# 超越真假：检索增强的复杂主张的层次分析

**Priyanka Kargupta***, **Runchu Tian***, **Jiawei Han**
Department of Computer Science, University of Illinois at Urbana-Champaign
{ pk36, runchut2, hanj } @illinois.edu

## Abstract

个人或实体提出的主张通常是微妙的，不能明确地标记为完全“真”或“假”——这在科学和政治主张中经常发生。然而，一个主张（例如，“疫苗 A 比疫苗 B 更好”）可以被分解为其基本的方面和子方面（例如，功效、安全性、分布），这些方面单独验证起来更容易。这使得能够提供更全面、结构化的响应，从而对给定问题提供一个全面的视角，同时也允许读者在主张中优先考虑特定的关注角度（例如，针对儿童的安全性）。因此，我们提出了 CLAIMSPECT，一个基于检索增强生成的框架，旨在自动构建一个层次结构，包含通常在处理主张时考虑的各个方面，并通过特定语料库的视角丰富它们。该结构以层次方式划分输入语料库以检索相关的部分，这些部分有助于发现新的子方面。此外，这些部分使得能够发现对主张某个方面的不同视角（例如，支持、中立或反对）及其各自的普遍性（例如，“多少生物医学论文认为疫苗 A 比疫苗 B 更易运输?”）。我们将 CLAIMSPECT 应用于我们构建的数据集中包含的各种真实世界的科学和政治主张，展示其在解构微妙主张和在语料库中表示视角方面的稳健性和准确性。通过真实案例研究和人类评估，我们验证了其在多个基准上的有效性。

## 1 介绍

科学和政治话题越来越以简洁、引人注目的断言形式被消费，这些断言缺乏表现复杂现实所需的细微差别。这些断言经常被过度简化或自信地陈述，尽管它们仅在特定条件下或从某些角度评估时才有效。例如，“疫苗 A 比疫苗 B 更好”这样的断言可能看起来很简单，但在考虑具体方面如效率、安全性和分发物流时，则变得具有内涵。此外，在这些平台上共享的信息通常模糊且片段化，常常使这些断言被歪曲或重新定义为“真实”或“虚假”，以支持冲突的叙述，从而使验证其有效性变得复杂。

立场检测将文本意见分类为支持、中立或反对相对于目标 (Mohammad et al., 2016)。然而，文档——特别是在科学领域的——经常在一个主张的各个方面呈现出不同的立场。例如，如图 1 所示，一项研究可能会发现疫苗 A 比疫苗 B 对成人更安全，同时强调其在广泛分发时存在显著更大的物流挑战。在这种情况下，论文支持关于“对成人的安全性”的主张（注意：不是“安全性”整体），但在分发方面反对它。这种复杂性使得在文档层面对细微和多方面的主张的立场检测无效。

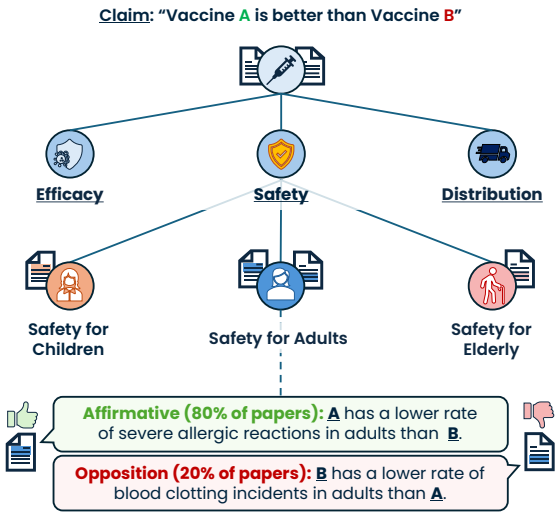


Figure 1: 一个细微主张的分解层次结构示例。每个节点都通过相关摘录进行丰富，包括支持/中立/反对的观点及其各自的证据。

事实核查模型通常通过从大型语料库中检索证据或使用网络集成语言模型 (Thorne et al., 2018; Popat et al., 2018; Zhang and Gao, 2023) 来验证声明。虽然一些方法现在提供多样化的真实性判断，如“基本属实”或“半真” (Zhang and Gao, 2023)，但在科学环境中，这些方法不够有效。特别是在不断发展的领域中，精细的科学声明可能没有得到证实，因为缺乏研究或科学共识，而不是彻底错误。这一区别至关重要，因为它突出了需要进一步探索的领域。例如，在图 1 中，映射到“成人安全”节点的相关论文摘录显示，附属与反对立场的 80:20 比例对于子方面声明意味着共识，而 60:40 比例或稀疏的数据则表明研究有限或存在分歧。这种见解对于理解知识差距至关重要，但常常被现有的事实核查框架忽视。

我们使用 CLAIMSPECT 解决这些挑战，这是一种通过利用大型语言模型（LLMs）系统地解构和分析声明的框架。ClaimSpect 将一个声明层次地分割成一个包含多个方面和子方面的树，从而实现结构化验证和发现观点。这是通过采用以下原则来完成

*Equal contribution.

的：

与单独考虑一个目标声明和整个文档相反，我们必须首先确定分析语料库中讨论的相关方面，以发现更有针对性的子声明。然而，保持这些方面的层级性质是至关重要的。图 1 中展示了这一点，其中一些难以验证的方面（例如，“安全性”）通常可以被拆分，直到它们到达更常考虑的“原子”子方面（例如，“儿童安全”、“成年人的安全”和“老年人的安全”）。此外，这些层级关系通常也反映在我们自然地对一个给定主题（个体或集体）形成观点的方法中：解析现有知识，考虑基于这些知识的不同子角度的问题，检索更多与子角度相关的知识，相应地发展我们的观点，并将它们聚合为一个高层次的观点 (Perony et al., 2013; Chen et al., 2022) 。因此，这引出了我们的下一个原则。

Principle # 2: Iterative, discriminative retrieval enhances LLM-based tree construction. 大型语言模型最近在自动分类拓展和扩充方面表现出很有前景的能力，将数据组织成类似我们目标方面层次结构 (Shen et al., 2024b; Zeng et al., 2024) 的类别和子类别的层次结构。然而，这些方法通常依赖于大型语言模型预训练数据集中存在的一般知识，忽视了对于 (1) 发现领域特定数据中普遍存在的细粒度子方面的至关重要的语料库特定见解，以及 (2) 确保与确定语料库范围的共识任务的一致性。为了解决这个问题，我们利用了检索增强生成 (RAG)，它最近通过在生成过程中整合外部语料库或数据库，在知识密集型任务中取得了进展 (Lewis et al., 2020; Gao et al., 2023) 。我们介绍了一种迭代的 RAG 方法，该方法通过为一个方面节点检索相关段落并使用它们来发现新的子方面，动态构建方面层次结构。这样可以确保分类法与语料库中特定的声明、方面和观点讨论紧密对齐。

我们注意到，噪声检索常常阻碍推理性能 (Shen et al., 2024a) 。在我们的设置中，当某些检索到的摘录与多个语义相似的方面节点重叠时（例如，“儿童安全”与“成人安全”），在仅为一个方面确定子方面时引入噪声。为减轻这种情况，我们引入了一种判别性排序机制，优先考虑深度讨论单一方面的段落，增强子方面的发现和最终方面层次结构。

对于层次结构中的每个方面节点，我们使用层次文本分类和立场检测来识别和聚类论文，基于它们的立场（肯定、中立、反对）。这些聚类不仅揭示了一致性的存在或缺失，还揭示了每种立场中的关键观点。例如，如图 1 所示，肯定的观点可能强调疫苗 A 在成人中较低的严重过敏反应率，而反对的观点关注其较高的血栓发生率。这些观点提供了透明性，揭示了潜在的研究空白（例如，如果 80 % 的肯定论文没有涉及这些血栓事件），并为提出细化的主张提供了关键背景。

总体而言，CLAIMSPECT 采用一种结构化的方法，将一个复杂的主张分解为一个层次化的方面体系，目标是采用全面的方法考虑所有可能用于验证根本主张的方面。该框架包括以下步骤：（1）辨别性方面检索，（2）迭代的子方面发现，以及（3）基于分类的观点发现。我们的贡献可以总结为：

- 据我们所知，CLAIMSPECT 是第一个将论点正式解构为各个方面的层次结构以确定共识的工作。
- 我们构建了两个新的数据集，这些数据集包括现实世界中具有科学和政治微妙性的主张及相应的语料。
- 通过对现实世界领域的实验和案例研究，我们证明了 ClaimSpect 在进行层次共识分析时明显比基线方法更有效。

可复现性：我们提供了我们的数据集和源代码 [*] ，以便于进一步研究。

## 2 相关工作

Fact Checking. 事实核查模型 (Thorne et al., 2018; Popat et al., 2018; Atanasova et al., 2019; Karadzhov et al., 2017) 利用外部证据来验证声明，但往往将声明视为一个整体的陈述。网络集成的方法 (Zhang and Gao, 2023; Karadzhov et al., 2017) 尝试用额外的上下文来丰富事实核查，但仍然无法处理那些如果不考虑多样的声明子方面及其不同层次的证据就无法明确验证的复杂声明。相反，CLAIMSPECT 认识到某些声明背后的复杂性，利用一个语料库帮助识别在验证声明时可能考虑到的各个方面，从而实现更具多面性和可解释性的分析。我们注意到 CLAIMSPECT 并不旨在验证给定的声明，它仅仅是旨在将声明分解为可用于验证的多层次的方面层次，提出对于声明方面的潜在的观点，这些观点基于语料库。我们还注意到邻近的证据检索任务，现有研究探索根据固定且平面的（非层次化的）一组方面来组织证据：人群、干预措施和结果 (Wadhwa et al., 2023) 。

最近在分类法生成方面的进展展示了大型语言模型在分层结构化信息方面的潜力。然而，这些方法通常依赖于静态的、与领域无关的知识，限制了它们构建丰富、精细分类法的适应性。通过语料库感知、方面辨别性检索和迭代子方面发现来解决这些限制，构建出与语料库一致的丰富方面分类法。这使我们能够识别出与给定方面相关的片段，同时也识别出对该方面的某种观点。

Stance Detection 传统立场检测 (Mohammad et al., 2016) 将意见分类为支持的、中立的或反对的针对一个目标（例如，声明）。然而，这些方法通常将整个文档分配一个单一的立场，忽视了许多声明中存在的细致、特定方面的立场，尤其是在科学和政治背景下。最近的工作 (Zhang and Gao, 2023) 引入了更细致的判断（例如，“大多数真实”），但与事实核查方法类似，它们通常无法捕捉某些立场背后的多方面性质和理由。通过利用其构建的方面层次结构，CLAIMSPECT 能够推断出针对某个方面及其相关文献的可行的支持、中立和反对观点。

如图 2 所示，CLAIMSPECT 包括以下步骤：(1) 方面判别检索，(2) 迭代子方面发现，(3) 基于分类的视角发现。

### 2.1 预备知识

我们假设用户提供一个主张 $t_0$（例如，“疫苗 A 优于疫苗 B”）和一个语料库 $D$ 作为输入。为了更

[*] https://github.com/pkargupta/claimspect

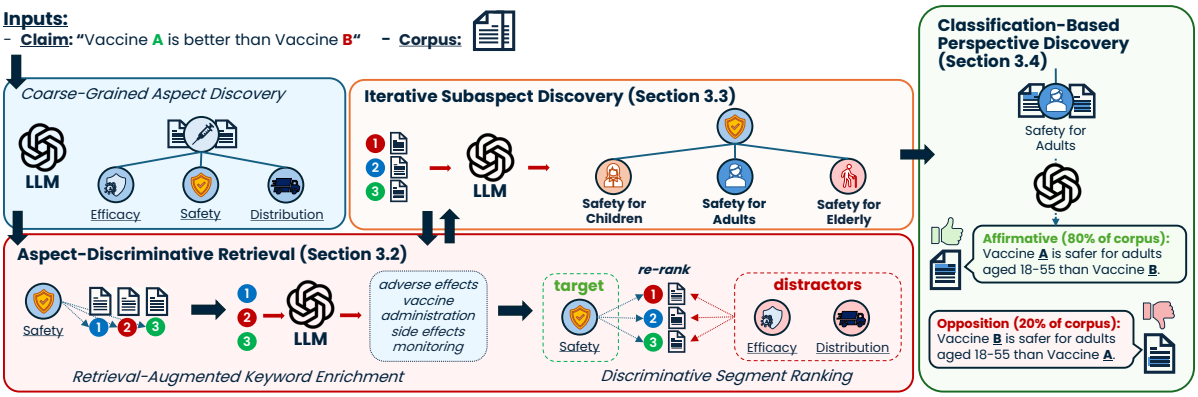


Figure 2: CLAIMSPECT 将一个复杂的论断分解为一个通常用于验证该论断方面的层次结构。我们自动从语料库中发现每个方面的观点集合。

好地反映现实世界的情况，我们不假设每个文档 $d \in D$ 与 $t_0$ 相关。

ClaimSpect 旨在输出一个方面层次结构 $T$，其中层次结构中的每个方面节点（例如，"安全性"）可以被视为用户指定的根主张 $t_0$ 的后代子主张 $t_i$（例如，"A 是一种比 B 更安全的疫苗"）。换句话说，每个方面节点 $t_i$ 应反映在评估根主张 $t_0$ 时重要的相关方面。

对于每个 $d \in D$，我们假设我们拥有其完整的文本内容（例如，一篇完整的科学论文）。为了在我们的框架中检索较小的、保留上下文的文本单位，我们使用广泛认可的文本分割方法 C99 (Choi, 2000) 将每个 $d$ 分割成块。此方法对句子进行标记，如果它们属于同一个主题组，则匹配标签，这有助于将有关某一方面的连续讨论保留在一个片段中。

在我们的弱监督设置中，仅提供了根主张 $t_0$，我们首先生成可靠的、粗粒度的方面来指导检索增强的层次结构构建。这些方面通常是常识性的，不需要领域专长来识别。初步实验确认，LLM 能够仅凭广泛的背景知识可靠地识别它们。因此，我们提示 LLM 生成粗粒度的方面 $t_i^0 \in T^0$（例如，图 1 中的功效、安全性和分布），这些将作为 $t_0 \in T$ 的子节点。对于每个方面 $t_i^0$，模型输出其标签、对 $t_0$ 的重要性以及一份 $n = 10$ 相关关键词的列表。这个初始子树构成了我们框架的基础。完整的提示在附录 A.1 中。

为了构建与语料库一致的丰富、由粗到细的方面层次结构，我们必须从语料库中识别出同样丰富的参考材料。通常，噪声检索常常妨碍推理性能，这可能会对发现给定节点的子方面产生负面影响。因此，为了发现一个方面节点的每个子方面，我们必须确定语料库中的哪些段落对此进行讨论。然而，并不是所有段落对发现子方面都同样具有信息价值。

具体来说，一个针对于节点 $t_i$ 的高质量、辨别性的片段 $s_i$ 包含以下特征：(1) $s_i$ 深入讨论 $t_i$，以及 (2) $s_i$ 没有广泛或深入地讨论 $t_i$ 的同级节点。例如，在图 1 中，一个关于疫苗 A 和 B 在儿童和成人临床试验中观察到的副作用的片段比只提到儿童时更深入地讨论了"安全性"。此外，为了发现"儿童安全性"的子方面，独立讨论儿童和成人安全性的片段会在子方面生成过程中引入额外的噪音。总的来说，将这些片段进行排名是重要的，这样我们可以选择一个集合，最小化我们在检索增强子方面发现过程中引入的噪音，同时最大化我们能够发现的子方面数量。我们在下面的章节中形式化我们的辨别性排名机制：

为了确定一个片段是否深入讨论了某个方面 $t_i$，我们首先需要进一步丰富我们对 $t_i$ 的理解。我们建议执行检索增强的基于关键词的 $t_i$ 的丰富，每一个关键词在与 $t_i$ 相关的片段中可能出现，因此明确或隐含地反映了 $t_i$ 的子方面。例如，对于"效能"方面，相应的关键词是：中和作用、免疫刺激、剂后抗体反应、以及免疫力减退。首先，我们使用检索嵌入模型从整个语料库中选择与 $t_i$ 特定查询（其根、名称、描述和关键词来自第 **??** 节）相关的、基于余弦相似性的前 $n$ 段落。

我们提供这些初始的顶部 $n$ 段，除了根主张 $t_0$，方面标签 $t_i$ 及其描述，以在上下文中帮助 LLM 识别 $2k$ 关键字。给定相同的信息和这些关键词，我们合并相似或重复的术语，同时过滤掉无关的术语——明确提示模型仅提供 $k$ 关键字。这组术语 $w \in W_i; |W_i| = k$，是我们用于节点 $t_i$ 的判别段排序的基础。我们在附录 A.2 中提供这两个提示。

为了确定对于方面节点 $t_i$ 具有最强判别力的段 $S_i$，我们首先使用与第 **??** 节相同的检索嵌入方法收集初始的大量段。我们的后续目标是根据其判别力对段 $s \in S_i$ 进行排序：

- 目标得分：根据它包含所有相关子方面的可能性奖励 $s$。
- 干扰项评分：根据它讨论的其他同级方面的程度和深度来惩罚 $s$。

我们假设 $t_i$ 的关键字 $W_i$ 隐式或显式地反映了其许多子方面。因此，我们使用它们来接近特定方面讨论的深度。我们将每个关键字 $w_i$ 转换为描述性查询："[ $w_i$ ] 相对于 [ $w_i$ 的所有祖先节点]"。通过将祖先节点整合到查询中，我们影响了 $t_i$ 的层次上下文的保留；例如，如果某段落讨论的是"Vaccine A and B 的安全性"，而不仅仅是"安全"，我们将特别给予奖励。我们使用检索嵌入模型嵌入每个关键字查询 $emb(w \in W_i)$，另外还嵌入每个段落 $emb(s) \in S_i$。

形式上，我们被给定一个方面节点 $t_i^h$ ，它是父节点 $t_h$ 的子节点和 $t_j \in T_{\neq i}^h$ 的兄弟节点。我们还提供一个段嵌入 $emb(s) \in S_i$ ，所有关键词查询嵌入 $t_i^h$ ， $emb(w) \in W_i$ ，以及所有兄弟关键词查询嵌入 $emb(w) \in W_{\neq i}^h$ 。我们基于以下内容计算判别排名：

**Definition 1 (TARGET SCORE)** 一个片段 $s_i$ 根据其与所有关键词 $w \in W_i$ 的相似度的加权平均值（$H$ ）进行奖励，这意味着对节点 $t_i$ 及其子方面进行了更深入的讨论。

$$p(s_i, W_i) = H\left(\left[sim\big(emb(s_i), emb(w)\big) \mid w \in W_i\right]\right),$$
$$\text{where } H(X) = \frac{\sum_{r=1}^{|X|} \frac{1}{r} x_r}{\sum_{r=1}^{|X|} \frac{1}{r}} \tag{1}$$

我们基于齐夫定律 (Powers, 1998) 计算加权平均，其中索引为 $r$ 位置的词会有权重 $1/r$ 。这种分段-关键词相似性的加权平均是基于一个假设，即模型将隐式地从最重要到最不重要生成关键词——换句话说，我们对第一个项 $w_1 \in W_i$ 赋予最高权重，而对 $w_{|X|}$ 赋予最低权重。例如，如果 $s_i$ 的相似性是 $[0.9, 0, 0]$ 到 $W_i = \{w_1, w_2, w_3\}$ ，那么 $p(s_i, W_i) = 0.5363$ 。另一方面，如果相似性是 $[0.7, 0.8, 0.7]$ ， $p(s_i, W_i) = 0.7272$ 。总体而言，目标分数将指示一个段落对方面节点 $t_i$ 的讨论深度——即它与多少关键词对齐以及达到何种程度。

**Definition 2 (DISTRACTOR SCORE)** 一个片段 $s_i$ 的惩罚基于讨论的兄弟节点的广度和深度。广度由 $s_i$ 与 $t_j \in T_{\neq i}^h$ 的每个 $W_j$ 之间的平均目标评分指标示。深度由 $s_i$ 与 $t_j \in T_{\neq i}^h$ 的每个 $W_j$ 之间的最大目标评分指标示。

$$n(s_i, T_{\neq i}^h) = 0.5 \times \left(\frac{1}{|T_{\neq i}^h|} \sum_{j=1}^{|T_{\neq i}^h|} p(s_i, W_j)\right) + 0.5 \times \left(max_{j=\left[1, |T_{\neq i}^h|\right]} \left(p(s_i, W_j)\right)\right) \tag{2}$$

我们利用目标和干扰项的得分来计算整体区分性得分，该得分衡量一个片段与其目标方面的接近程度，相对于其与干扰项、同类方面的整体和个体接近程度。

**Definition 3 (DISCRIMINATIVENESS SCORE)** 一个片段 $s_i$ 的奖励基于其与所有关键词 $w \in W_i$ 的相似度的加权平均值（$H$ ），同时基于讨论的兄弟节点的广度和深度进行惩罚。

$$d(s_i, W^h) = \frac{\beta \times p(s_i, W_i^h)}{\gamma \times n(s_i, T_{\neq i}^h)} \tag{3}$$

在公式 3 中， $\mathrm{d}\,(s_i, W^h)$ 的增长与目标得分成正比，而与干扰项得分成反比。我们为每个项目包括了 $\beta$ 和 $\gamma$ 缩放因子，以便用户可以根据需要自定义他们的奖励或惩罚的程度。最终，我们基于每段的可判别性得分对其进行排序，选取前 $k$ 段，它们包含了对目标方面 $t_i$ 的最丰富讨论，以便发现其子方面。

为了扩展我们的方面层级结构，我们迭代地利用我们的方面区分性检索作为知识基础，支持 LLM 的子方面发现。给定方面节点 $t_i$ ，其描述，其对应的区分性片段 $S_i$ ，以及根声明 $t_0$ ，我们提示模型为方面 $t_0$ 确定至少两个且最多 $k$ 个子方面的集合。我们在附录 A.3 中提供了此提示。每个子方面以第 **??** 节中指定的相同方式表示：其标签、描述和关键词。我们继续以自顶向下的方式构建我们的方面层级结构，详见算法 1 。

**Algorithm 1** 迭代子面相发现

**Require:** Root Claim $t_0$ , Corpus $D$ , max_depth= $l$
$T$ = initialize_tree( $t_0$ ) { $T$ .depth = 0 }
$t_i^0 \in T^0 \leftarrow$ coarse_grained_aspects( $t_0$ ) {Section **??**}
$q$ = queue( $T^0$ )
**while** $len(q) > 0$ and $T$.depth $\leq l$ **do**
  $t_i \leftarrow pop(q)$
  enrich_node( $t_0, t_i, D$ ) {Section **??**}
  $S_i \leftarrow$ rank_segments( $t_0, t_i, D$ ) {Section **??**}
  $t_j^i \in T^i \leftarrow$ subaspect_discovery( $t_0, t_i, S_i$ )
  $q$ .append( $T^i$ )
**end while**
**return** $T$

最终，算法 1 的输出是我们的最终方面层次结构，作为我们共识确定和视角发现过程的基础。

## 2.2 基于分类的视角发现

随着方面层次结构的构建，我们必须识别 (1) 与根主张 $t_0$ 相关且 (2) 与层次结构 $T$ 中的一个方面节点对齐的完整语料段落集。锁定讨论方面节点 $t_i$ 的论文可以让我们推断它们对 $t_i$ 的看法，并评估共识的存在和程度。然而，正如在第 **??** 节中所指出的，我们不能假设所有语料段落都与根主张相关——这种假设在基于 LLM 的分类系统中已被使用。因此，我们必须首先剔除与主张无关的段落。

Filtering. 一个简单的方法是通过上下文提示确定每个节点的片段相关性，但这种方法的扩展性很差。相反，我们将相关性过滤设定为一个二分搜索问题，从而识别相关性与不相关性的边界。具体来说，我们嵌入了声明标签 $t_0$ ( $emb(t_0)$ ) 和每个子方面 $t_i^0 \in T^0$ ( $emb($ “ 关于 [ $t_0$ ]” ) )，计算声明表示为：

$$c_0 = \frac{1}{2}\left(emb(t_0) + \frac{\sum_{i=1}^{|T^0|} emb(t_i^0)}{|T^0|}\right) \tag{4}$$

我们通过余弦相似性对编码的片段进行排序以 $c_0$ ，并使用二分查找来找到索引 $r$ ，其中在 $\pm n$ 窗口中的片段少于 $\delta\%$ 个是相关的。这个排名 $r$ 作为我们的阈值，过滤掉排名较低的片段，仅保留那些与 $t_0$ （ $S_0'$ ）相关的片段。此优化显著减少了必要的相关性判断数量；相关提示在附录 A.3 中。

有了 $S_0'$ 和 $T$ ，我们应用分类学指导的层次分类来确定每个方面节点 $t_i \in T$ 的 $S_i'$ 。由于我们的重点是检索指导的方面层次结构构建而不是分类，我们采用了最近的基于 LLM 的层次分类模型 (Zhang et al., 2024a) ，该模型通过丰富分类学节点（例如，添加关键词）来支持其从上到下的 $S_i'$ 到 $t_i$ 的分类。

Perspective & Consensus Discovery. 我们流程的最后一步是确定针对每个方面节点 $t_i$ 的主要视角 $P_i = \{a_i, o_i\}$ ，其中 $a_i$ 是总体支持性视角， $o_i$ 是反对性视角。我们还寻求识别持有这些视角的论文

($D_i = D_i^{\text{supp}} \cup D_i^{\text{opp}} \cup D_i^{\text{neutral}}$)，并包括那些对 $t_i$ 没有明确视角的论文。

**Definition 4 (PERSPECTIVE)** 以隐含或明确的立场对权利要求 $t_0$ 的特定方面 $t_i$ 表达的一种描述性观点，涉及对 $t_i$ 的态度（例如，支持、中立或反对）以及可选的理由。

我们不假设 $D_i^{\text{supp}}, D_i^{\text{opp}}$, and $D_i^{\text{neutral}}$ 是非重叠的，因为它们可能有多个段落表示不同的观点。例如，一个映射到“老年人安全”的段落 $s_i' \in S_i'$ 可能讨论了一项临床试验，显示老年患者在接种疫苗 A 时增加了过敏性休克。然而，同一篇文章中的另一段可能也提到疫苗 B 导致严重的荨麻疹。因此，我们允许这种灵活性。

最近研究表明，LLMs 在立场检测方面表现出强大的能力 (Zhang et al., 2024b; Lan et al., 2024) 。因此，为了发现这些观点，我们提示模型首先确定每个片段的立场 $s_i' \in S_i'$ ：

- 支持声明：$s_i'$ 无论是隐含地还是明确地表明，关于 $t_i$ 的声明是正确的。
- 中立于断言：$s_i'$ 与该断言和方面相关，但并未表明该断言对于 $t_i$ 是否为真。
- 反对断言：$s_i'$ 以明示或暗示的方式指出关于 $t_i$ 的断言是错误的。

这构成了段落集：$S_i^{\text{supp}}$ 、$S_i^{\text{neutral}}$ 和 $S_i^{\text{opp}}$ 。我们要求模型总结每个段落集的观点（立场和理由）：$P_i$ 。所有提示信息均在附录 A.4 中提供。由于我们保留了每个段落的原始论文来源，我们能够为每个节点 $t_i$ 构建 $D_i$ 。这表示共识；例如，$D$ 中有多少篇论文在 $t_i$ 方面持有 $p_i^{\text{supp}}$ 观点。作为我们的最终输出，我们有方面层次结构 $T$ ，观点集 $P_i$ 及其对应的论文 $D_i$ 。

我们探讨了 CLAIMSPECT 在开源模型 `Llama-3.1-8B-Instruct` ( ) 上的性能。我们从位于顶部的 1 % 采样，并根据给定任务的性质设置温度（所有样本都使用相同的设置）；这些设置包括在附录 B 中。我们将方面层次结构的最大深度设置为 $l = 3$ 。

### 2.3 数据集

为了评估 CLAIMSPECT 将细微主张解构为一个方面层级并识别其相应观点的能力，我们构建了两个新颖的大规模数据集，专门用于我们的任务，应用于政治（世界关系）和科学（生物医学）领域。为了构建这个数据集，我们首先手动收集了 ~ 50 个参考材料，讨论（1）与安全相关的国际冲突，以及（2）与生物医学安全相关的研究。然后，我们使用 `GPT-4o` (OpenAI et al., 2024) 基于这些材料生成细微的主张。随后，我们使用 Semantic Scholar API (Allen Institute for AI, 2025) 收集与这些主张相关的文献的元信息。然后，基于这些元信息，我们过滤所收集的文献并检索全文。通过这种方式，对于每个主张，我们获得了对应的文献库。我们在表格 1 中展示了这些数据集的统计信息。关于数据集构建的更多细节可以在附录 C 中找到，包括在附录 **??** 中用于验证生成主张及其关联论文质量的人类研究。

| Dataset | Claims | Papers | Segments |
|---|---|---|---|
| World Relations | 140 | 9,525 | 1,081,241 |
| Biomedical | 50 | 3,719 | 428,833 |
| Total | 190 | 13,244 | 1,510,074 |

Table 1: 每个数据集的"claims"、"papers"和"segments" 的数量。

我们的主要动机是展示 CLAIMSPECT 在将一个复杂的陈述分解为一个层次结构并识别相应观点的能力。然而，目前没有现有的方法来处理这一新颖任务。因此，我们选择实现并比较我们的方法与 RAG 驱动和仅限 LLM 的方法，详细信息如下。我们使用 Llama ( ) 和 GPT-4o-mini ( ) 运行每个基线：

我们还进行了一项消融研究，称为没有辨别力 (No Disc)，在这项研究中，我们移除辨别性排序，改用基于语义相似度的排序。为此，我们计算每个段和我们从第 **??** 节中的 $t_i$ 特定查询之间的语义相似度。

### 2.4 评估指标

我们设计了一个全面的自动化评估套件，使用 `GPT-4o-mini` 来确定我们生成的分类法的质量，同时使用节点级和分类法级的指标。对于每个判断，我们要求 LLM 提供额外的理由：

- （节点级）节点相关性：对于层次结构中的每个方面节点 $t_i$ 及其相应路径，它与论点 $t_0$ 的相关性如何？评分为 0/1。
- （节点级）路径粒度：到节点 $t_i$ 的路径是否保留其实体之间的层次关系（每个子节点 $t_j^i$ 是否比父节点 $t_i$ 更具体）？评分为 0/1。
- （按层级划分）兄弟粒度：对于层次结构中每组兄弟节点 $T^i$ ，整体集合相对于其父节点 $t_i$ 的特定程度是否相同？评分从 1 到 4（全部不同 → ，有些不同 → ，大多数不同 → ，全部相同）。
- （节点层次）唯一性：方面节点 $t_i$ 是否在层次结构 $T$ 中有其他重叠节点？评分 0/1。
- （节点级）段质量：有多少段落 $s \in S_i'$ 与声明 $t_0$ 和方面 $t_i$ 相关？我们计算每个节点相关段落的平均比例。

除了自动评估我们的方面层次结构之外，我们还对 CLAIMSPECT 从语料库中识别出的 50 个观点及其抽样段落进行了补充的人为评估（第 3.2 节）。

## 3 实验结果

### 3.1 整体性能 & 分析

表 2 - 3 展示了对比基线，CLAIMSPECT 在世界关系和生物医学数据集上的各种节点和层级指标中的几个关键优势。CLAIMSPECT 能够强烈地维护生成的方面层级的层次结构，同时保持与语料库的相关性。下面，我们展示我们的核心发现和见解。我们还在附录 **??** 中提供了 ClaimSpect 计算效率的详细分析。最后，我们在附录 **??** 中进行了人机自动评估一致性研究。

Table 2: ClaimSpect 与所有基准的比较。兄弟节点粒度（Sib）已被标准化；所有其他度量均缩放为 100。由于 Iterative Zero-Shot 没有与语料库关联，因此每个节点没有相关的片段。因此，我们省略其片段相关性得分（Seg）。我们将最高得分加粗显示，第二高得分用 underline 表示。

| Method | World Relations | | | | | Biomedical | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rel | Path | Sib | Unique | Seg | Rel | Path | Sib | Unique | Seg |
| Iterative Zero-Shot | 97.85 | 41.94 | 58.01 | 72.96 | — | 98.33 | 44.44 | 57.04 | 77.17 | — |
| Iterative RAG | 97.18 | 45.34 | 59.01 | 74.25 | 42.79 | 97.14 | 45.93 | 59.08 | 76.17 | 27.11 |
| Iterative Zero-Shot | 98.60 | 42.88 | 64.04 | 76.01 | — | 97.89 | 41.56 | 62.09 | 77.55 | — |
| Iterative RAG | 97.40 | 52.30 | 66.45 | 76.59 | 46.93 | 94.37 | 50.07 | 64.21 | 77.05 | 31.82 |
| CLAIMSPECT | 95.30 | 78.24 | 85.26 | 87.62 | 43.23 | 97.95 | 75.10 | 74.80 | 86.26 | 27.39 |
| CLAIMSPECT - No Disc | 99.00 | 79.75 | 82.64 | 85.43 | 49.47 | 96.07 | 76.26 | 74.39 | 87.69 | 39.03 |

Table 3: 针对每个数据集的所有方法进行成对比较。每个值表示在每个数据集内被认为该方法更好的样本百分比。E-Tie 表示显式平局；I-Tie 表示隐式平局。

| Method Pair (A vs. B) | World Relations | | | | Biomedical | | | |
|---|---|---|---|---|---|---|---|---|
| | A Wins | B Wins | E-Tie | I-Tie | A Wins | B Wins | E-Tie | I-Tie |
| Zero-Shot vs RAG | 0.00 | 33.06 | 0.00 | 66.94 | 2.22 | 22.22 | 0.00 | 75.55 |
| Zero-Shot vs CLAIMSPECT | 0.00 | 97.58 | 0.00 | 2.42 | 0.00 | 95.55 | 2.22 | 2.22 |
| RAG vs CLAIMSPECT | 0.81 | 90.32 | 0.00 | 8.87 | 0.00 | 95.55 | 0.00 | 4.44 |
| No Disc vs CLAIMSPECT | 21.43 | 30.00 | 0.00 | 48.57 | 24.00 | 28.00 | 0.00 | 48.00 |
| Zero-Shot vs RAG | 0.00 | 36.00 | 0.00 | 64.00 | 0.71 | 47.14 | 0.00 | 52.14 |
| Zero-Shot vs CLAIMSPECT | 0.00 | 98.00 | 0.00 | 2.00 | 0.00 | 96.43 | 0.00 | 3.57 |
| RAG vs CLAIMSPECT | 0.00 | 90.00 | 0.00 | 10.00 | 7.14 | 72.14 | 0.71 | 20.00 |

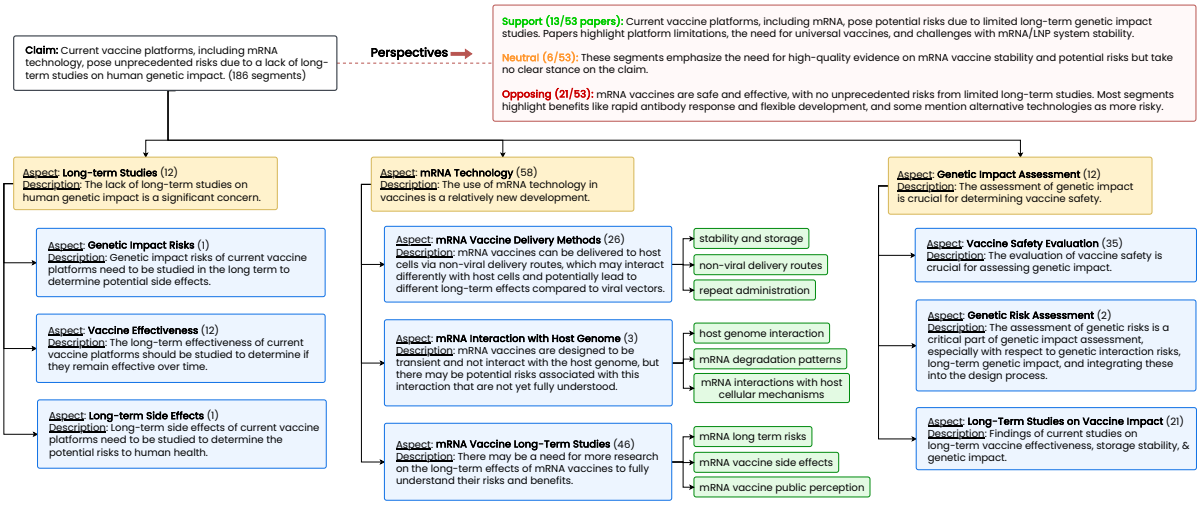


Figure 3: 构建的生物医学方面层次结构。包括所有节点及其从第 1 层到第 2 层的 # 段；突出显示了第三级的一部分。括号中提供了映射到每个视角的论文的 #。

如表格 2 所示，CLAIMSPECT 在与节点层次结构相关的指标上显著优于基准，特别是在保持层次关系（路径粒度）上比 Iterative RAG 的表现好 72.6% 和 63.51%，并在保持同级特异性均匀性（同级粒度）上分别在两个数据集上表现出色，超过了 44.48% 和 26.61% 。这表明该方法能够在目标粒度水平上检索和组织方面。即使依赖于一个闭源模型，这种提升在基于 GPT 的基准中也同样明显。我们将这种提升归因于 ClaimSpect 的迭代子方面发现（章节 **??** ）与其方面判别检索机制（章节 **??** ）的整合，其中，基础子方面发现的片段池根据给定的方面节点进行迭代更新。我们可以看到，无判别消融在粒度质量上确实有一些损失。需要注意的是，无判别有时表现出竞争力，甚至更好的性

能；这可能是因为它考虑了更多的片段，这些片段可能涉及或不涉及多个方面。相比之下，基准方法检索到的是更广泛的、关注度较低的片段，减少了它们发现细粒度子方面的能力。总体而言，这表明 ClaimSpect 能够将一个主张分解成一个结构良好的方面层次结构。

在表 2 中，我们观察到 ClaimSpect 构建的层次结构的节点在每个数据集上分别比顶级基线更加 14.40% 和 11.23% 独特。这表明 ClaimSpect 的层次结构在方面质量上更加丰富，体验到树中不同方面之间的重叠更少，并且伴随着段落质量的提升。尽管独特性显著提高，但 ClaimSpect 仅在世界关系和生物医学数据集上的方面节点相关性较顶级基线分别下降了 3.35% 和 0.386% 。这强调了 ClaimSpect 的检索增强型关键词扩充和方面判别检索（章节 **??** 和 **??** ）的强大功能，它优先考虑深入讨论单个方面的段落，而不是浅显描述多个方面的段落。这使我们能够在层次结构的每个级别上发现一组更加丰富且相关的独特子方面。

CLAIMSPECT is overwhelmingly preferred over baselines. 表 3 展示了 CLAIMSPECT 和基线方法之间的成对比较。这些比较由一个大型语言模型（LLM）进行判断，该模型展示的是方法 $A$ 和 $B$ 的方面层级输出，且以两种可能的顺序进行：$A$ vs. $B$ 和 $B$ vs. $A$ 。LLM 可能会 (1) 偏好方法 $A$ 或 $B$ 其中之一，(2) 宣布一个明确的平局 (E-Tie)，或者 (3) 表示一个隐含的平局 (I-Tie)，这种情况发生在根据展示顺序而改变偏好方法时（例如，在 $A$ vs. $B$ 中 $A$ 获胜，但在 $B$ vs. $A$ 中 $B$ 获胜）(Shi et al., 2024) 。

在两个数据集中，CLAIMSPECT 展现了明显的优势，被偏好为 92.95% 的时间，在所有设置和数据集中的平均不一致率为 6.69% 。具体来说，与 Zero-Shot 相比，CLAIMSPECT 在案例和生物医学数据集中分别被认为在 97.58% 和 95.55% 的情况下更为优越。即使与 RAG 相比，CLAIMSPECT 在样本中的表现也优于 90.00% 和 72.14% 。这与两个基线之间缺乏强烈偏好的形成鲜明对比，显示为 64.66 % 平均隐式平局率——暗示两者之间没有明显的质量偏好。最后，我们显示 ClaimSpect 和 No Disc 同样受到偏好，且 ClaimSpect 略微更常被偏好。总体而言，这些结果验证了 CLAIMSPECT 构建的论点方面层次结构在相关性上显著更有意义。

### 3.2 视角发现分析

在图 3 中，我们展示了一种细化主张方面层次结构的定性分析，突出显示了某些子树和根节点的提取视角。我们观察到，每一个粗粒度方面（黄色节点）都很好地代表了在验证根主张时考虑的各个角度：目前存在的长期疫苗研究是什么，当前的 mRNA 技术是什么，以及如何评估遗传影响？我们看到，路径特定的依赖性反映在每个方面的描述中（例如，“mRNA 与宿主基因组的相互作用”涉及 mRNA 技术和潜在的遗传影响风险）。此外，即使在层次结构的最终层，这些层次关系和主张相关性仍得以保留（例如，“mRNA 与宿主基因组的相互作用” → “mRNA 降解模式”）。最后，我们看到映射到根节点的视角是信息丰富的，为每种立场提供了正当理由。注意，ClaimSpect 将段落映射到每个视角，允许我们识别原始论文来源，并最终提供一个特定语料库的共识估计。总体来说，这种解构的主张视图提供了一种识别某些方面被探索过的哪些以及在何种程度上的手段（例如，mRNA 技术在语料库中相对于遗传影响评估被更多地探索过）。

| $k$ | World Relations | Biomedical |
|---|---|---|
| 5 | 72 % (50) | 72 % (50) |
| 10 | 80 % (20) | 82 % (20) |
| 15 | 85 % (19) | 89 % (9) |

Table 4: 人类对由 CLAIMSPECT 发现的视角百分比的验证，这些视角至少在一个 $k$ 关联片段中有依据。考虑的片段数为 $k = \#$。在括号中提供每种设置的样本数量。

Human annotators validate the grounding of discovered perspectives. To assess the validity of the perspectives discovered by CLAIMSPECT, we apply human evaluation to evaluate whether these perspectives are effectively grounded in the corpus. We randomly sampled perspectives along with their associated $k$ segments (each aspect node has three ass from the generated results across two datasets. The evaluation metric used was whether at least one segment in $k$ could provide grounding background knowledge for the corresponding perspective . As shown in Table 4, we found that the vast majority of cases (85 % and 89 % for each dataset respectively) are supported by specific literature segments. Furthermore, we can see that the support rate steadily increases as we retrieve more segments that are mapped to the perspective. This shows the perspectives identified by CLAIMSPECT are largely supported by the corpus .

## 4 结论

我们的工作引入了 CLAIMSPECT ，这是一个将复杂主张分解为语料库特定方面和观点层次结构的新框架。通过结合迭代的、方面区分的检索、层次子方面发现和观点聚类，CLAIMSPECT 提供了复杂主张的结构化、全面视图。我们在两个新的大规模数据集上的实验表明，CLAIMSPECT 构建了丰富且与语料库对齐的方面层次结构，并且充满了多样且具有信息量的观点。这突显出其作为一个可扩展和适应性强的方法在跨领域细致主张分析中的有效性。

## 5 限制与未来工作

CLAIMSPECT 的主要贡献是我们提出的用于构建与验证细致主张相关的方面层次结构的检索增强框架。为了展示该层次结构的潜力，我们将其应用于视角发现任务中，涉及（1）识别语料库中与给定方面节点相关的段落，（2）确定段落对主张和方面的立场（或缺乏立场），以及（3）发现每个基于立场的段落簇的潜在视角。因此，此步骤在很大程度上依赖于现有的层次分类模型 (Zhang et al., 2024a) ，因为我们不声称在分类方面有创新。同样，我们基于分类的视角发现（第 2.2 节）依赖于 LLM 的细粒度立场检测能力——尽管之前的工作 (Zhang

et al., 2024b; Lan et al., 2024) 已显示出其能力的先例。因此，层次分类和立场检测的性能是我们方法性能的瓶颈。例如，如果基于 LLM 的立场检测在检测支持主张某方面的段落时具有很高的召回率但精度较低，那么这种方法可能会高估语料库中某一视角背后的共识。同样，如果检测具有很高的精度但较低的召回率，可能会低估共识。尽管如此，我们的工作总体上旨在激励在直接进行验证之前需要构建某些细致主张的方面结构。

分层分析细微的主张为许多新的研究方向打开了大门。首先，CLAIMSPECT 可以与更系统和/或工具集成的事实验证系统整合，以努力构建一个更健壮的事实核查系统。此外，CLAIMSPECT 可以应用于更有针对性的检索或问答任务，在这些任务中，类似于细微主张的问题难以被简单回答，并且可能从更结构化的输出（类似于方面层次结构）中受益。

## 6 致谢

本工作得到了国家科学基金研究生研究奖学金的支持。部分工作还得到了 BRIES 项目编号 HR0011-24-3-0325 的资助。本研究使用了由国家科学基金会（奖项 OAC 2320345）和伊利诺伊州支持的 DeltaAI 高级计算和数据资源。DeltaAI 是伊利诺伊大学厄巴纳-香槟分校及其国家超级计算应用中心的共同努力成果。我们感谢 Aptima 公司中的 Peter Bautista、Spencer Lynch 和 Svitlana Volkova 对我们工作的讨论。我们还感谢 Mihir Kavishwar 在早期创意讨论中的贡献。

## References

Allen Institute for AI. 2025. Semantic Scholar API. Accessed: 2025-02-15.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Zhen-Song Chen, Xuan Zhang, Rosa M Rodríguez, Witold Pedrycz, Luis Martínez, and Miroslaw J Skibniewski. 2022. Expertise-structure and risk-appetite-integrated two-tiered collective opinion generation framework for large-scale group decision making. *IEEE Transactions on Fuzzy Systems*, 30(12):5496–5510.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Google Scholar. 2025. Google Scholar. https://scholar.google.com. Accessed: 2025-02-15.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria. INCOMA Ltd.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Saif M Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jason Alan Palmer. 2024. pdftotext.

Nicolas Perony, René Pfitzner, Ingo Scholtes, Claudio J Tessone, and Frank Schweitzer. 2013. Enhancing consensus under opinion bias by means of hierarchical decision making. *Advances in Complex Systems*, 16(06):1350020.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

David M. W. Powers. 1998. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*.

PubMed. 2025. PubMed. https://pubmed.ncbi.nlm.nih.gov. Accessed: 2025-02-15.

Xiaoyu Shen, Rexhina Blloshmi, Dawei Zhu, Jiahuan Pei, and Wei Zhang. 2024a. Assessing "implicit" retrieval robustness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8988–9003, Miami, Florida, USA. Association for Computational Linguistics.

Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. 2024b. A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. *arXiv preprint arXiv:2402.13405*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. RedHOT: A corpus of annotated medical questions, experiences, and claims on social media. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 809–827, Dubrovnik, Croatia. Association for Computational Linguistics.

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2024. Codetaxo: Enhancing taxonomy expansion with limited examples via code language prompts. *arXiv preprint arXiv:2408.09070*.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024a. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *arXiv preprint arXiv:2403.00165*.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024b. Llm-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378.

我们进行了一项人工评估研究，以展示人类和 `GPT-4o-mini` 之间的评价一致性。我们使用两名人工评估员来评估我们提出的每个度量标准的随机样本案例：节点相关性、路径粒度、兄弟粒度、唯一性和片段质量。我们的人工评估员是两位志愿的研究生（一个博士生和一个硕士生），其中一位具有生物学和自然语言处理背景（这对我们的生物医学任务评估至关重要）。我们使用以下指示来帮助指导他们完成评估任务，评估员们首先经历了一个“培训”阶段，以便熟悉任务（例如，方面分类和它们的预期层次关系）并彼此讨论示例，整个评估阶段则是独立进行的。

由于 ClaimSpect 在“节点相关性”和“唯一性”方面始终保持高分（整个分类体系中没有太多节点是无关的或重叠的），因此来自 LLM 和评估者的分数方差都非常低。这极大地限制了 Cohen 的和类内相关系数（高不稳定性）。尽管如此，评估者分别有 100% 和 96.97% 的协议率。我们根据度量的顺序/连续性与分类性质选择 Cohen 的 $\kappa$ 或 ICC。

1. 路径粒度的加权 $\kappa$ 为 0.62 → 显著一致
2. 兄弟粒度的 $ICC1k$ 具有 0.7806 → 良好的可靠性
3. 段质量的 $ICC2k$ 是 0.7578 → 良好的可靠性

我们还在表 5 中展示了 100 个不同样本在所有指标上的一致率。

| | Rel | Path | Sib | Unique | Seg |
|---|---|---|---|---|---|
| Agreement Rate | 100 % | 85 % | 87.5 % | 96.97 % | 82 % |

Table 5: 人类-自动化一致性在不同指标中的比率。

我们可以看到，LLM 评估与人工评审者具有高度的一致性。我们注意到，段落质量的对齐度最低（尽管仍然相对较高），这可能是由于验证段落与父节点对齐所需的更细粒度的文本理解能力（通常情况下，段落之间在语义上可能相当不同，或仅讨论节点的“子方面”）。尽管如此，透过这些结果，我们可以看到 ** 我们的自动评估是可靠的 **。

# A 提示模板

在本节中，我们介绍了在 CLAIMSPECT 的不同模块中使用的提示。

## A.1 粗粒度方面发现

这是用于生成根主张的粗粒度方面的提示，包括它们的标签、描述和相关关键词，以构建初步的检索增强层次结构。

> Prompt
>
> For the topic, { topic } , output the list of up to { k } aspects in JSON format.

## A.2 检索增强关键词扩展

以下是用于检索增强关键词丰富的提示，指导 LLM 优化和筛选特定方面的关键词，以改进段落排名。

> Prompt (Extraction)
>
> The claim is: { claim } . You are analyzing it with a focus on the aspect { aspect_name } . The aspect, { aspect_name } , can be described as the following: { aspect_description }
> Please extract at most { 2*max_keyword_num } keywords related to the aspect { aspect_name } from the following documents: { contents }
> Ensure that the extracted keywords are diverse, specific, and highly relevant to the given aspect.
> Only output the keywords and seperate them with comma. Your output should be in JSON format.

> Prompt (Filtering)
>
> Our claim is ' { claim } '. With respective to the target aspect ' { aspect_name } ', identify { min_keyword_num } to { max_keyword_num } relevant keywords from the provided list: { keyword_candidates } .
> { aspect_name } : { aspect_description }
> Merge terms with similar meanings, exclude relatively irrelevant ones, and output only the final keywords separated by commas.
> Your output should be in JSON format.

## A.3 迭代子方面发现

以下是用于迭代引导大型语言模型（LLM）发现和扩展每个方面节点的子方面的提示，该提示基于判别检索和根声明上下文。

> Prompt
>
> Output the list of up to { k } subaspects of parent aspect { aspect } that would be considered when evaluating the claim, { topic } . claim: { topic } parent_aspect: { aspect } ; { aspect_description } path_to_parent_aspect: { aspect_path } Provide your output in the following JSON format.

以下是用于相关性筛选的提示，利用基于余弦相似度排序的二分搜索来高效识别并仅保留每个方面最相关的片段。

Prompt

I am currently analyzing a claim based on a segment from the literature from several different aspects. The segment is: { segment } The claim is: { claim } The aspects are: { aspects } Please help me determine whether this segment is related to the claim so that I can analyze this claim based on it from at least one of these aspects. Your output should be 'Yes' or 'No' in JSON format.

### A.4 视角发现

以下是用于确定段落立场（支持、中立或反对）和总结每个方面的观点（包括基本原理）的提示。

Prompt

You are a stance detector, which determines the stance that a segment from a scientific paper has towards an aspect of a specific claim. Oftentimes, scientific papers do not provide explicit, outright stances, so your job is to figure out what stance the data or statement that they are presenting implies. Segment: { segment.content }
What is the segment's stance specifically with respect to { aspect_name } for if { claim } ? { aspect_name } can be described as { aspect_description } . Claim: { claim } Aspect to consider: { aspect_name } : { aspect_description } Path to aspect: { aspect_path }
Your stance options are the following: - supports_claim: The segment either implicitly or explicitly indicates that claim is true specific to the given aspect. - neutral_to_claim: The segment is relevant to the claim and aspect, but does not indicate whether the claim is true specific to the given aspect. - opposes_claim: The segment either implicitly or explicitly indicates that the claim is false specific to the given aspect. - irrelevant_to_claim: The segment does not contain relevant information on the claim and the aspect.

## B 生成设置

本节详细说明了在我们过程的各个阶段中使用的温度值及其各自的作用。

### B.1 温度设置概述

- 粗粒度方面发现 (0.3)：用于生成与观点相关的高层次方面。较低的温度确保结构化和确定性的输出。
- 子方面发现 (0.7)：用于从排序段中识别子方面。更高的温度允许在保持连贯性的同时增加多样性。
- OpenAI 聊天模型（GPT-4o (OpenAI et al., 2024)，GPT-4o-mini (OpenAI et al., 2024)）(0.3)：应用于使用 GPT-4o 模型的各个阶段（例如，方面生成、分类），以确保响应的一致性。
- 子方面发现（方面排名和检索）(0.7)：用于从排名段中提取子方面，以平衡创造性与相关性。

### B.2 一般趋势

- 较低的温度（0.3）用于结构化和确定性的任务，例如方面生成和分类。
- 更高的温度（0.7）被应用于子方面的发现，在这种情况下，多样性和探索是有益的。

## C 数据集构建

为了评估 CLAIMSPECT 的有效性，我们建立了两个涵盖关键领域的数据集：政治（世界关系）和科学（生物医学）。数据集的构建过程包括以下步骤：

### C.1 人工种子采集

我们首先从可靠来源（例如 Google Scholar (Google Scholar, 2025) 和 PubMed (PubMed, 2025)）手动收集一组种子论点。具体来说，我们从世界关系领域的 7 篇论文和生物医学领域的 50 篇论文中收集材料。这些初始材料作为生成细致论点的背景或特定主题。

利用前一步中收集的文献并将细微主张的定义作为背景，我们提示 `GPT-4o` (OpenAI et al., 2024) 生成与这些论文主题相关的细微主张。为确保主张视角的多样性，我们使用两组提示：一组用于生成与文献中观点一致的主张，另一组用于生成与之相异的主张。具体的提示如下所示。

Positive Claim Generation Prompt

Scientific or political claims are often nuanced and multifaceted, rarely lending themselves to simple "yes" or "no" answers. To answer such questions effectively, claims must be broken into specific aspects for in-depth analysis, with evidence drawn from relevant scientific literature. We are currently studying such claims using this corpus:
{ context }
Task: Generate 10 nuanced and diverse claims based on this corpus. The claims should adhere to the following criteria:
1. Diversity: The claims should be sufficiently varied: they should involve diverse sub-topics in the context
2. Complexity: The claims should be complex and controversial (and not necessarity true), requiring multi-aspect analysis rather than simplistic treatment. Avoid overly straightforward or simplistic claims.
3. Research Feasibility: The claims should not be too specific and should pertain to topics with a likely body of existing literature to support evidence-based exploration.
4. Concision: The claims should be concise and focused in one short sentence.
5. Completeness: The claims should be complete and not require additional context to understand.
Output: Provide the claims as a list.

Negative Claim Generation Prompt

Scientific or political claims are often nuanced and multifaceted, rarely lending themselves to simple "yes" or "no" answers. To answer such questions effectively, claims must be broken into specific aspects for in-depth analysis, with evidence drawn from relevant scientific literature. We are currently studying such claims using this corpus:
{ context }
Task: Generate 10 nuanced and diverse claims based on this corpus. The claims should adhere to the following criteria:
1. Diversity: The claims should be sufficiently varied: they should involve diverse sub-topics in the context
2. Complexity: The claims should be complex and controversial (and not necessarity true), requiring multi-aspect analysis rather than simplistic treatment. Avoid overly straightforward or simplistic claims.
3. Research Feasibility: The claims should not be too specific and should pertain to topics with a likely body of existing literature to support evidence-based exploration.
4. Concision: The claims should be concise and focused in one short sentence.
5. Completeness: The claims should be complete and not require additional context to understand.
6. The claims should be against the point of view in the context.
Output: Provide the claims as a list.

我们发现生成的细致声明质量很高，它们内容丰富、具体，并且难以简单地归类为真或假，这与我们的任务要求非常吻合。以下是我们数据集中的一些示例性声明。

Claims for World Relations

1. International collaborations under the Global Nuclear Security Program prioritize geopolitical alliances over immediate nuclear threat reduction.

2. Counteracting WMDs through international partnerships creates dependency and may hinder national self-sufficiency in threat reduction capabilities.

3. The effectiveness of the biological threat reduction component is questionable given the rise and global spread of emerging biological threats.

Claims for Biomedical Domain

1. COVID-19 vaccine safety evaluations are compromised by inconsistent application of evidence standards across different data sources like RCTs and VAERS.

2. The rigid adherence to optimized distribution plans might inhibit the flexibility needed to respond to unforeseen disruptions in the vaccine supply chain.

3. Keeping manufacturing costs secret is essential for protecting proprietary processes and innovations in the pharmaceutical industry.

为了支持基于语料库的每个论点分析，我们使用 Semantic Scholar API (Allen Institute for AI, 2025) 检索相关文献。

由于我们的论断非常微妙，并涉及多个概念，直接搜索论断本身并不能根据文献的标题和摘要找到有用的匹配。为了解决这个问题，我们首先对每个论断进行关键词提取。然后，我们使用提取的关键词来查询 Semantic Scholar API，并为每个论断检索最多 1000 个相关的文献条目。

在获取文献元数据后，我们首先筛除缺少字段的条目，并根据相关性保留最相关的前 100 篇论文。然后，我们利用提供的 PDF URL 下载所选文献的全文，并使用 pdftotext (Palmer, 2024) 将其转换为纯文本。结果，我们为每个论点获得了一个全面的文本文献库，以确保进一步分析的丰富上下文基础。

这种结构化的方法确保了一个稳健的数据集，适合在各个领域进行细微的索赔分析。

我们进行了一个人为评估研究，用于验证总共 40 个论断——每个数据集 20 个。我们为论断验证定义了以下二元标准：

我们在表 6 中展示了验证结果，表明生成的论点具有细微的特征、其相关性，以及与每个论点对齐的 $k$ 论文的存在。

| Metric | World Relations | Biomedical |
|---|---|---|
| Nuanced | 0.9 | 1.0 |
| Relevant | 1.0 | 1.0 |
| Corpus-Aligned@5 | 0.95 | 0.8 |
| Corpus-Aligned@10 | 0.65 | 0.65 |

Table 6: 跨数据集的人类对主张质量的验证。

我们在下面指定 ClaimSpect 框架的组件及其在整个流程中的相应计算效率。我们将完整方面层次结构中的节点数量视为 $n$，将语料库中的总段数视为 $S$。此外，我们基于平均样本提供粗略的时间估计。

- 粗粒度方面发现（第 **??** 节）
  - 单次 LLM 调用：$O(1)$
- 特征辨别检索（第 **??** 节）
  - 检索增强的关键词扩展（第 **??** 节）
    - 使用检索模型嵌入段落花费最多的时间（$O(S)$），但这可以在给定知识库的情况下离线计算。检索本身是相当高效的，因为它是基于嵌入的，我们使用余弦相似性来确定相关性（在高维场景中这种计算尤其高效）。
    - 丰富每个节点：$O(N) \rightarrow$ 每节点 10 秒
  - 判别段排序 (第 **??** 节)
    - 我们只计算前 100 个片段的排名，因此此操作的效率是常数：$O(1)$
    - 目标分数和干扰项分数计算根据整棵树的方面数量进行缩放（因为它们关联的关键词的 # 是恒定的）：每个节点 $O(N) \rightarrow 6$ 秒
- 迭代子方面发现 (第 **??** 节)
  - 层级结构中每个方面节点的单一提示：$O(N)$
- 基于分类的视角发现（第 2.2 节）
  - 如第 442-446 行所述，“我们将相关性过滤重新框架为一个二进制搜索问题”，这是为了有意优化该模块的效率。因此，我们不是逐个确定每个段落的声明相关性，而是对段落进行排序（这些段落的字符数超过 500），并使用二进制搜索（$O(\log S)$）来找到相关与不相关的界限。由于 Python 优化的原因，这种方式因排序函数在运行时耗时 $\rightarrow 5$ 秒从而提升了 $O(S \log S)$ 效率。
    - 观点发现牵涉到为每个经过筛选的片段提供大型语言模型的立场：$O(S) \rightarrow$ 3 分钟

我们特别使用 `vLLM` 来优化我们的 LLM 批量生成。构造一个具有 39 个节点和最大深度为 3 的声明，大约需要 20 分钟时间在两台 NVIDIA RTX A6000 显卡上运行。我们可以看到，总共核心框架的操作需要 13 分钟 42 秒，其余时间用于嵌入计算（这些计算可以离线进行）。