、透明度有限等风险，并强调需要严谨的方法 [10, 16, 30]。与此同时，综述和立场论文警示，依赖大语言模型作为评估者可能导致偏差强化、可复现性挑战以及方法不一致性 [16, 17]。实证错误分析进一步揭示了大语言模型判断的弱点。Alaofi 等人 [1]发现，模型可能会受到词汇重叠、对抗性线索和过度推断的误导；他们还观察到，大语言模型往往过度预测相关性，而其不相关标签则更为保守可靠。比较研究发现，大语言模型评估者彼此之间的一致程度往往高于其与人工评审者之间的一致程度 [19]。在 TREC 2024 检索增强生成赛道中，Upadhyay 等人 [35]主张自动化判断可以取代人工标签，但这一主张很快便受到质疑。Clarke 和 Dietz [16]指出了循环性和操纵风险，而 Balog 等人 [6] 则提供了实证证据，表明大语言模型评估者存在偏差，倾向于偏好其他大语言模型生成的检索运行。

这些发现凸显了我们不仅需要了解不一致的程度，还需要了解它在查询与文档的语义空间中何处以及为何出现。此前的基准评测工作 [3, 16, 23] 侧重于对多样化判断取平均的宏观层面统计，因此不一致的结构及其起源仍未得到解答。

近年来，研究重点已转向诊断合成评估中的偏差与有效性。Rahmani 等人 [27] 表明，合成查询在风格和长度上都与人工查询存在差异，而 GPT-4的评判则系统性地偏宽松，由此引入正向偏差，抬高绝对性能分数，同时相对排名却比较稳定。Arbabi 等人 [4]提出了相对偏差框架，将嵌入分析与大语言模型评审评估相结合，以

量化不同主题子空间中的过度标注和标注不足倾向。这些研究共同表明，偏差具有结构性且依赖于具体情境，而非随机产生。这些发现凸显了以下重要性：不仅要考察不一致的程度，还要探究它出现在哪里 以及为何 出现；这是一个悬而未决的问题，单靠全局信息检索评估指标无法充分解答。

在信息检索评估中，一致性可靠性至关重要，因为系统有效性取决于相关性判断的可靠性。Cohen 的 κ [11] 一直是占主导地位的测量方法， [3, 11, 19], 但它在类别不平衡情况下不稳定，这一问题被称为“κ 悖论” [9]。在类别不平衡的数据集中， κ 即使评估者之间达成了较高程度的一致，相关方法也可能低估一致性，因为其中很大一部分一致性会被视为偶然一致。除信息检索之外，Vidgen 等人 [36] 将 Gwet 的 AC1 应用于仇恨言论标注，展示了其在标签分布偏斜情况下的稳健性。Haley 等人 [21] 强调了 AC1 作为信息检索场景中合适替代方法的优势，尽管它尚未得到广泛采用。这表明人们已经意识到 AC1 的优势，但其在信息检索评估中的系统性使用仍未得到探索。

经典的信息检索研究确立了这样一个观点：相关性本质上具有主观性，会受到意图、任务和歧义的影响 [14, 28]。这种主观性促使我们不再将相关性视为仅属于查询或仅属于文档的属性，而是将其视为查询-文档对的关系属性。近期研究强调，需要采用以语义为基础的方法，不仅揭示大语言模型与人工之间存在多大程度的不一致，还要揭示这种不一致发生在何处以及为何发生。 ++我们的研究沿着这一方向推进，从超越标签层面的比较入手，转向一种基于嵌入的方法，将不一致定位在语义空间中。通过分析联合查询-文档嵌入空间中的邻域、考察簇级一致性稳定性，并识别偏差模式，我们揭示了语料库层面的平均值所掩盖的不一致的潜在结构。

3 方法与实验设置

我们概述了通过对查询–文档（Q–D）对进行语义聚类来分析相关性判断的方法与设置。 [1]

3.1 数据集

我们在 TREC 深度学习 2019（DL-2019） [13]和 TREC 深度学习 2020（DL-2020） [12],基于 MS MARCO 段落集合构建的基准数据集上进行评估 [5], , 并使用 Fröbe 等人发布的大语言模型相关性判断 [19]。作为具有代表性的大语言模型评判者，我们选择了 Claude-3-haiku、Gemini-1.5-flash-8b、GPT-4o 和 Llama-3.1，因为 Fröbe 等人报告称，这些模型在一致性、相似度和排名表现方面始终出色。表1总结了这些数据集的特征。我们的实验采用二元相关性设置，因此按照 TREC 的建议和相关规程，将分级相关性判断二值化：把分数 2 和 3 映射为相关（1），把分数 0 和 1 映射为不相关（0） [15]。

[1] 为确保可复现性，所有嵌入模型设置和指令均可在线获取： [REDACTED FOR BLIND REVIEW]。

表 1：TREC 数据集特征

数据集名称	查询	文档	相关性判断	相关（%）	不相关（%）
DL-2019	43	9139	9260	27.0	73.0
DL-2020	54	11224	11386	15.0	85.0

3.2 查询–文档对嵌入

我们采用 INSTRUCTOR [32] 作为嵌入模型。它是一种指令微调编码器，旨在生成任务感知嵌入。INSTRUCTOR 会根据指定预期任务的自然语言指令来调整嵌入过程，使语义一致的输入彼此靠近，同时让不相关的输入彼此远离。这一设计使其特别适用于搭配精心选择的任务指令进行信息检索和相关性评估。据我们所知，这是首次使用 INSTRUCTOR 嵌入来分析人工与大语言模型相关性判断之间的一致性。

由于 INSTRUCTOR 允许嵌入根据不同指令进行调整，我们系统地探索了十九种不同表述，涵盖二元分类、面向判断的任务、语义相似度和检索导向主题。每种指令变体都会生成一组嵌入，我们使用 HDBSCAN [8] 对其进行聚类。重要的是，聚类针对的是语义表示，而非相关性标签。只有在聚类完成后，我们才附加二值化人工标签，以评估各个邻域内人工判断的分布情况。为评估嵌入质量，我们测量了簇纯度，区分由相关与不相关标签占主导的簇，并检视其纯度直方图。为便于跨指令比较，我们采用簇纯度的第80百分位数作为稳健标准，它表示至少80%的簇所达到的纯度水平。例如，使用指令 “评判文档与查询的相关性，以进行临时检索” 和 INSTRUCTOR-XL 时，不相关簇和相关簇的第80百分位数纯度分别达到 ≈ 0.70 和 ≈ 0.61 。由于与其他方案相比，该指令与人工判断的对齐程度更高且更稳定（对于 INSTRUCTOR-Large 和 INSTRUCTOR-XL 均如此），因此我们选择将其与 INSTRUCTOR-XL 搭配，作为所有后续分析所用的嵌入配置。

为验证这一选择，我们还测试了其他模型。E5-Large-v2 [37],虽然在非对称检索中表现出色，但并不适合我们的分析，因为分析需要联合的 Q–D 嵌入。Qwen3-Embedding-8B [39], 尽管基准测试表现强劲，但生成的簇一致性较差，多数标签对齐也较弱。这些结果证实，采用所选指令的 INSTRUCTOR-XL 是本研究中最稳定、最有效的配置。

3.3 聚类

在嵌入步骤之后，我们使用 HDBSCAN 对 Q–D 对进行聚类，因为它能够识别嵌入空间中的稠密区域，而无需假设其形状为球形或均匀分布。因此，它非常适合语义多样且分布不均的查询–文档嵌入。值得注意的是，HDBSCAN 会将一些点分配到

噪声簇（C_{-1}）中，以捕捉语义含糊、非典型或与任何稠密邻域连接较弱的 Q–D 对。这些情况不会被丢弃，而是能提供有用信号，帮助识别易产生偏差或不稳定的区域；在这些区域中，人工判断与大语言模型判断最有可能出现分歧。

图 1 展示了查询–文档对的一种聚类示例。每个彩色组（C_1、 C_2、 C_3）代表一个稠密的语义邻域，而灰色点（C_{-1}）表示被归为噪声的点。十字标记表示查询 q_1对应的示例 Q–D 对；根据其关联文档的语义关系，这些对会出现在不同的簇中。例如，$\langle q_1$、 $D_1\rangle$ 和 $\langle q_1$、 $D_3\rangle$ 分别落入不同的簇，反映出它们不同的语义对齐方式。这表明，聚类会区分某个查询的文档中彼此连贯的子集，而不是将同一查询的所有文档归为一组。值得注意的是，同一文档可能会因对应不同查询而出现在不同簇中（例如 q_2，星形标记），这表明查询–文档嵌入捕捉的是语义上下文，而非由文档内容主导。聚类后，我们将二值化的人工和大语言模型相关性标签附加到簇成员上，从而能够评估全局和局部的一致性模式。这些赋予的判断如图 $J_{q、\,d}$ 所示，例如 $\langle q_1$、 D_1、$J_{1、\,1}\rangle$ 图 1所示。

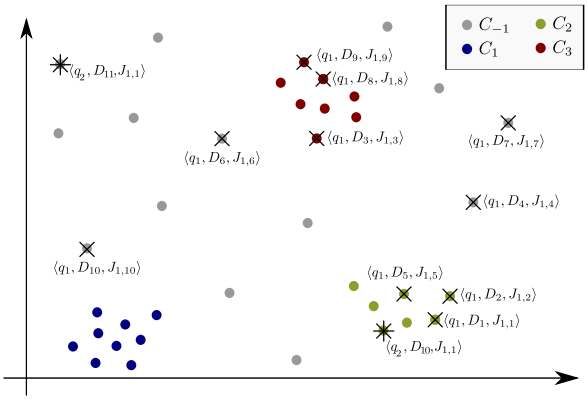

图 1：使用 HDBSCAN 对查询–文档（Q–D）嵌入进行聚类。不同颜色的组（C_1、 C_2、 C_3）表示稠密语义邻域，灰色点（C_{-1}）表示噪声。十字和星号分别标示查询 q_1 和 q_2 对应的 Q–D 对，说明文档和查询会因语义对齐情况不同而落入不同的簇。

3.4 一致性度量

我们使用 Gwet 的 AC1 作为主要的一致性度量，衡量人工判断与大语言模型判断之间的一致性。虽然 κ 仍是信息检索领域的主流度量，但在类别不平衡时会失效，而这在相关性评估中十分常见。相比之下，AC1 以不一致而非偶然一致为基准，因此在标签分布偏斜时能提供更稳定的估计。医学和临床研究的证据表明，AC1 在不平衡情况下表现稳定 [38]，模拟研究也进一步证实了它在各种二元

情境 [24, 29]下的稳健性。这些特性使 AC1 非常适用于信息检索评估，因为相关文档数量稀少；也适用于我们的研究设置，因为语义簇可能存在严重的类别不平衡。

为说明这一效应，图2 展示了一个示例聚类在 DL-2019 上的 Gwet 一致性得分，该示例使用 GPT-4o 相关性判断。每个点代表一个簇，颜色表示标签熵，用于衡量平衡程度。颜色较深的点代表熵较低、因而不平衡程度较高的簇。在这些区域，Cohen 的 κ 趋近于零，而 Gwet 的 AC1 在不平衡情况下不会趋近于零，并且在这些较深色的点之间仍能呈现有意义的变化。AC1 能捕捉到一定程度的一致性，而非默认其为零。这既凸显了簇级分析中不平衡现象的普遍性，也强调了 κ的不稳定性，进一步印证了我们选择 AC1 作为主要指标的合理性。

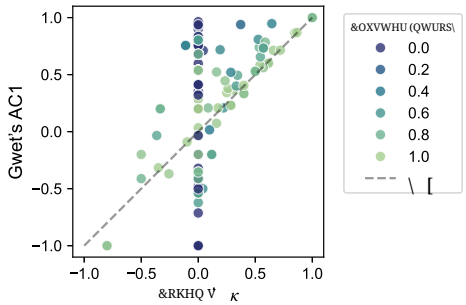

图2：GPT-4o 对 DL-2019 的评判结果中 Cohen 的 κ 和 Gwet 的 AC1 的一致性得分。每个点对应 HDBSCAN 结果中的一个簇。

3.5 基于簇的一致性变化

在聚类与一致性设置的基础上，我们定义基于簇的一致性变化这一度量。对于查询 q，我们考察包含其关联 Q–D 对的所有簇，并计算每个簇内人工评判者与大语言模型评审 j 之间的 Gwet 的 AC1。由此得到一组按簇划分的一致性值 $\{AC1_j(q,c)\}$，其中 c 为包含 q的簇的索引。随后，我们将这些值的极差定义为变化幅度，即 $\Delta AC1_j(q) = \max_c AC1_j(q,c) - \min_c AC1_j(q,c)$。该指标反映了给定查询的一致性在不同语义邻域间的变化程度，能够突出那些在某些簇中评判结果稳定、而在其他簇中则明显分歧的查询。

单个查询 q 可根据语义关系分布到多个簇中，如图1所示。例如，文档 d_1 和d_3 与 q_1 相关联，因而落入不同的簇；因此，必须分别评估每个簇内人工判断与大语言模型判断的一致性；较大的$\Delta AC1_j(q)$ 则表明不同语境下的对齐情况存在差异，并揭示全局一致性度量无法呈现的语义区域。

3.6 启发式偏差定位

基于变异度量 $\Delta AC1_j(q)$（第 3.5 节）中引入的启发式方法，我们定义了一种启发式方法，用于标记偏差易发的查询–评审者对。其目标是将变异

转化为一种诊断工具，用于识别导致系统性分歧的条目。对于每个查询 q 和大语言模型评审 j 在（q, j）中，变异度 $\Delta AC1_j(q)$ 会与混合判定条件进行比较；如果某个 Q–D 对满足以下任一条件，就会被标记为偏差易发：

1) $\Delta AC1_j(q) \geq \tau_{\text{abs}}$，绝对阈值：无论数据集特定变异如何，只要变异足够大，即可表明存在偏差；2) $\Delta AC1_j(q) \geq$ 中位数(Δ)+1.5 IQR(Δ)，稳健的数据自适应阈值，其中 Δ 是所有变异的集合，中位数(Δ)是它们的中位数，IQR(Δ)是它们的四分位距。该阈值会标记相对于全局变异分布而言属于统计离群值的查询；3) $\max_c AC1_j(q,c) > 0.8$ 和 $\min_c AC1_j(q,c) < 0.2$，用于捕捉方向性反转：一致性在某些簇中极高、在另一些簇中极低，表明存在极端的不稳定性。

随后，我们在查询层面对这些局部标记进行汇总：如果某个查询被多个评判器（两个或更多）或多数评审（超过 50%）标记，则将其标记为易产生偏差。为了对已标记的查询进行排名，我们定义偏差严重度评分 为 BSS$_j(q) = \mathbb{1}_{\text{D}}(q) + \Delta AC1_j(q)$，其中$\mathbb{1}_{\text{D}}(q)$是指示函数；若查询 q 被诊断为存在方向性偏差，则其值为 1，否则为 0。

4 实验与结果

本节介绍我们为回答RQ1 和RQ2而开展的实验。我们首先考察全局一致性模式，然后在查询和簇层面定位不一致，最后诊断易产生偏差的查询并分析其模式。这一结构体现了我们的贡献（C1–C3），从全局统计逐步深入到细粒度的上下文感知诊断。

4.1 全局偏差分析

我们首先从全局视角出发，评估大语言模型评判器相对于人工是否存在一致的偏差，以及这种偏差在整个数据集上的稳定程度。为此，我们使用 Bland–Altman（BA）图 [2], ，这是一种比较两种测量程序的标准方法。BA 图通过方法间的平均差异来刻画系统性偏差，显示其中一种方法是否持续给出更高或更低的评分；它还通过一致性界限（LoA）刻画变异性，说明个体差异的波动程度。在我们的研究中，BA 图揭示了大语言模型与人工的一致性是否与人类之间的一致性存在系统性差异，以及这些差异在不同查询之间是否稳定。

我们通过对每个查询的簇进行划分来分析这两种条件（如适用），即噪声簇（C_{-1}）和非噪声簇。分别计算每个簇的查询级 AC1；当查询跨越多个非噪声簇时，则根据簇大小进行加权平均，以汇总得分。随后，使用汇总后的查询级得分推导偏差的 Bland–Altman 估计值和一致性界限，如图 3所示。

结果呈现出一致的对比。在噪声簇条件下，偏差较小，一致性界限更窄，这表明稀疏语义区域更容易被

更一致地评判。在非噪声条件下，偏差较大，一致性界限更宽，这表明密集的语义邻域会导致系统性分歧。图3展示了 GPT-4o 在 DL-2019 和 DL-2020 上的这一模式。这些发现表明，全局偏差集中在嵌入空间中语义丰富的区域，因此有必要从全局汇总转向以查询为中心的诊断，以更精确地定位不一致的来源。

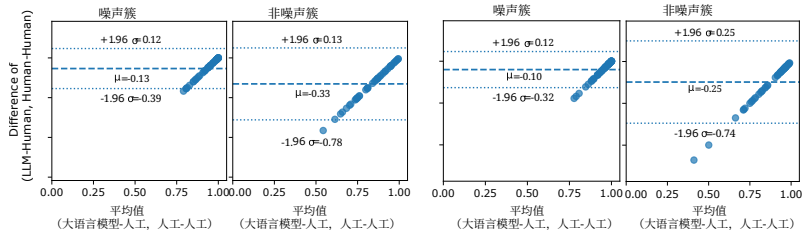

图3：GPT-4o 在 DL-2019 和 DL-2020 数据集的噪声聚类条件与非噪声聚类条件下的 Bland–Altman 图。每个点代表一个查询，虚线表示平均偏差，点线表示一致性界限（LoA）。

4.2 查询级偏差定位

超越全局概述，我们采用查询级视角来确定哪些查询导致系统性分歧。我们使用基于簇的一致性变化 $\Delta AC1(q)$ （在第3.5 节中介绍），以评估人类与大语言模型的一致性如何在单个查询的语义邻域间发生变化。图4 展示了 $\Delta AC1(q)$ 值在 DL-2019 和 DL-2020 数据集上不同评判器间的分布。较高的值表示查询在不同簇间的一致性变化显著。这类情况揭示了大语言模型评估中依赖上下文的弱点；较低的值则对应着跨簇一致性稳定的查询，这表明其行为更加稳健。这些发现表明，聚类感知分析比全局一致性度量提供了更细致的视角，并突显了两个数据集中最容易出现语义偏差的查询。

4.3 易产生偏差的查询的启发式方法诊断

我们应用第3.5 节定义的启发式方法来聚类一致性变化。如果至少两位评判者识别出某个查询，则将其标记为易产生偏差。我们采用保守的绝对阈值 $\tau_{\text{abs}} = 0.5$。由于 AC1 的取值范围为 –1 至 1，大多数实际值介于 0 至 1 之间。0.5 的偏差覆盖了这一有效范围的一半，代表显著且系统性的变化。这样可以确保只有不稳定性较大的查询才会被标记，优先考虑精确率（捕捉真正不稳定的情况），而非召回率（纳入轻微波动）。图5 展示了启发式方法在 DL-2019 和 DL-2020 中识别出的排名前 10 的易产生偏差的查询，并按其在各评判者间的平均偏差严重度评分排序。每个柱形代表

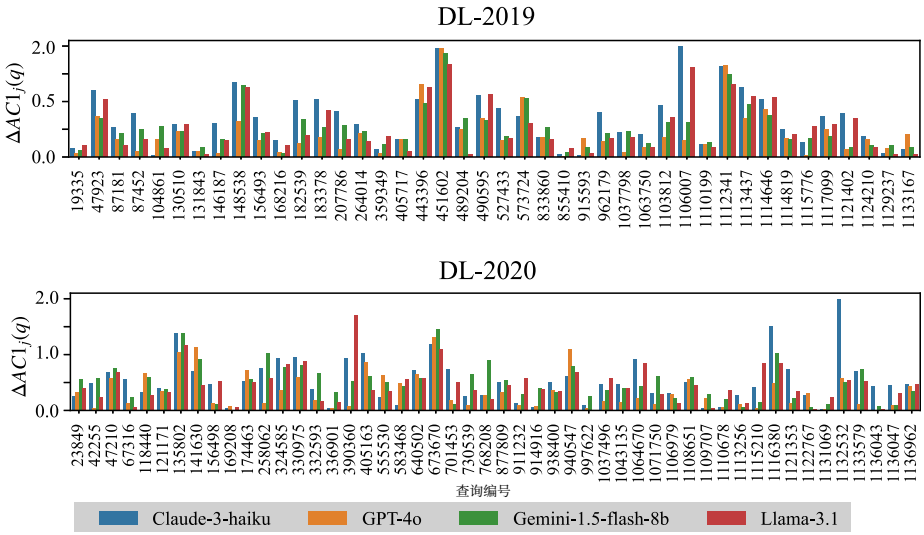

图 4：基于簇的一致性变化 $\Delta AC1_j(q)$，根据所有查询计算得出

某位评判者的变化 $\Delta AC1_j(q)$。下方各面板总结了主要偏差原因，其中标签 A（绝对阈值）、R（稳健离群值）和 D（方向性偏差）表示触发查询标记的启发式条件。由于绝对阈值适用于所有查询，排名靠前的查询都会带有 (A) 标记；它们的区别在于是否还表现出稳健离群值行为 (R) 或方向性反转 (D)。

偏差严重度评分通过结合方向性偏差与变化幅度来确定偏差排名。排名较高的查询通常会累积所有启发式方法标记，反映出强烈且系统性的不稳定性。此诊断提供了一份具体的易产生偏差查询清单，这些查询会导致模型与人工的不一致，并为发现更广泛的偏差模式奠定基础。

这种基于范围的启发式方法为识别可采取行动的偏差案例提供了实用工具。它补充了 Bland–Altman分析，将描述性总结推进为一种具体的检测机制。它能筛选出最可能扭曲评估结果的查询和模型判断，并为定性分析导致不一致的语义因素提供依据。

4.4 偏差模式的定性分析

基于第4.3节的结果，我们分析了每个数据集排名前 10 的易产生偏差的查询，以考察不一致的来源以及语义簇中的结构化偏差模式。在 DL-2019 和 DL-2020 中，我们发现大语言模型失败并非随机发生，而是遵循反复出现的模式，最常涉及定义僵化、语境歧义以及政策或角色不匹配。

在 DL-2019 中，定义性判断失败最为常见。对于寻求定义的信息型查询（例如“云杉的外观描述”“泛昔洛韦的处方用途”

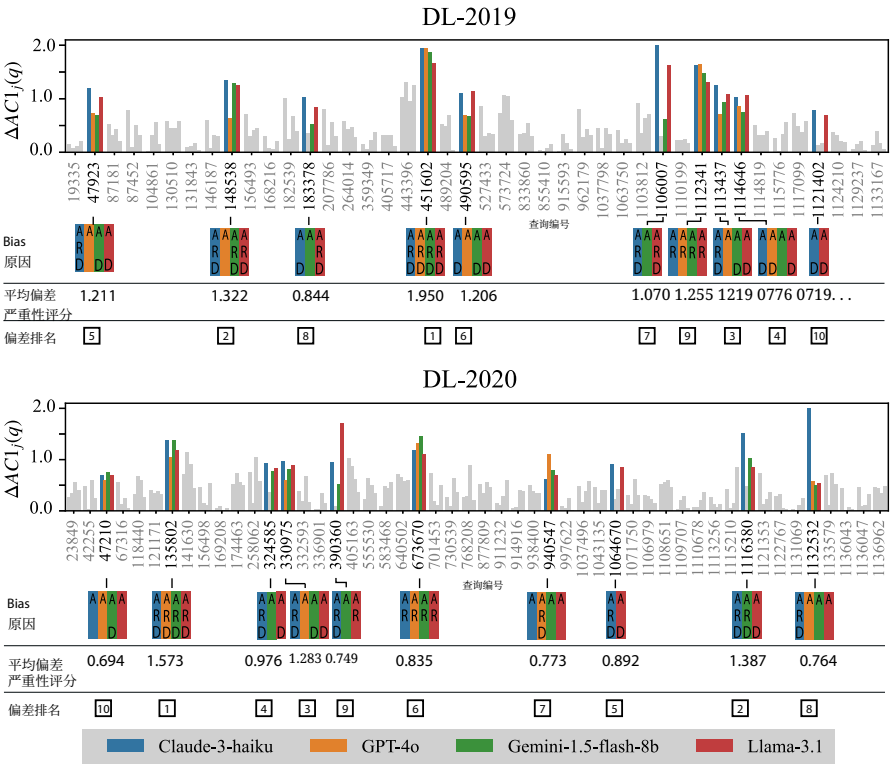

图 5：启发式方法识别出的 DL-2019 和 DL-2020 中最易产生偏差的前 10 个查询

、“轴突末梢”“RSA 定义键”“外显子定义生物学”），模型往往坚持过于严格的词典式解释，拒绝人类标记为相关的描述性或应用类段落，却有时会纳入错误的词汇匹配，例如金融领域中的“终端”。当查询需要更广泛的政策定义时，模型默认采用狭义的技术定义，从而导致政策与承保范围不匹配。例如，对于“Medicare 对机械通气的定义”，模型将涉及持续气道正压通气和双水平气道正压通气承保范围的段落判为不相关，而人类依据 Medicare 的界定认为这些段落相关，造成了明显的方向性反转。专业角色区别和领域歧义也带来了问题。在“注册护士与护理学学士”这一查询中，模型混淆了人类认为相关的细微差异；而对于“泰国人的日常生活”和“等高耕作”等查询，人类能够灵活接受多种表述方式，大语言模型却采用僵化的启发式方法，导致其在不同簇中的判断不稳定。

在 DL-2020 中，失败再次呈现出反复出现的模式。对于“admirable”的定义、后缀“ia”的含义和什么是 alm 等定义类查询，大语言模型表现出过于狭窄的词典式或技术性解读。它们召回不足，未能检索到人工评判者认为相关的较广泛段落，有时还偏离到无关领域，例如将“alm”误解为“施舍”。领域歧义

以及文化框架也造成了显著的不稳定性。对于“地球科学中的不整合”这一查询，人工评判者始终将地质学相关段落标记为相关内容，但大语言模型却将其误判为社会反对，而非地质不整合。对于“摇滚乐始于何时？”和“猎人为什么要给霰弹枪配上瞄准花纹？”等查询，人工评判者接受多种相互重叠的解读，而大语言模型则强行给出单一、固定的答案，或纳入关联较弱的内容。这导致了簇级翻转。对于“风力涡轮机安装成本”“婚纱修改成本”和“励志演讲者薪资”等实际信息需求，也显示出另一种持续存在的弱点。人工评判者将描述性或包含丰富背景信息的段落视为相关内容，而大语言模型要么将判断局限于狭窄的数字信息，要么纳入无关的财务材料。这些失败与 DL-2019 中观察到的根本原因相同。过度严格的过滤导致系统性召回不足，而过度泛化的词汇匹配则导致纳入过多无关内容。

在两个数据集中，始终出现两个根本原因。过度严格的过滤会抑制对有效语境的召回，而过于宽泛的词汇匹配则会纳入不相关材料。Claude-3 和 LLaMA-3.1 的分歧最大，在定义性和存在歧义的查询中，AC1 往往接近于零或为负。GPT-4o 和 Gemini-1.5 则更稳定，但始终倾向于召回不足。

5 结论与启示

本研究考察了联合查询—文档嵌入和聚类如何揭示基于大语言模型的相关性判断中的系统性偏差。在 DL-2019 和 DL-2020 中，我们发现，不一致并非随机出现，而是遵循集中于语义簇中的结构化模式。在每个数据集中分析的排名前 10 的易受偏差影响的查询中，定义僵化、语境歧义以及政策或角色不匹配反复导致召回不足、过度严格的过滤和簇级翻转等错误。

这些观察结果对在信息检索评估中使用大语言模型具有明确的启示。大语言模型非常适合在嵌入空间的稳定区域生成可扩展的判断；在这些区域中，它们与人工评估的一致性很高。而对于寻求定义、涉及政策或存在歧义的查询，大语言模型的可靠性较低，因为这些查询不稳定性突出，可能导致评估结果失真。尽管我们的基于聚类的框架需要人工判断来验证，但原则上，也可以将揭示偏差的相同信号（例如跨领域重叠、定义框架或语义歧义）用作预筛选工具，而无需任何人工标签。这样，实践者便能预判哪些查询可能引发不一致，并通过仅对不稳定案例进行人工评估、对大多数稳定案例依赖大语言模型来降低标注成本。未来的工作可以将这一框架扩展到更多信息检索任务、数据集和模型家族，从而推动更具偏差意识的评估实践，并使基于大语言模型的判断更可靠地融入测试集构建。

参考文献

[1] Alaofi, M., Thomas, P., Scholer, F., Sanderson, M.：大语言模型可能会被诱导将文档标记为相关：我附近最好的咖啡馆；这篇论文完全相关。载于：2024年亚太地区国际ACM SIGIR信息检索研究与发展会议论文集（SIGIR-AP '24）。第32–41页（2024）。https://doi.org/10.1145/3673791.3698431, https://doi.org/10.1145/3673791.3698431[2] Altman, D.G.：医学测量：方法比较研究分析。《英国皇家统计学会杂志：D辑（统计学家）》32，307–317（1983）。https://doi.org/10.2307/2987937, https: //doi.org/10.2307/2987937[3] Arabzadeh, N., Clarke, C.L.：基于大语言模型的相关性判断方法基准测试。载于：第48届国际ACM SIGIR信息检索研究与发展会议论文集（SIGIR '25）。第3194–3204页（2025）。https://doi.org/10.1145/3726302.3730305, https: //doi.org/10.1145/3726302.3730305[4] Arbabi, A., Kerschbaum, F.：相对偏差：量化大语言模型偏差的比较框架（2025），https://arxiv.org/abs/2505.17131[5] Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., 等：MS MARCO：人工生成的机器阅读理解数据集（2016），https://arxiv. org/abs/1611.09268[6] Balog, K., Metzler, D., Qin, Z.：排序器、评判器与助手：探索大语言模型在信息检索评估中的相互作用。载于：第48届国际ACM SIGIR信息检索研究与发展会议论文集（SIGIR '25）。第3865–3875页（2025）。https://doi.org/10.1145/3726302.3730348, https://doi.org/10.1145/3726302.3730348[7] Bauer, C., Carterette, B., Ferro, N., Fuhr, N., Beel, J., Breuer, T., Clarke, C.L., Crescenzi, A., Demartini, G., Di Nunzio, G.M., 等：研究与教育领域的信息访问实验研究前沿（Dagstuhl研讨会23031）。《Dagstuhl报告》13(1)，68–91（2023）。https://doi.org/10.4230/DagRep.13.1.68, https://doi.org/10.4230/DagRep.13.1.68[8] Campello, R.J.G.B., Moulavi, D., Sander, J.：基于层次密度估计的密度聚类。载于：《知识发现与数据挖掘进展：第17届亚太会议论文集》（PAKDD 2013）。第160–172页。Springer（2013）。https://doi.org/10.1007/978-3-642-37456-2_14, https://doi.org/10.1007/978-3-642-37456-2_14[9] Cicchetti, D.V., Feinstein, A.R.：一致性高但κ值低：I. 两个悖论的问题。《临床流行病学杂志》43(6)，543–549（1990）。https://doi.org/10.1016/0895-4356(90)90158-L

[10]Clarke, C.L.A., Dietz, L.: 基于大语言模型的相关性评估仍无法取代人工相关性评估。arXiv 预印本 arXiv:2412.17156 (2024)。
https://doi.org/10.48550/arXiv.2412.17156, https://doi.
org/10.48550/arXiv.2412.17156[11] Cohen, J.: 名义量表的一致性系数。
Educational and Psy- chological Measurement 20(1), 37–46 (1960年)。
https://doi.org/10.1177/ 001316446002000104,
https://doi.org/10.1177/001316446002000104[12] Craswell, N., Mitra, B.,
Yilmaz, E., Campos, D.: TREC 2020 深度学习赛道概述 (2021),
https://arxiv.org/abs/2102.07662[13] Craswell, N., Mitra, B., Yilmaz, E.,
Campos, D., Voorhees, E.M.: TREC 2019 深度学习赛道概述 (2020),
https://arxiv.org/abs/ 2003.07820[14] Cuadra, C.A., Katter, R.V.: 揭开“相关性”的黑箱。Jour- nal of Documentation 23(4), 291–303 (1967年)。
https://doi.org/10.1108/ eb026436, https://doi.org/10.1108/eb026436[15]
Damessie, T.T., Nghiem, T.P., Scholer, F., Culpepper, J.S.: 使用评审者间一致性衡量相关性评估的质量。载于：第40届国际 ACM SIGIR 信息检索研究与发展会议论文集（SIGIR ’17）。第1089–1092页。计算机协会 (2017)。
https://doi.org/10.1145/3077136.3080661,
https://doi.org/10.1145/3077136.3080661[16] Dietz, L., Zendel, O., Bailey, P.,
Clarke, C., Cotterill, E., Dalton, J., Hasibi, F., Sanderson, M., Craswell, N.: 大语言模型评估话语模式：关于大语言模型评估有效性的观点 (2025),
https://arxiv.org/abs/2504.19076[17] Faggioli, G., Dietz, L., Clarke, C.L.,
Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M.,
Stein, B., Wachsmuth, H.: 关于大语言模型进行相关性判断的观点。载于：2023年 ACM SIGIR 信息检索理论国际会议论文集（ ICTIR ’23 ）。第 39–50 页 (2023)。
https://doi.org/10.1145/ 3578337.3605136,
https://doi.org/10.1145/3578337.3605136[18] Faggioli, G., Dietz, L., Clarke,
C.L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast,
M., Stein, B., Wachsmuth, H.: 谁来决定什么是相关的？人工还是人工智能？为何不能兼顾？Communica- tions ofthe ACM 67(4),31–34 (2024)。
https://doi.org/10.1145/3624730, https://doi.org/10.1145/3624730[19] Fröbe,
M., Parry, A., Schlatt, F., MacAvaney, S., Stein, B., Potthast, M., Hagen, M.: 大语言模型相关性评估者彼此之间的一致性高于与人工评估员之间的一致性。载于：第48届国际ACM SIGIR信息检索研究与发展会议论文集（SIGIR ’25）。第2858–2863页 (2025)。https://doi.org/10.1145/ 3726302.3730218,
https://doi.org/10.1145/3726302.3730218[20] Gwet, K.L.: 高一致性情况下如何计算评审者间信度及其方差。British Journal of Mathematical and Statistical Psychol- ogy 61(1), 29–48 (2008年)。
https://doi.org/10.1348/000711006X126600,
https://doi.org/10.1348/000711006X126600

[21] Haley, D.T., Thomas, P., Petre, M., De Roeck, A.: 使用一种新的评审者间信度统计量。技术报告：技术报告 2008/15，开放大学计算系，英国米尔顿凯恩斯 (2008年)，https: //oro.open.ac.uk/90234/1/TR2008_15.pdf[22] Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., 等：从生成到评判：大语言模型作为评审的机遇与挑战 (2024)，https://arxiv. org/abs/2411.16594[23] Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., Liu, Y.: 大语言模型作为评审者：基于大语言模型的评估方法综合综述 (2024)，https://arxiv.org/abs/2412.05579[24] Minozzi, S., Cinquini, M., Gianola, S., Gonzalez-Lorenzo, M., Banzi, R.: Kappa 和 AC1/2 统计量：超越悖论。Journal of Clinical Epi- demiology 142, 328–329 (2022)。https://doi.org/10.1016/j.jclinepi. 2021.12.004, https://doi.org/10.1016/j.jclinepi.2021.12.004[25] Page, L., Brin, S., Motwani, R., Winograd, T.: PageRank 引文排名：为网络带来秩序。技术报告 1999-66，斯坦福 InfoLab (1999年11月)，http://ilpubs.stanford.edu:8090/422/，之前编号= SIDL-WP-1999-0120[26] Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: 我们信任 Google：用户对排名、位置和相关性的判断。Journal of Computer-Mediated Communication 12(3), 801–823 (2007)。https://doi.org/10.1111/j.1083-6101.2007.00351.x[27] Rahmani, H.A., Ramineni, V., Craswell, N., Mitra, B., Yilmaz, E.: 探索评估合成数据中的偏差。arXiv 预印本 (2025)，https://arxiv.org/abs/2506.10301[28] Rees, A.M., Schultz, D.G.: 以现场实验方法研究相关性评估与文档检索的关系：提交给美国国家科学基金会的最终报告。第二卷，附录。技术报告 ED027910，ERIC (1967年)，https://eric.ed.gov/?id=ED027910[29] Silveira, P.S.P., Siqueira, J.O.: 与其和糟糕的伙伴一致，不如追求一致：对多种类 Kappa 检验的批判性分析。Behavior Re- search Methods 55(7), 3326–3347 (2023)。https://doi.org/10.3758/ s13428-023-02124-5, https://doi.org/10.3758/s13428-023-02124-5[30] Soboroff, I.: 不要用大语言模型作出相关性判断。Information Re- trieval Research Journal (2024)。https://doi.org/10.54195/irrj.19625, https://doi.org/10.54195/irrj.19625[31] Soboroff, I.: 不要用大语言模型作出相关性判断。信息检索研究 1(1), 29–46 (2025)。https://doi.org/10.54195/irrj. 19625, https://doi.org/10.54195/irrj.19625[32] Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.t., Smith, N.A., Zettlemoyer, L., Yu, T.: 一个嵌入器，适用于任何任务：指令微调文本嵌入。载于：Rogers, A.,Boyd-Graber, J., Okazaki, N. (编) 计算语言学协会论文集：ACL 2023。第1102–1121页。计算语言学协会，多伦多，加拿大 (7月

2023年）。https://doi.org/10.18653/v1/2023.findings-acl.71, https://aclanthology.org/2023.findings-acl.71/[33] Thomas, P.、Spielman, S.、Craswell, N.、Mitra, B.：大语言模型能够准确预测搜索者偏好。载于：第47届国际 ACM 信息检索研究与发展会议论文集（SIGIR '24）。第1930–1940页（2024年）。https://doi.org/10.1145/3626772.3657707, https://doi.org/10.1145/3626772.3657707[34] Upadhyay, S.、Pradeep, R.、Thakur, N.、Campos, D.、Craswell, N.、Soboroff, I.、Dang, H.T.、Lin, J.：使用 UMBRELA 和大语言模型进行相关性评估的大规模研究。载于：2025年国际 ACM SIGIR 信息检索创新概念与理论会议论文集（ICTIR '25）。第358–368页（2025年）。https://doi.org/10.1145/3731120.3744605, https://doi.org/10.1145/3731120.3744605[35] Upadhyay, S.、Pradeep, R.、Thakur, N.、Craswell, N.、Lin, J.：Umbrela：Umbrela 是必应相关性评估器的（开源复现版）（2024年），https://arxiv.org/abs/2406.06519[36] Vidgen, B.、Thrush, T.、Waseem, Z.、Kiela, D.：从最糟糕的案例中学习：通过动态生成的数据集改进在线仇恨检测。载于：第59届计算语言学协会年会论文集（ACL 2021）。第1667–1682页。计算语言学协会（2021年）。https://doi.org/10.18653/v1/2021.acl-long.132, https://arxiv.org/pdf/2106.09775[37] Wang, L.、Yang, N.、Huang, X.、Jiao, B.、Yang, L.、Jiang, D.、Majumder, R.、Wei, F.：通过弱监督对比预训练获得的文本嵌入（2022年），https://arxiv.org/abs/2212.03533[38] Wongpakaran, N.、Wongpakaran,T.、Wedding, D.、Gwet, K.L.：计算评定者间信度系数时 Cohen κ 与 Gwet 的 AC1 的比较：一项基于人格障碍样本开展的研究。BMC医学研究方法学**13**（1），61（2013年）。https://doi.org/10.1186/1471-2288-13-61, https://doi.org/10.1186/1471-2288-13-61[39] Zhang, Y.、Li, M.、Long, D.、Zhang, X.、Lin, H.、Yang, B.、Xie, P.、Yang, A.、Liu, D.、Lin, J. 等：Qwen3 嵌入：借助基础模型推进文本嵌入与重排序（2025年），https://arxiv.org/abs/2506.05176[40] Zhuang, H.、Qin, Z.、Hui, K.、Wu, J.、Yan, L.、Wang, X.、Bendersky, M.：超越是与否：通过为细粒度相关性标签评分，改进零样本大语言模型排序器。载于：NAACL-HLT 2024（短文）论文集。第358–370页（2024年）。https://doi.org/10.18653/v1/2024.naacl-short.31, https://aclanthology.org/2024.naacl-short.31/