

# Beyond True or False: Retrieval-Augmented Hierarchical Analysis of Nuanced Claims

Priyanka Kargupta\*, Runchu Tian\*, Jiawei Han

Department of Computer Science, University of Illinois at Urbana-Champaign

{pk36, runchut2, hanj}@illinois.edu

## Abstract

Claims made by individuals or entities are oftentimes nuanced and cannot be clearly labeled as entirely “true” or false—as is frequently the case with scientific and political claims. However, a claim (e.g., “vaccine A is better than vaccine B”) can be dissected into its integral aspects and sub-aspects (e.g., efficacy, safety, distribution), which are individually easier to validate. This enables a more comprehensive, structured response that provides a well-rounded perspective on a given problem while also allowing the reader to prioritize specific angles of interest within the claim (e.g., safety towards children). Thus, we propose CLAIMSPECT, a retrieval-augmented generation-based framework for automatically *constructing* a hierarchy of aspects typically considered when addressing a claim and *enriching* them with corpus-specific perspectives. This structure hierarchically partitions an input corpus to retrieve relevant segments, which assist in discovering new sub-aspects. Moreover, these segments enable the discovery of varying perspectives towards an aspect of the claim (e.g., support, neutral, or oppose) and their respective prevalence (e.g., “how many biomedical papers believe vaccine A is more *transportable* than B?”). We apply CLAIMSPECT to a wide variety of real-world scientific and political claims featured in our constructed dataset, showcasing its robustness and accuracy in deconstructing a nuanced claim and representing perspectives within a corpus. Through real-world case studies and human evaluation, we validate its effectiveness over multiple baselines.

## 1 Introduction

Scientific and political topics increasingly being consumed in the form of concise, attention-grabbing claims which lack the nuance needed to represent complex realities (Vosoughi et al., 2018; Allcott and Gentzkow, 2017; Lazer et al., 2018). Such claims are frequently oversimplified or confidently stated, despite being valid only under specific conditions or when evaluated from

\*Equal contribution.

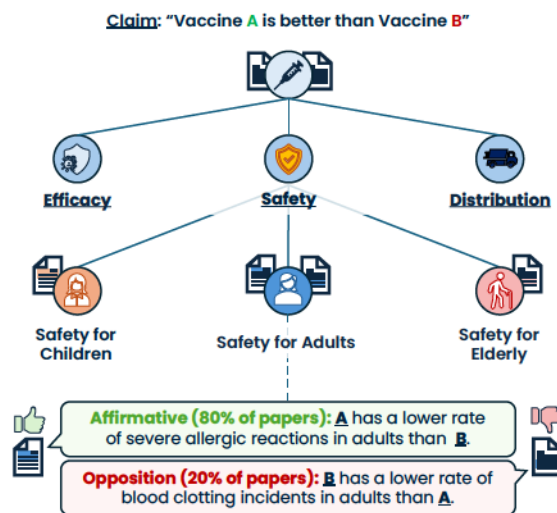


Figure 1: An example hierarchy of a nuanced claim being deconstructed into aspects. Each node is enriched with relevant excerpts, the affirmative/neutral/opposing perspectives, and their respective evidence.

certain perspectives. For instance, a claim like “vaccine A is better than vaccine B” may appear straightforward but becomes inherently nuanced when specific aspects, such as efficacy, safety, and distribution logistics, are considered. Moreover, the ambiguous and fragmented nature of information shared on such platforms often allows such claims to be twisted or reframed as “true” or “false” to support conflicting narratives, complicating the task of verifying their validity (Sharma et al., 2019; Pennycook and Rand, 2021).

Stance detection categorizes textual opinions as supportive, neutral, or opposing relative to a target (Mohammad et al., 2016). However, documents—especially those in a scientific domain—often present a *range of stances* across various aspects of a claim. For instance, as illustrated in Figure 1, a study might find Vaccine A safer for adults than Vaccine B while highlighting its significantly greater logistical challenges for widespread distribution. In this case, the paper supports the claim regarding “safety for adults” (note: not “safety” in its entirety) but opposes it concerning distribution. This complexity renders stance detection at the document level ineffective for nuanced, multifaceted claims.

Fact-checking models often validate claims by retriev-

ing evidence from large corpora or using web-integrated language models (Thorne et al., 2018; Popat et al., 2018; Zhang and Gao, 2023). While some methods now offer varied factuality judgements like “mostly true” or “half-true” (Zhang and Gao, 2023), these are less effective in scientific contexts. Especially in evolving areas, fine-grained scientific claims may be *unsubstantiated* due to a *lack of research* or *scientific consensus*, rather than being outright false. This distinction is vital, as it highlights areas needing further exploration. For example, in Figure 1, relevant paper excerpts mapped to the “Safety for Adults” node show that an 80:20 ratio of affirmative to opposing stances towards the sub-aspect claim suggests consensus, whereas a 60:40 ratio or sparse data signals limited research or disagreement. Such insights, crucial for understanding gaps in knowledge, are often **overlooked by existing fact-checking frameworks**.

We address these challenges using CLAIMSPECT, a framework which systematically deconstructs and analyzes claims by leveraging large language models (LLMs). ClaimSpect hierarchically partitions a claim into a tree of aspects and sub-aspects, enabling structured validation and the discovery of perspectives. This is accomplished by adopting the following principles:

**Principle #1: Claim trees capture the multidimensionality inherent in nuanced topics.** As opposed to considering a single target claim and the full document, we must first determine the relevant aspects discussed within the corpus itself in order to discover more targeted subclaims. However, it is essential to retain the hierarchical nature of such aspects. This is demonstrated in Figure 1, where certain aspects that are difficult to validate (e.g., “safety”) can typically be partitioned until they reach “atomic” sub-aspects that are more commonly considered (e.g., “safety for children”, “safety for adults”, and “safety for elderly”). Furthermore, these hierarchical relationships are often also reflected in how we naturally navigate formulating our own perspective towards a given topic (either individually or collectively): parse through the existing knowledge on a topic, consider different sub-angles of the problem based on this knowledge, retrieve more sub-angle specific knowledge, develop our opinions accordingly, and aggregate them to a high-level opinion (Perony et al., 2013; Chen et al., 2022). Thus, this brings us to our next principle.

**Principle #2: Iterative, discriminative retrieval enhances LLM-based tree construction.** LLMs have recently shown promise in automatic taxonomy enrichment and expansion, organizing data into hierarchies of categories and subcategories similar to our target aspect hierarchy (Shen et al., 2024b; Zeng et al., 2024b). However, these approaches often rely on general knowledge existing within the LLM’s pre-training dataset, overlooking corpus-specific insights crucial for (1) uncovering fine-grained sub-aspects prevalent in domain-specific data, and (2) ensuring alignment with the task of determining corpus-wide consensus. To address this, we leverage retrieval-augmented generation (RAG), which

has recently made advances in knowledge-intensive tasks by integrating external corpora or databases into the generation process (Lewis et al., 2020; Gao et al., 2023). We introduce an *iterative* RAG approach, which dynamically constructs the aspect hierarchy by retrieving relevant segments for an aspect node, using them to *discover new sub-aspects*. This ensures the taxonomy aligns closely with corpus-specific discussions of claims, aspects, and perspectives.

We note that noisy retrieval often hinders reasoning performance (Shen et al., 2024a). In our setting, this may occur when certain retrieved excerpts overlap multiple semantically similar aspect nodes (e.g., “safety for children” vs. “safety for adults”), introducing noise when determining sub-aspects for only one aspect. To mitigate this, we introduce a discriminative ranking mechanism that prioritizes segments discussing a *single aspect in-depth*, enhancing sub-aspect discovery and the final aspect hierarchy.

**Principle #3: Perspectives enrich understanding beyond stance and consensus.** For each aspect node in the hierarchy, we identify and cluster papers based on their stance (affirmative, neutral, opposing) using hierarchical text classification and stance detection. These clusters reveal not only the presence or absence of consensus but also the key perspectives within each stance. For example, as shown in Figure 1, the affirmative perspective might highlight Vaccine A’s lower rate of severe allergic reactions in adults, while the opposition focuses on its higher incidence of blood clotting. These perspectives offer transparency, uncover potential research gaps (e.g., if 80% of the affirmative papers do not address these blood clotting incidents), and provide critical context for framing nuanced claims.

Overall, CLAIMSPECT utilizes a structured approach to deconstruct a nuanced claim into a hierarchy of aspects, targeting a holistic approach considering all aspects which could be used to validate the root claim. The framework comprises the following steps: (1) *aspect-discriminative retrieval*, (2) *iterative sub-aspect discovery*, and (3) *classification-based perspective discovery*. Our contributions can be summarized as:

- From the best of our knowledge, CLAIMSPECT is the *first work* to formally deconstruct claims into a hierarchical structure of aspects to determine consensus.
- We construct **two novel datasets** of real-world, scientific and political *nuanced* claims and corresponding corpora.
- Through **experiments and case studies on real-world domains**, we demonstrate that ClaimSpect performs hierarchical consensus analysis significantly more effectively than the baselines.

**Reproducibility:** We provide our dataset and source code<sup>1</sup> to facilitate further studies.

<sup>1</sup><https://github.com/pkargupta/claimspect>



## 2 Related Works

**Fact Checking.** Fact-checking models (Thorne et al., 2018; Popat et al., 2018; Atanasova et al., 2019; Karadzhov et al., 2017) have leveraged external evidence to validate claims, but often treat claims as monolithic statements. Web-integrated methods (Zhang and Gao, 2023; Karadzhov et al., 2017) attempt to enrich fact-checking with additional context, but still fail to account for *nuanced* claims that cannot be clearly validated without considering a diverse range of claim sub-aspects and their varying levels of evidence. In contrast, CLAIMSPECT acknowledges the nuance behind certain claims, utilizing a corpus to help identify the various aspects that would be considered when validating a claim—enabling a more multi-faceted and interpretable analysis. We note that CLAIMSPECT does not aim to validate a given claim—it simply aims to *deconstruct* the claim into a hierarchy of aspects which *could be used to validate it, posing potential perspectives* towards the aspect of the claim, *grounded* in the corpus. We also note the adjacent task of evidence retrieval, where existing work explores organizing evidence according to a fixed and flat (non-hierarchical) set of aspects: populations, interventions, and outcomes (Wadhwa et al., 2023).

**LLM-Based Taxonomy Generation.** Recent advances in taxonomy generation (Shen et al., 2024b; Zeng et al., 2024b; Chen et al., 2023; Zeng et al., 2024a; Sun et al., 2024) have demonstrated the potential of large language models for structuring information hierarchically. However, these methods typically rely on static, domain-agnostic knowledge, limiting their adaptability to construct rich, fine-grained taxonomies (Sun et al., 2024). CLAIMSPECT addresses these limitations through corpus-aware, aspect-discriminative retrieval and iterative sub-aspect discovery, constructing a rich *taxonomy* of aspects that is aligned with a corpus. This allows us to identify the relevant segments to both a given aspect but also a perspective towards that aspect.

**Stance Detection** Traditional stance detection (Mohammad et al., 2016) classifies opinions as supportive, neutral, or opposing towards a target (e.g., claim). However, these approaches typically assign a single stance to an entire document, overlooking the nuanced, aspect-specific stances present within many claims, especially in scientific and political contexts. Recent works (Zhang and Gao, 2023) have introduced more fine-grained judgments (e.g., “mostly true”), but similar to fact-checking methods, they often fail to capture the multi-faceted nature and *rationale* behind certain stances. By exploiting its constructed aspect hierarchy, CLAIMSPECT is able to infer viable *supportive*, *neutral*, and *opposing perspectives towards an aspect* and its associated papers.

## 3 Methodology

Illustrated in Figure 2, CLAIMSPECT consists of the following steps: (1) *aspect-discriminative retrieval*, (2) *iterative sub-aspect discovery*, and (3) *classification-based perspective discovery*.

### 3.1 Preliminaries

#### 3.1.1 Task Definition

We assume that as input, the user provides a claim  $t_0$  (e.g., “Vaccine A is better than Vaccine B”) and a corpus  $D$ . In order to better reflect real-world settings, we *do not* assume that each document  $d \in D$  is relevant to  $t_0$ .

**Definition 1 (CLAIM)** *A statement or assertion that expresses a position, which may require validation or scrutiny. It often encapsulates multiple dimensions that contribute to its overall truthfulness or validity.*

**Definition 2 (ASPECT)** *A specific component or dimension of a claim that can be independently analyzed or evaluated.*

ClaimSpect aims to output a hierarchy of aspects  $T$ , where each aspect node (e.g., “safety”) within the hierarchy can be considered as a descendant subclaim  $t_i$  of the root user-specified claim,  $t_0$  (e.g., “A is a safer vaccine than B”). In other words, each aspect node  $t_i$  should reflect a relevant aspect that is important to consider when evaluating the root claim  $t_0$ .

#### 3.1.2 Document Preprocessing

For each  $d \in D$ , we assume we have its full textual content (e.g., a full scientific paper). In order to have smaller, context-preserving units of text for our framework to retrieve, we segment each  $d$  into chunks using the widely-recognized text segmentation method, C99 (Choi, 2000). This method labels sentences with matching tags if they pertain to the same topical group, which assists with retaining consecutive discussion of an aspect to a single segment.

#### 3.1.3 Initial Coarse-Grained Aspect Discovery

Given our weak supervision setting, where only the root claim  $t_0$  is provided, we first generate reliable, coarse-grained aspects to guide the retrieval-augmented hierarchy construction. These aspects are typically common-sense and do not require domain expertise to identify. Preliminary experiments confirm that LLMs can reliably identify them based on their expansive background knowledge alone. Thus, we prompt an LLM to generate coarse-grained aspects  $t_i^0 \in T^0$  (e.g., efficacy, safety, and distribution in Figure 1) that will serve as the children of  $t_0 \in T$ . For each aspect  $t_i^0$ , the model outputs its label, significance to  $t_0$ , and a list of  $n = 10$  relevant keywords. This initial subtree forms the foundation of our framework. The full prompt is in Appendix B.1.

### 3.2 Aspect-Discriminative Retrieval

In order to construct a rich, coarse-to-fine aspect hierarchy that is *aligned with* the corpus, we must identify similarly rich reference material *from our* corpus. In general, noisy retrieval often hinders reasoning performance (Shen et al., 2024a), which may negatively impact discovering subaspects of a given node. Thus, in order to discover each subaspect  $t_j^i$  of an aspect node  $t_i$ , we must determine which segments  $S_i$  from our corpus

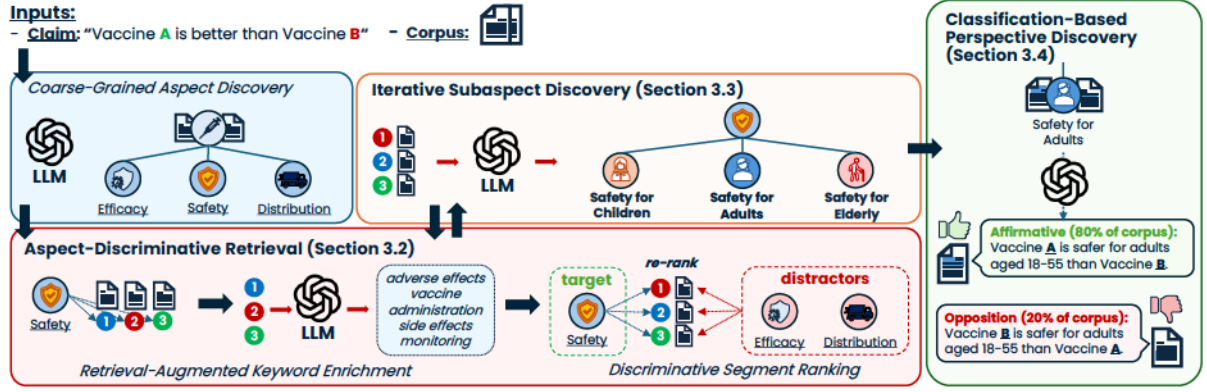


Figure 2: CLAIMSPECT deconstructs a nuanced claim into a hierarchy of aspects typically considered for validating the claim. We automatically discover the set of perspectives towards each aspect from the corpus.

$D$  discuss  $t_i$ . However, not all segments are equally informative for discovering subaspects.

Specifically, a high-quality, *discriminative* segment  $s_i$  for node  $t_i$  contains the following features: (1)  $s_i$  discusses  $t_i$  in *depth* and (2)  $s_i$  *does not* discuss  $t_i$ ’s siblings in *breadth* or *depth*. For instance, in Figure 1, a segment regarding the side effects observed within a clinical trial of Vaccine A and B on both children and adults *discusses “safety” in more depth* than if it only mentioned children. Furthermore, for discovering subaspects of “safety for children”, a segment which independently discusses the safety for *both* children and adults would introduce additional noise into the subaspect generation process. Overall, it is important to rank these segments such that we select a set which *minimizes the noise* we introduce into the retrieval-augmented discovery of subaspects, while *maximizing the number of subaspects* which we can discover. We formalize our discriminative ranking mechanism in the sections below:

### 3.2.1 Retrieval-Augmented Keyword Enrichment

In order to determine whether a segment discusses an aspect  $t_i$  in *depth*, we must first further enrich our understanding of  $t_i$ . We propose performing a retrieval-augmented keyword-based enrichment of  $t_i$ , where each keyword is likely to occur within segments relevant to  $t_i$  and, thus, reflects either explicitly or implicitly the subaspects of  $t_i$ . For example, for the “*efficacy*” aspect, the corresponding keywords are: *neutralization, immune stimulation, post-dose antibody response, and waning immunity*. First, we use a retrieval embedding model to select the top- $n$  segments (based on cosine-similarity) from the entire corpus that are relevant to a  $t_i$ -specific query (its root, name, description, and keywords from Section 3.1.3):

**Claim:**  $[t_0]$ ; **Aspect:**  $[t_i]$ : [generated description of  $t_i$ ]; **Aspect Keywords:** [generated keywords of  $t_i$ ].

We provide these initial top  $n$  segments in addition to the root claim  $t_0$ , the aspect label  $t_i$ , and its description to the LLM in-context to identify  $2k$  keywords. Given the same information and these keywords, we then

merge similar or duplicate terms, while filtering irrelevant terms—explicitly prompting the model to provide solely  $k$  keywords. This set of terms  $w \in W_i$ ;  $|W_i| = k$ , grounds our discriminative segment ranking for node  $t_i$ . We provide these two prompts in Appendix B.2.

### 3.2.2 Discriminative Segment Ranking

In order to determine the most discriminative segments  $S_i$  for aspect node  $t_i$ , we first collect an initial large pool of segments using the same retrieval embedding-based method as Section 3.2.1. Our subsequent goal is to rank a segment  $s \in S_i$  based on its discriminativeness:

- **Target Score:** Reward  $s$  based on its likelihood to contain all relevant subaspects  $t_j^i$  of  $t_i$ .
- **Distractor Score:** Penalize  $s$  based on the degree and depth of other sibling aspects that it discusses.

We assume that  $t_i$ ’s keywords  $W_i$  implicitly and/or explicitly reflect many of its subaspects. Thus, we use them to approximate the depth of an aspect-specific discussion. We convert each keyword  $w_i$  into a descriptive query: “[ $w_i$ ] with respect to [all ancestor nodes of  $w_i$ ]”. By integrating the ancestors into the query, we influence the retention of  $t_i$ ’s hierarchical context; for example, we specifically *reward* a segment if it discusses “the safety of Vaccine A and B”, as opposed to merely “safety”. We embed each keyword query  $emb(w \in W_i)$  using the retrieval embedding model, in addition to embedding each segment  $emb(s) \in S_i$ .

More formally, we are given an aspect node  $t_i^h$ , which is a child of parent node  $t_h$  and sibling node of  $t_j \in T_{\neq i}^h$ . We are also provided with a segment embedding  $emb(s) \in S_i$ , all keyword query embeddings of  $t_i^h$ ,  $emb(w) \in W_i$ , and all sibling keyword query embeddings,  $emb(w) \in W_{\neq i}^h$ . We compute the discriminative rank based on the following:

**Definition 3 (TARGET SCORE)** A segment  $s_i$  is rewarded based on a weighted average ( $H$ ) of its degree of similarity to all keywords  $w \in W_i$ , implying a deeper discussion of node  $t_i$  and its subaspects.



$$\mathbf{p}(s_i, W_i) = H\left(\left[\text{sim}(\text{emb}(s_i), \text{emb}(w)) \mid w \in W_i\right]\right),$$

$$\text{where } H(X) = \frac{\sum_{r=1}^{|X|} \frac{1}{r} x_r}{\sum_{r=1}^{|X|} \frac{1}{r}} \quad (1)$$

We compute a weighted average based on Zipf’s Law (Powers, 1998), where a word indexed at the  $r$ -th position will have a weight of  $1/r$ . This weighted average of the segment-keyword similarities is based on the assumption that the model will implicitly generate the keywords from most to least significant—in other words, we weight the first term  $w_1 \in W_i$  the highest, while weighing  $w_{|X|}$  the lowest. For example, if  $s_i$  had similarities of  $[0.9, 0, 0]$  to  $W_i = \{w_1, w_2, w_3\}$ , then  $\mathbf{p}(s_i, W_i) = 0.5363$ . On the other hand, if the similarities were  $[0.7, 0.8, 0.7]$ ,  $\mathbf{p}(s_i, W_i) = 0.7272$ . Overall, the target score will indicate a segment’s discussion *depth* of aspect node  $t_i$ —how many keywords it aligns with and to what degree.

**Definition 4 (DISTRATOR SCORE)** A segment  $s_i$  is penalized based on the breadth and depth of siblings discussed. The breadth is indicated by the mean target scoring between  $s_i$  and each  $W_j$  of  $t_j \in T_{\neq i}^h$ . The depth is indicated by the max target scoring between  $s_i$  and each  $W_j$  of  $t_j \in T_{\neq i}^h$ .

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

We utilize the target and distractor scores to compute our overall discriminativeness score, which weighs the proximity between a segment and its target aspect, relative to its overall and individual proximity to its distractor, sibling aspects.

**Definition 5 (DISCRIMINATIVENESS SCORE)** A segment  $s_i$  is rewarded based on a weighted average ( $H$ ) of its degree of similarity to all keywords  $w \in W_i$ , while being penalized based on the breadth and depth of siblings discussed.

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

In Equation 3,  $\mathbf{d}(s_i, W^h)$  grows proportional to the target score, while falling proportional to the distractor score. We include the  $\beta$  and  $\gamma$  scaling factors for each in case users would like to customize their degree of reward or penalty. Ultimately, we rank each segment  $s \in S_i$  based on its discriminativeness score, taking the top- $k$  segments which feature the richest discussion of target aspect  $t_i$  in order to discover its subaspects.

### 3.3 Iterative Subaspect Discovery

In order to expand our aspect hierarchy, we iteratively exploit our aspect-discriminative retrieval as knowledge which grounds the LLM’s subaspect discovery. Given the aspect node  $t_i$ , its description, its corresponding

discriminative segments  $S_i$ , and the root claim  $t_0$ , we prompt the model to determine a set of at minimum two and at maximum  $k$  subaspects for aspect  $t_0$ . We provide this prompt in Appendix B.3.

**Definition 6 (SUBASPECT)** A more granular component of a parent aspect  $t_i$  that further refines  $t_i$ ’s evaluation and would be considered when specifically addressing the root claim  $t_0$ .

Each subaspect is represented in the same manner specified in Section 3.1.3: its label, description, and keywords. We continue constructing our aspect hierarchy in a top-down fashion, as detailed in Algorithm 1.

---

#### Algorithm 1 Iterative Subaspect Discovery

---

**Require:** Root Claim  $t_0$ , Corpus  $D$ , max\_depth= $l$   
1:  $T = \text{initialize\_tree}(t_0)$  { $T.\text{depth} = 0$ }  
2:  $t_i^0 \in T^0 \leftarrow \text{coarse\_grained\_aspects}(t_0)$  {Section 3.1.3}  
3:  $q = \text{queue}(T^0)$   
4: **while**  $\text{len}(q) > 0$  and  $T.\text{depth} \leq l$  **do**  
5:    $t_i \leftarrow \text{pop}(q)$   
6:    $\text{enrich\_node}(t_0, t_i, D)$  {Section 3.2.1}  
7:    $S_i \leftarrow \text{rank\_segments}(t_0, t_i, D)$  {Section 3.2.2}  
8:    $t_j^i \in T^i \leftarrow \text{subaspect\_discovery}(t_0, t_i, S_i)$   
9:    $q.\text{append}(T^i)$   
10: **end while**  
11: **return**  $T$

---

Ultimately, the output of Algorithm 1 is our final aspect hierarchy, serving as the basis for our consensus determination and perspective discovery process.

### 3.4 Classification-Based Perspective Discovery

With the aspect hierarchy constructed, we must identify the *complete* set of corpus segments that (1) pertain to the root claim  $t_0$  and (2) align with an aspect node in hierarchy  $T$ . Pinpointing papers discussing aspect node  $t_i$  allows us to infer their *perspective* on  $t_i$  and assess the *presence* and *extent of consensus*. However, as noted in Section 3.1.1, we cannot assume all corpus segments are relevant to the root claim—an assumption made in LLM-based taxonomy-guided hierarchical classification works (Zhang et al., 2024a). Thus, we must first filter out claim-irrelevant segments.

**Filtering.** A naive approach determines segment relevance per node via in-context prompting, but this scales poorly. Instead, we frame relevance filtering as a *binary search* problem, identifying the relevance-irrelevance boundary. Specifically, we embed the claim label  $t_0$  ( $\text{emb}(t_0)$ ) and each child aspect  $t_i^0 \in T^0$  ( $\text{emb}([[\text{aspect\_label}]]$  with respect to  $[t_0]$ )), computing the claim representation as:

$$\mathbf{c}_0 = \frac{1}{2} \left( \text{emb}(t_0) + \frac{\sum_{i=1}^{|T^0|} \text{emb}(t_i^0)}{|T^0|} \right) \quad (4)$$

We rank the encoded segments by cosine-similarity to  $\mathbf{c}_0$  and use binary search to find the index  $r$  where fewer than  $\delta\%$  of segments in a  $\pm n$  window are relevant. This rank  $r$  serves as our threshold, filtering out lower-ranked segments and retaining only those relevant to  $t_0$  ( $S_0'$ ).

This optimization significantly reduces the quantity of relevance judgments necessary; the relevancy prompt is in Appendix B.4.

**Hierarchical Text Classification.** With  $S'_0$  and  $T$ , we apply *taxonomy-guided hierarchical classification* to determine  $S'_i$  for each aspect node  $t_i \in T$ . Since our focus is retrieval-guided aspect hierarchy construction rather than classification, we adopt a recent LLM-based hierarchical classification model (Zhang et al., 2024a), which enriches taxonomy nodes (e.g., adding keywords) to support its top-down classification of  $S'_i$  to  $t_i$ .

**Perspective & Consensus Discovery.** The final step of our pipeline is to determine the primary perspectives  $P_i = \{a_i, o_i\}$  towards each aspect node  $t_i$ , where  $a_i$  is the overarching supportive perspective and  $o_i$  is the opposing perspective. We also seek to identify the papers which hold each of these perspectives ( $D_i = D_i^{\text{supp}} \cup D_i^{\text{opp}} \cup D_i^{\text{neutral}}$ ), accounting for papers which do not hold any clear perspective towards  $t_i$ .

**Definition 7 (PERSPECTIVE)** *A descriptive viewpoint expressed toward a specific aspect  $t_i$  of claim  $t_0$  in the form of an implicit or explicit stance towards  $t_i$  (e.g., support, neutral, or oppose) and optionally, a rationale.*


We do not assume that  $D_i^{\text{supp}}$ ,  $D_i^{\text{opp}}$ , and  $D_i^{\text{neutral}}$  are non-overlapping, as they may have multiple segments indicating different perspectives. For example, a segment  $s'_i \in S'_i$  mapped to “Safety for Elders” may discuss a clinical trial showing increased anaphylactic shock in older patients when taking Vaccine A. However, another segment from the same paper may also note severe hives from Vaccine B. Thus, we allow for such flexibility.

Recent studies have shown LLMs demonstrate powerful abilities in stance detection (Zhang et al., 2024b; Lan et al., 2024). Consequently, in order to discover these perspectives, we prompt the model to first determine the stance of each segment  $s'_i \in S'_i$ :

- **Supports Claim:**  $s'_i$  either implicitly or explicitly indicates that the *claim is true* with respect to  $t_i$ .
- **Neutral to Claim:**  $s'_i$  is relevant to the claim and aspect, but *does not indicate* whether the claim is true with respect to  $t_i$ .
- **Opposes Claim:**  $s'_i$  either implicitly or explicitly indicates that the *claim is false* with respect to  $t_i$ .

This forms the segment sets:  $S_i^{\text{supp}}$ ,  $S_i^{\text{neutral}}$ , and  $S_i^{\text{opp}}$ . We ask the model to summarize the perspective (stance and rationale) of each segment set:  $P_i$ . Both prompts are provided in Appendix B.5. Since we retain the original paper source of each segment, we are able to construct  $D_i$  for each node  $t_i$ . This indicates consensus; for instance, how many papers in  $D$  held perspective  $p_i^{\text{supp}}$  towards aspect  $t_i$ . **As our final output, we have the aspect hierarchy  $T$ , the set of perspectives  $P_i$ , and their corresponding papers  $D_i$ .**

## 4 Experimental Design

We explore CLAIMSPECT’s performance on an open-source model, Llama-3.1-8B-Instruct (.

We sample from the top 1% of the tokens and set the temperature based on the nature of the given task (same setting across all samples); we include these settings in Appendix C. We set the maximum depth of the aspect hierarchy to  $l = 3$ .



### 4.1 Dataset

In order to evaluate CLAIMSPECT’s abilities to deconstruct *nuanced claims* into a hierarchy of aspects and identify their corresponding perspectives, we construct **two novel, large-scale datasets** specific to our task, applied to both political (**World Relations**) and scientific (**Biomedical**) domains. To construct this dataset, we first manually collect  $\sim 50$  reference materials discussing (1) security-related international conflicts, and (2) biomedical safety-related studies. Then, we used GPT-4o (OpenAI et al., 2024) to generate nuanced claims based on these materials. Subsequently, we used the Semantic Scholar API (Allen Institute for AI, 2025) to collect meta information relevant literature based on these claims. Then, based on this meta information, we filtered the collected literature and retrieved the full texts. This way, for each claim, we obtained a corresponding literature repository. We show the statistics of each of these datasets in Table 1. More details about the dataset construction can be found in Appendix D, including a human study for validating the quality of the generated claims and their associated papers in Appendix D.5.

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: # of claims, papers, and segments per dataset.







### 4.2 Baselines

Our primary motivation for CLAIMSPECT is to demonstrate its capabilities of deconstructing a nuanced claim into an aspect hierarchy and identifying corresponding perspectives. However, no existing methods tackle this novel task. Consequently, we choose to implement and compare our method with both *RAG-driven* and *LLM-only* approaches, detailed below. We run each baseline using both Llama () and GPT-4o-mini ()

1. **RAG-Based:** Given a claim and definition of an aspect hierarchy, we use the claim as a query to retrieve relevant documents. We then provide the documents in-context when prompt the LLM to generate the aspect hierarchy.
2. **Iterative RAG-Based:** Given the claim, the definition of an aspect hierarchy, and the name/description of the current node  $t_i$ , we construct a detailed query to retrieve node-specific



Table 2: Comparison between ClaimSpect and all baselines. Sibling granularity (*Sib*) is normalized; all others are scaled by 100. Since Iterative Zero-Shot is not grounded with a corpus, there are no associated segments to each node. Thus, we omit its segment relevance scores (*Seg*). We **bold** the top score and underline the second-highest.

Method	World Relations					Biomedical				
	<i>Rel</i>	<i>Path</i>	<i>Sib</i>	<i>Unique</i>	<i>Seg</i>	<i>Rel</i>	<i>Path</i>	<i>Sib</i>	<i>Unique</i>	<i>Seg</i>
Iterative Zero-Shot 	97.85	41.94	58.01	72.96	—	<b>98.33</b>	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 	<u>98.60</u>	42.88	64.04	76.01	—	97.89	41.56	62.09	77.55	—
Iterative RAG 	97.40	52.30	66.45	76.59	<u>46.93</u>	94.37	50.07	64.21	77.05	<u>31.82</u>
CLAIMSPECT 	95.30	<u>78.24</u>	<b>85.26</b>	<b>87.62</b>	43.23	<u>97.95</u>	<u>75.10</u>	<b>74.80</b>	<u>86.26</u>	27.39
CLAIMSPECT - No Disc 	<b>99.00</b>	<b>79.75</b>	<u>82.64</u>	<u>85.43</u>	<b>49.47</b>	96.07	<b>76.26</b>	<u>74.39</u>	<b>87.69</b>	<b>39.03</b>

relevant documents. We then provide these documents in-context to prompt the LLM for generating the children subaspects  $t_j^i$  of aspect  $t_i$ .

We also conduct an ablation study, *No Discriminative* (No Disc), where we remove discriminative ranking and instead replace it with a semantic similarity-based ranking. For this, we compute the semantic similarity between each segment and our  $t_i$ -specific query from Section 3.2.1.

### 4.3 Evaluation Metrics

We design a thorough automatic evaluation suite using GPT-4o-mini to determine the quality of our generated taxonomies, using both node-level and taxonomy-level metrics. For each judgment, we ask the LLM to provide additional rationalization:

- **(Node-Wise) Node Relevance:** For each aspect node  $t_i$  and its respective path within the hierarchy, what is its relevance to the claim  $t_0$ ? Scored 0/1.
- **(Node-Wise) Path Granularity:** Does the path to node  $t_i$  preserve the hierarchical relationships between its entities (is each child  $t_j^i$  more specific than the parent  $t_i$ )? Scored 0/1.
- **(Level-Wise) Sibling Granularity:** For each set of siblings  $T^i$  within the hierarchy, does the overall set reflect the same level of specificity relative to their parent aspect  $t_i$ ? Scored from 1 to 4 (all different  $\rightarrow$  some  $\rightarrow$  most  $\rightarrow$  all same).
- **(Node-Wise) Uniqueness:** Does the aspect node  $t_i$  have other overlapping nodes within the hierarchy  $T$ ? Scored 0/1.
- **(Node-Wise) Segment Quality:** How many segments  $s \in S_i'$  are relevant to the claim  $t_0$  and aspect  $t_i$ ? We compute the average proportion of relevant segments per node.

In addition to automatically evaluating our aspect hierarchy, we also conduct a supplementary human evaluation on 50 perspectives and their sampled segments, which CLAIMSPECT identifies from the corpus (Section 5.2).

## 5 Experimental Results

### 5.1 Overall Performance & Analysis

Tables 2-3 demonstrate several key advantages of CLAIMSPECT over the baselines across various node and level-wise metrics for both the *World Relations* and *Biomedical* datasets. CLAIMSPECT is able to strongly enforce the hierarchical structure of the generated aspect hierarchy while preserving relevance to the corpus. Below, we present our core findings and insights. We also provide a breakdown of ClaimSpect’s computational efficiency in Appendix E. Finally, we additionally conducted a human-automatic evaluation agreement study in Appendix A.


**CLAIMSPECT excels in granular aspect discovery.** As shown in Table 2, CLAIMSPECT significantly outperforms the baselines in metrics associated with node-level structure, particularly outperforming Iterative RAG  by 72.6% and 63.51% in preserving hierarchical relationships (path granularity) and by 44.48% and 26.61% in maintaining uniform sibling-level specificity (sibling granularity) for both datasets respectively. This demonstrates the method’s ability to *retrieve and organize aspects at targeted levels of granularity*. These gains are similarly observed with the GPT-based baselines, despite relying on a closed-source model. We attribute this gain to ClaimSpect’s iterative subaspect discovery (Section 3.3) being integrated with its *aspect-discriminative retrieval mechanism* (Section 3.2), where the pool of segments grounding the subaspect discovery is iteratively updated based on the given aspect node. We can see that the *No Disc* ablation does experience some loss in granularity quality. It is important to note that *No Disc* does experience competitive and, at times, better performance; this is likely due to it considering more segments, which may or may not discuss multiple aspects. In contrast, the baseline methods retrieve broader, less focused segments, reducing their ability to discover fine-grained sub-aspects. Overall, this demonstrates that ClaimSpect is able to *deconstruct a claim into a well-structured hierarchy of aspects*.

Table 3: Pairwise comparisons between all methods for each dataset. Each value is the *percentage* of samples within each dataset where the method is considered better. *E-Tie* denotes Explicit Tie; *I-Tie* denotes Implicit Tie.

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 $\infty$ vs RAG $\infty$	0.00	33.06	0.00	<b>66.94</b>	2.22	22.22	0.00	<b>75.55</b>
Zero-Shot $\infty$ vs CLAIMSPECT $\infty$	0.00	<b>97.58</b>	0.00	2.42	0.00	<b>95.55</b>	2.22	2.22
RAG $\infty$ vs CLAIMSPECT $\infty$	0.81	<b>90.32</b>	0.00	8.87	0.00	<b>95.55</b>	0.00	4.44
No Disc $\infty$ vs CLAIMSPECT $\infty$	21.43	30.00	0.00	<b>48.57</b>	24.00	28.00	0.00	<b>48.00</b>
Zero-Shot $\odot$ vs RAG $\infty$	0.00	36.00	0.00	<b>64.00</b>	0.71	47.14	0.00	<b>52.14</b>
Zero-Shot $\odot$ vs CLAIMSPECT $\infty$	0.00	<b>98.00</b>	0.00	2.00	0.00	<b>96.43</b>	0.00	3.57
RAG $\odot$ vs CLAIMSPECT $\infty$	0.00	<b>90.00</b>	0.00	10.00	7.14	<b>72.14</b>	0.71	20.00

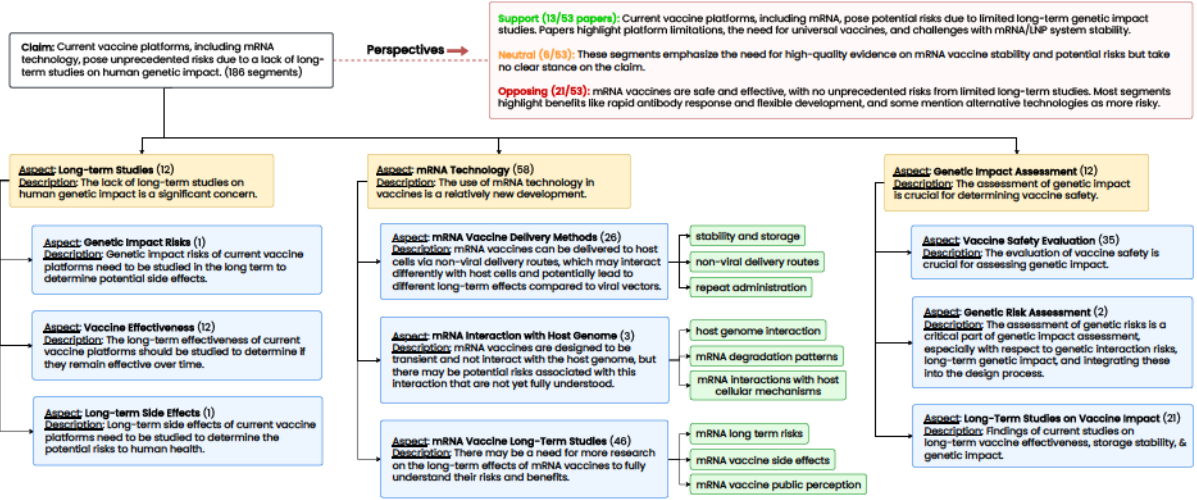


Figure 3: A constructed Biomedical aspect hierarchy. All nodes and their # of segments from levels 1-2 are included; a subset of the third level is highlighted. The # of papers mapped to each perspective is provided in parentheses.

**CLAIMSPECT constructs a rich aspect hierarchy while preserving relevance.** In Table 2, we observe that ClaimSpect’s constructed hierarchy features nodes that are 14.40% and 11.23% more unique than the top baseline on each dataset, respectively. This indicates that ClaimSpect’s hierarchies are *richer* in aspect quality, experiencing less overlap between aspects across the tree and supported by an increase in segment quality. Despite this significant boost in uniqueness, ClaimSpect only experiences a 3.35% and 0.386% drop from the top baseline in aspect node relevance for the World Relations and Biomedical datasets, respectively. This highlights the strength of ClaimSpect’s retrieval-augmented keyword enrichment and aspect-discriminative retrieval (Sections 3.2.1 and 3.2), which prioritize segments that *thoroughly discuss a single aspect* rather than *shallow* descriptions of *multiple* aspects. This allows us to *discover a richer set of unique and relevant subaspects at each level, throughout the hierarchy*.

**CLAIMSPECT is overwhelmingly preferred over baselines.** Table 3 presents pairwise comparisons between CLAIMSPECT and the baseline methods. These

comparisons are judged by an LLM that is shown the aspect hierarchy outputs of methods *A* and *B* in both possible orders: *A* vs. *B* and *B* vs. *A*. The LLM may (1) prefer either method *A* or *B*, (2) declare an explicit tie (*E-Tie*), or (3) indicate an implicit tie (*I-Tie*), which occurs when the preferred method changes depending on the order of presentation (e.g., *A* wins in *A* vs. *B*, but *B* wins in *B* vs. *A*) (Shi et al., 2024).

Across both datasets, CLAIMSPECT exhibits a clear advantage, being preferred 92.95% of the time, with a 6.69% average inconsistency rate across all settings and datasets. Specifically, when compared with Zero-Shot  $\infty$ , CLAIMSPECT is judged superior in 97.58% and 95.55% of cases for World Relations and Biomedical datasets, respectively. Even against RAG  $\odot$ , CLAIMSPECT outperforms in 90.00% and 72.14% of samples. This is a stark contrast from the lack of strong preference between the baselines themselves, indicated by the 64.66% average implicit tie rate—implying that there is no obvious qualitative preference between the two. Finally, we show that ClaimSpect and No Disc are similarly preferred, with ClaimSpect preferred slightly more



often. Overall, these results validate that CLAIMSPECT *constructs significantly more meaningful aspect hierarchies relevant to the claim*.

## 5.2 Perspective Discovery Analysis

**CLAIMSPECT identifies nuanced, corpus-specific perspectives.** We showcase a qualitative analysis of a nuanced claim’s aspect hierarchy, highlighting certain subtrees and the root node’s extracted perspectives, in Fig. 3. We observe each coarse-grained aspect (yellow nodes) well represents the various angles of the root claim that would be considered in validating it: what long-term vaccine studies currently exist, what is the current mRNA technology, and how is genetic impact currently assessed? We see that the path-specific dependencies are reflected within the descriptions of each aspect (e.g., “*mRNA Interaction with Host Genome*” involves both mRNA technology *and* potential genetic impact risks). Furthermore, these hierarchical relationships and claim relevance are preserved even in the final layer of the hierarchy (e.g., “*mRNA Interaction with Host Genome*” → “*mRNA degradation patterns*”). Finally, we see that the perspectives mapped to the root node are informative, providing justification behind each stance. Note that ClaimSpect maps segments to each perspective, allowing us to identify the original paper sources and ultimately provide a *corpus-specific estimate of the consensus*. Overall, this deconstructed view of the claim provides a means to identify *which and to what degree certain aspects have been explored* (e.g., *mRNA Technology* has been more explored within the corpus compared to *Genetic Impact Assessment*).

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

Table 4: Human validation on the percentage of perspectives discovered by CLAIMSPECT which are grounded in at least one of  $k$  associated segments.  $k$  = # of segments considered. We provide the number of samples for each setting in parenthesis.

**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*.

## 6 Conclusion

Our work introduces CLAIMSPECT, a novel framework for deconstructing nuanced claims into a hierarchy of corpus-specific aspects and perspectives. By integrating iterative, aspect-discriminative retrieval with hierarchical sub-aspect discovery and perspective clustering, CLAIMSPECT provides a structured, comprehensive view of complex claims. Our experiments on two novel, large-scale datasets demonstrate that CLAIMSPECT constructs rich, corpus-aligned aspect hierarchies that are enriched with diverse and informative perspectives. This highlights its effectiveness as a scalable and adaptable method for nuanced claim analysis across domains.

## 7 Limitations & Future Work

The primary contribution of CLAIMSPECT is our retrieval-augmented framework for constructing an aspect hierarchy relevant for validating a nuanced claim. In order to demonstrate the hierarchy’s potential, we apply it to the task of perspective discovery, involving (1) identifying which segments from the corpus are *relevant to a given aspect node*, (2) determining the *stance* (or lack thereof) of the segment towards the claim and aspect, and (3) discovering the potential *perspective* of each of the stance-based segment clusters. Consequently, this step *relies heavily upon an existing hierarchical classification model* (Zhang et al., 2024a), as we do not claim novelty with respect to classification. Similarly, our classification-based perspective discovery (Section 3.4) is reliant on the LLM’s fine-grained stance detection abilities— although prior work (Zhang et al., 2024b; Lan et al., 2024) has shown precedence for its capabilities. Thus, the performance of the hierarchical classification and stance detection serves as a bottleneck to our method’s performance. For example, if the LLM-based stance detection has a *high recall but low precision* for detecting segments which *support* the aspect of claim, then the method may *overestimate* the consensus behind a certain perspective within the corpus. Likewise, if the detection has a high precision but lower recall, it may *underestimate the consensus*. Nonetheless, our work aims to, overall, motivate the need to structure the aspects of certain nuanced claims *before diving straight into their validation*.

Hierarchically analyzing nuanced claims opens up doors to many new avenues of research. First, CLAIMSPECT can be integrated with more systematic and/or tool-integrated fact validation systems, in an effort to build a more robust fact-checking system. Furthermore, CLAIMSPECT can be applied to more targeted retrieval or question answering tasks where a question, similar to a nuanced claim, cannot easily be answered and may

benefit from a more structured output (similar to an aspect hierarchy).

## 8 Acknowledgements

This work was supported by the National Science Foundation Graduate Research Fellowship. The work was also supported in part by the BRIES Program No. HR0011-24-3-0325. This research used the DeltaAI advanced computing and data resource, which is supported by the National Science Foundation (award OAC 2320345) and the State of Illinois. DeltaAI is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. We thank Peter Bautista, Spencer Lynch, and Svitlana Volkova from Aptima, Inc. for their discussions on our work. We also thank Mihir Kavishwar for early ideation discussions.

## References

- Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.
- 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.
- Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 588–596. IEEE.
- Zhen-Song Chen, Xuan Zhang, Rosa M Rodríguez, Witold Pedrycz, Luis Martínez, and Mirosław 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*.
- 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.
- David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.
- 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. [pdfotext](#).
- Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.
- 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.
- Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Minghui Zhang, and Yan Liu. 2019. Combating



- fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.
- Xiaoyu Shen, Rexhina Biloshmi, 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*.
- Yushi Sun, Hao Xin, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2024. Are large language models a good replacement of taxonomies? *arXiv preprint arXiv:2406.11131*.
- 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.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- 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, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. 2024a. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3093–3102.
- Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2024b. Code-taxo: 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.

## A Human Annotator Background and Human-Automatic Alignment for Evaluation

We perform a human evaluation study to show the alignment between humans and GPT-4o-mini for evaluation. We use two human evaluators to evaluate randomly sampled cases for each of our proposed metrics: node relevance, path granularity, sibling granularity, uniqueness and segment quality. Our human evaluators are two volunteer graduate researchers (one PhD student and Masters student), with one having a background in Biology + NLP (critical for our biomedical task evaluation). We used the following instructions to help guide them in the evaluation task, with an initial “training” period where the evaluators could familiarize themselves with the task (e.g., aspect taxonomies and their expected hierarchical relationships) and discuss examples with one another, but the full evaluation period was conducted independently:

1. **General Instruction:** *Claims made by individuals or entities are often nuanced and cannot always be strictly categorized as entirely 'true' or 'false', particularly in scientific and political contexts. Instead, a claim can be broken down into its core aspects and sub-aspects, which are easier to evaluate individually.*
2. **Node Relevance:** *Given the claim: [claim], decide whether this path from the aspect tree is relevant to the analysis of the claim: [path] → relevant or irrelevant?*
3. **Path Granularity:** *Given the claim: [claim], decide whether this path from the aspect tree has good granularity: [path]. Check whether the child node is a more specific subaspect of the parent node. → granular or non-granular?*
4. **Sibling Granularity:** *Given the claim: [claim]', decide whether these siblings from parent node [parent] have good granularity. → <proportion of granular siblings>*
5. **Uniqueness:** *Normally, we want the aspects and sub-aspects to be unique in the taxonomy. Given the claim: [claim], count how many nodes in this taxonomy are largely overlapping or almost equivalent. → <total count of overlapping nodes>*
6. **Segment Quality:** *Given the claim: [claim], evaluate the quality of these segments for aspect [aspect node label]. [list of mapped segments] → <total count of relevant segments to aspect node>*

Since ClaimSpect has consistently high scores on “Node Relevance” and “Uniqueness” (not very many nodes are irrelevant or overlapping across the entire taxonomy), the scores from both the LLM and evaluators have very low variance. This greatly limits both Cohen’s

and Intraclass Correlation Coefficient (high instability). Nonetheless, the evaluators have a 100% and 96.97% agreement rate respectively. We choose Cohen’s  $\kappa$  or ICC based on the ordinal/continuous versus categorical nature of the metric.

1. The weighted  $\kappa$  for Path Granularity is 0.62 → **substantial agreement**
2. The  $ICC1k$  for Sibling Granularity is 0.7806 → **good reliability**
3. The  $ICC2k$  for Segment Quality is 0.7578 → **good reliability**

We also show the agreement rate below on 100 different samples across all metrics in Table 5.

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

Table 5: Human-Automatic Agreement rates across different metrics.

We can see that the LLM evaluation and human evaluators have a high degree of alignment. We note that segment quality features the lowest alignment (albeit still a relatively high rate), likely due to the more fine-grained text understanding abilities required for verifying segment alignment to the parent node (where oftentimes, segments can be quite semantically dissimilar or only discuss a “sub-aspect” of the node). Nonetheless, through these results, we can see that **our auto-evaluation is reliable**.

## B Prompt Template

In this section, we present the prompts used in different modules of CLAIMSPECT.

### B.1 Coarse-Grained Aspect Discovery

This is the prompt used to generate coarse-grained aspects for the root claim, including their labels, description, and relevant keywords to structure the initial retrieval-augmented hierarchy.

#### Prompt

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

### B.2 Retrieval-Augmented Keyword Enrichment

Following are the prompts used for retrieval-augmented keyword enrichment, instructing the LLM to refine and filter aspect-specific keywords for improved segment ranking.



#### **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 separate them with comma. Your output should be in JSON format.

#### **Prompt (Filtering)**

Our claim is '{claim}'. With respect 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.

### **B.3 Iterative Subaspect Discovery**

Following is the prompt used to iteratively guide the LLM in discovering and expanding subaspects for each aspect node based on discriminative retrieval and root claim context.

#### **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.

### **B.4 Relevance Filtering**

Following is the prompt used for relevance filtering, leveraging binary search on cosine-similarity rankings to efficiently identify and retain only the most relevant segments for each aspect.

#### **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.

### **B.5 Perspective Discovery**

Following are prompts used to for determining segment stances (support, neutral, or oppose) and summarizing perspectives, including rationales, for each aspect.

#### **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.

## **C Generation Settings**

This section details the temperature values used in various stages of our process and their respective roles.

### **C.1 Overview of Temperature Settings**

- **Coarse-Grained Aspect Discovery (0.3):** Used to generate high-level aspects related to the claim. A lower temperature ensures structured and deterministic output.
- **Subaspect Discovery (0.7):** Used for identifying subaspects from ranked segments. A higher temperature allows for more diversity while maintaining coherence.

- **OpenAI Chat Models** (GPT-4o (OpenAI et al., 2024), GPT-4o-mini (OpenAI et al., 2024)) (0.3): Applied in various stages where GPT-4o models are used (e.g., aspect generation, classification), ensuring consistent responses.
- **Subaspect Discovery (Aspect Ranking and Retrieval)** (0.7): Used when extracting subaspects from ranked segments to balance creativity with relevance.

## C.2 General Trends

- **Lower temperature** (0.3) is used for structured and deterministic tasks such as *aspect generation and classification*.
- **Higher temperature** (0.7) is applied to *subaspect discovery*, where diversity and exploration are beneficial.

## D Dataset Construction

To evaluate the effectiveness of CLAIMSPECT, our nuanced claims analysis, we constructed two datasets covering two key domains: **political (World Relations)** and **scientific (Biomedical)**. The dataset construction process consists of the following steps:

### D.1 Manual Seed Collection

We begin by manually collecting a set of seed claims from reliable sources such as Google Scholar (Google Scholar, 2025) and PubMed (PubMed, 2025). Specifically, we collect material from 7 papers in the World Relations domain and 50 papers in the Biomedical domain. These initial materials serve as a context or specific topics for generating nuanced claims.

### D.2 Nuanced Claims Generation

Using the literature collected in the previous step and definition of nuanced claims as context, we prompt GPT-4o (OpenAI et al., 2024) to generate nuanced claims related to the topics within these papers. To ensure diversity in claim perspectives, we employ two sets of prompts: one for generating claims that align with the perspectives in the literature and another for generating claims that diverge from them. The specific prompts used are detailed below.

#### 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 necessarily 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 necessarily 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.

We find that the generated nuanced claims are of high quality. They are content-rich, specific, and difficult to classify as simply true or false, aligning well with our task requirements. Below are some example claims from our datasets.

### 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.

### D.3 Meta Information Collection

To support the corpus-based analysis of each claim, we retrieve relevant literature using the Semantic Scholar API (Allen Institute for AI, 2025).

Since our claims are highly nuanced and involve multiple concepts, directly searching for claims themselves does not yield useful matches based on literature titles and abstracts. To address this, we first perform keyword extraction for each claim. We then use the extracted keywords to query the Semantic Scholar API and retrieve up to 1000 related literature entries for each claim.

### D.4 Filtering and Full-Text Collection

After obtaining the literature metadata, we first filter out entries with missing fields and retain the top 100 most relevant papers based on relevance. We then utilize the provided PDF URLs to download the full-text of the selected literature and convert them into plain text with pdftotext (Palmer, 2024). As a result, we obtain a comprehensive textual literature repository for each claim, ensuring a rich contextual foundation for further analysis.

This structured approach ensures a robust dataset suitable for nuanced claims analysis across the domains.

### D.5 Human Validation of Generated Claims and Assigned Papers

We conducted a human evaluation study for validating 40 total claims—20 on each dataset. We define the following binary criteria for claim validation:

1. **Nuanced:** Is the claim obviously true or false?
2. **Relevant:** Is the claim relevant to the topic (biomedical/world relations)?
3. **Corpus-Aligned@k:** At least  $k$  papers are relevant to the claim within the corpus (this is computed on  $k = 5$  and  $k = 10$ )

We show the validation results in Table 6, which demonstrates the nuanced nature of the generated claims, their relevancy, and the presence of  $k$  papers aligned to each claim.

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: Human validation of claim quality across datasets.

## E Computational Efficiency

We specify the components of CLAIMSPECT’s framework and their corresponding computational efficiency across the entire pipeline below. We consider the number of nodes within a full aspect hierarchy as  $n$  and the total number of segments within the corpus as  $S$ . We additionally provide rough time estimates based on an average sample.

- **Coarse-Grained Aspect Discovery (Section 3.1.3)**

- A single LLM call:  $O(1)$

- **Aspect-Discriminative Retrieval (Section 3.2)**

- *Retrieval-Augmented Keyword Enrichment (Section 3.2.1)*

- \* Embedding the segments using the retrieval model takes the most amount of time ( $O(S)$ ), but this can be computed offline given a knowledge base. Retrieval itself is quite efficient since it is embedding-based, and we use cosine-similarity to determine relevance (an efficient computation, especially in high-dimensional scenarios).

- \* Enrich each node:  $O(N) \rightarrow 10$  seconds per node

- *Discriminative Segment Ranking (Section 3.2.2)*

- \* We only compute the ranking on the top-100 segments, so this operation’s efficiency is constant:  $O(1)$

- \* The target score and distractor score computation scales according to the number of aspects throughout the tree (since their # of associated keywords is constant):  $O(N) \rightarrow 6$  seconds for each node

- **Iterative Subaspect Discovery (Section 3.3)**

- A single prompt per aspect node in hierarchy:  $O(N)$

- **Classification-Based Perspective Discovery (Section 3.4)**

- As mentioned in lines 442–446, “we reframe relevance filtering as a binary search problem” intentionally to optimize the efficiency of this

module. Thus, instead of sequentially determining claim relevancy for each segment, we sort the segments (which have more than 500 characters) and use binary search ( $O(\log S)$ ) to find the boundary of relevance-irrelevance. This leads to  $O(S \log S)$  efficiency due to the sorting function  $\rightarrow 5$  seconds during runtime, due to Python optimizations.

- \* Perspective Discovery involves prompting the LLM for each filtered segment’s stance:  $O(S) \rightarrow 3$  minutes

We specifically use vLLM to optimize our LLM batched generation. To construct a claim with 39 nodes and a max depth of 3, it takes approximately **20 minutes** to run on two NVIDIA RTX A6000s. We can see that in total, the core framework operations take **13 minutes and 42 seconds**, with the remaining time dedicated to embedding computations (which can be done offline).