How Grounded is Wikipedia?
A Study on Structured Evidential Support and Retrieval

Anonymous ACL submission

Abstract

Wikipedia is a critical resource for modern NLP, serving as a rich repository of up-to-date and citation-backed information on a wide variety of subjects. The reliability of Wikipedia—its groundedness in its cited sources—is vital to this purpose. This work analyzes both how grounded Wikipedia *is* and how readily fine-grained grounding evidence can be retrieved. To this end, we introduce PEOPLEPROFILES[1]—a large-scale, multi-level dataset of claim support annotations on biographical Wikipedia articles. We show that: (1) $\sim$22% of claims in Wikipedia *lead* sections are unsupported by the article body; (2) $\sim$30% of claims in the article *body* are unsupported by their publicly accessible sources; and (3) real-world Wikipedia citation practices often differ from documented standards. Finally, we show that complex evidence retrieval remains a challenge—even for recent reasoning rerankers.

1 Introduction

Long an essential ingredient for LLM pretraining, Wikipedia is now widely used during inference as a repository of high-quality, citation-backed information for RAG applications (Lewis et al., 2020; Chen et al., 2020; Fan et al., 2024, *i.a.*). In parallel, Wikipedia has played a major role in advancing automated *fact* or *claim verification* (Dmonte et al., 2024), enabling the creation of many notable benchmarks for these tasks, such as FEVER (Thorne et al., 2018a,b), WikiFactCheck-English (Sathe et al., 2020), VitaminC (Schuster et al., 2021), and WiCE (Kamoi et al., 2023). But whereas these works treat Wikipedia articles as sets of claims or passages to sample from for dataset curation, this work studies Wikipedia articles as *whole, structured documents*—relied upon as trustworthy sources for information-seeking tasks.

[1]Data is in the supplementary materials. Code available upon paper acceptance.

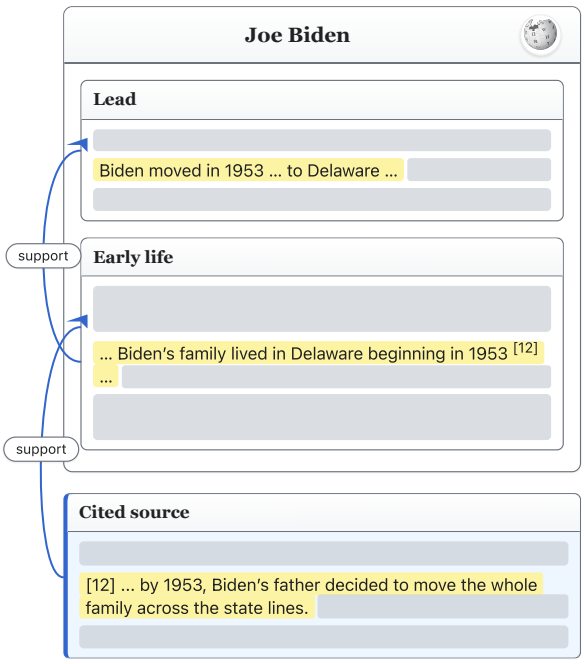

Figure 1: PEOPLEPROFILES features fine-grained, multi-level evidential relations and scalar support labels (not shown) on Wikipedia articles—from cited sources to claims in the article body (bottom arrow), and from the body to claims in the article lead (top arrow).

First, we ask how *grounded* claims in Wikipedia are. Acknowledging Wikipedia's distinction between an article's *lead* section (i.e. the intro) and its *body* (the remaining sections), we are the first to jointly explore both how claims in the lead are grounded in the body (*article-internal* support; Figure 1, top arrow) and how claims in the body are in turn grounded in publicly accessible cited sources (*article-external* support; Figure 1, bottom arrow). Second, we ask how well both standard first-stage retrievers and new reasoning-based reranking models can recover evidence about these claims—either from the body (for lead claims) or from cited source documents (for body claims). In answering these questions, we make the following contributions:

- We release PEOPLEPROFILES, a new dataset

of *structured* Wikipedia claim support judgments for *all* lead claims and *all* body claims with scrapable citations from 1.5K articles about people, covering nearly 50K lead claims and 100K body claims with fine-grained scalar support labels and associated evidence.

- We show that: (1) a surprising proportion of lead claims ($\sim$22%) are *unsupported by the body of the same article*; (2) an even higher proportion of body claims ($\sim$30%) are *unsupported by scrapable sources*; and, more generally, (3) actual Wikipedia citation practices often differ from documented standards.
- We show that evidence for these claims is often *complex*, involving multiple premises, and that retrieval of such evidence remains challenging even for new reasoning rerankers.

2 Background

Wikipedia-Based Claim Verification *Fact* or *claim verification* is an active area of research within NLP with many supporting datasets, including several that draw heavily on Wikipedia.[2] Of these, FEVER (**F**; Thorne et al., 2018a,b) is the most widely studied, featuring 185K claims written by human annotators, given randomly chosen Wikipedia sentences as prompts. For each claim, a different set of annotators then curated evidence passages—sets of sentences from one or more Wikipedia pages—as well as claim support labels (SUPPORTED, REFUTED, NOTENOUGHINFO).

WikiFactCheck-English (**WFC**; Sathe et al., 2020) consists of 125K claims also derived from Wikipedia, but uses binary (SUPPORTED, REFUTED) labels. SUPPORTED claims are Wikipedia sentences with at least one citation and REFUTED claims are obtained via manual edits to the original (true) claims to make them false. Evidence consists of full documents cited by the claim sentence.

VitaminC (**VC**; Schuster et al., 2021) uses factual revisions made by Wikipedia editors on a large set of articles as well as synthetic revisions to evidence sentences from **F** to build a set of *contrastive* evidence sentence pairs, based on which annotators then write SUPPORTED or REFUTED claims—yielding 326K claim-evidence pairs.

Finally, the work most similar to ours is WiCE (**W**; Kamoi et al., 2023), which annotates ternary support labels on *subclaims* automatically decomposed from Wikipedia sentences drawn from the SIDE dataset (Petroni et al., 2023).[3] Each subclaim is paired with annotated sets of evidence sentences from associated cited web articles. In total, WiCE contains roughly 5.4K subclaim-evidence pairs.

Claim Decomposition A wealth of recent work on *claim decomposition* has argued that the appropriate units for assessment of evidential support are *subclaims*, i.e., sub-sentence-level propositions (Kamoi et al., 2023; Min et al., 2023; Wanner et al., 2024a,b; Gunjal and Durrett, 2024, *i.a.*). Broadly, this argument contends that because subclaims assert single atomic propositions, they are easier and less ambiguous to evaluate than sentences, which may assert (or presuppose) multiple propositions. In practice, the optimal conception of atomicity is contested, and a variety of decomposition techniques have been proposed (Wanner et al., 2024b; Gunjal and Durrett, 2024). While we do not contribute to this debate, we accept the consensus that undergirds it—that subclaims are to be preferred to sentences—leveraging an existing claim decomposition method to construct PEOPLEPROFILES (§3).

Our Work Our PEOPLEPROFILES dataset differs from the resources discussed above in several key ways. First, our claims are *ecologically valid*, as they are sourced from Wikipedia articles, and not synthetically constructed or perturbed by crowdworkers (*contra* **VC**, **WFC**). Second, we annotate support on atomic *subclaims*, not full sentences, making it unambiguous which proposition's support is being assessed (*contra* **F**, **VC**, **WFC**). Third, we obtain *scalar* (not categorical) support judgments to capture variation in the degree of partial support that subclaims may have. Finally, our annotations capture *multiple levels of evidence*—enabling us to trace support from key claims in Wikipedia lead sections down into the body, and from there into cited sources. These last two features are unique to PEOPLEPROFILES.

3 Data Collection

3.1 Methodology

We obtain evidence for Wikipedia claims and scalar [-1,1] judgments of the degree of support/refutation

[2] We refer the reader to Dmonte et al. (2024) for a general survey of claim verification and focus only on Wikipedia here.

[3] Table 8 summarizes key differences between our work and **W**. Differences with **F/WFC/VC** are discussed in §2.

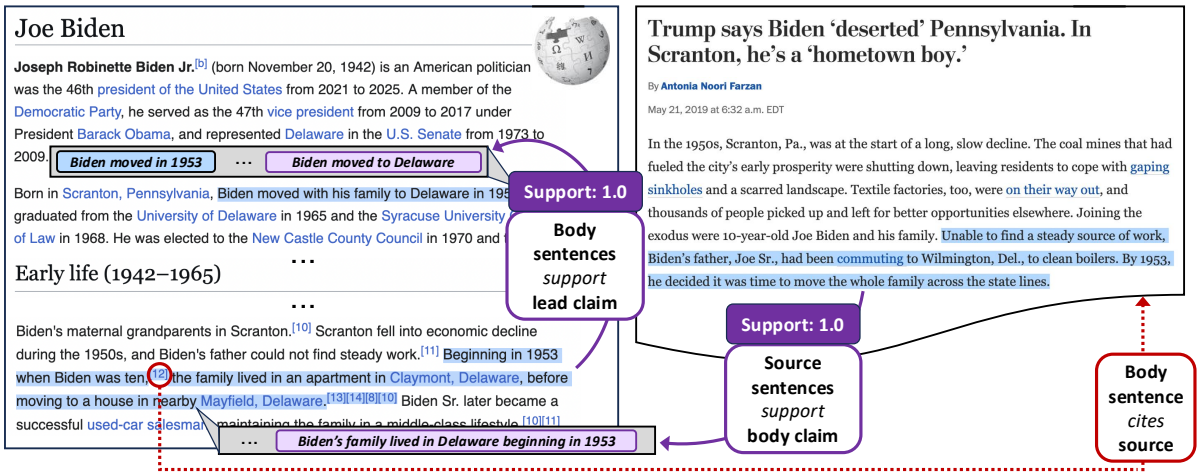

Figure 2: A more detailed view of the *multi-level structure* of PEOPLEPROFILES annotations. Claims in the *lead* of a Wikipedia article (top left) are supported by sentences in the *body* (bottom left), whose claims in turn are supported by evidence in cited sources (right). Prior work on Wikipedia claim verification has not considered this structure.

for those claims given that evidence.[4] We divide annotation into two phases—one for claims appearing in the article's *lead* and a second for claims appearing in its *body*. This is motivated by the different guidelines Wikipedia establishes for these two parts of an article: while citations are *required* for contentful claims in the body (e.g. quotations, statistics),[5] "it is common for citations to appear in the body and not the lead," since "significant information should not appear in the lead if it is not covered in the remainder of the article."[6] Thus, for lead claims, we seek evidence in the body, and for body claims, we seek evidence in cited sources. Following prior work (Kamoi et al., 2023), we define the evidence for a claim as a set of (possibly non-contiguous) sentences. We annotate up to three sentences that together provide the strongest evidence for or against each target claim.

3.2 Claims

Per §2, we adopt the view advocated in work on *claim decomposition* that the appropriate units for assessment of evidential support are *subclaims*, i.e., sub-sentence-level statements expressing an atomic proposition (Kamoi et al., 2023; Min et al., 2023; Wanner et al., 2024a,b; Gunjal and Durrett, 2024, *i.a.*).[7] We use the "DND" method of Wanner et al. (2024b) to jointly decompose each Wikipedia sentence into two sets of subclaims: a *contextualized* set decomposed from the sentence alone and a *decontextualized* set that inserts into each subclaim relevant extra-sentential context that may help resolve ambiguities in the contextualized version (e.g. unresolved pronouns). Annotators annotate *all* subclaims decomposed from a target sentence and can toggle between the contextualized and decontextualized versions of a subclaim when identifying evidence and assessing support. Following Wanner et al. (2024b), we perform the decomposition using GPT-4o-mini (OpenAI, 2024).[8]

	Train	Dev	Test
Articles	965	256	264
Lead Claims	30,331	9,272	9,351
Body Claims	60,107	19,712	18,712
Sources	10,539	3,298	3,485

Table 1: PEOPLEPROFILES summary statistics.

3.3 Data Source

Much prior work on claim decomposition and verification has focused on biographies of notable people—*entities*—as a natural test domain in which the ground truth (e.g. birth and death dates, educational background) is generally uncontroversial and where sources of evidence are numerous (Min et al., 2023; Jiang et al., 2024; Gunjal and Durrett,

[4]While refutation is uncommon in Wikipedia, we wanted to be able to capture the cases where it occurs.

[5]See Wikipedia:When_to_cite

[6]See Wikipedia:Manual_of_Style/Lead_section

[7]When we refer to *claims* in this work, we mean *subclaims* in this sense. (The claim decomposition literature often uses *claim* synonymously with *sentence*.)

[8]See Appendix A for prompts and details of DND.

2024). Given this, biographical Wikipedia pages also likely provide a rough upper bound on the degree of evidential support for claims in Wikipedia articles from other domains, which may be subject to wider debate or disagreement (e.g. theoretical frameworks, religions, or political movements). This upper bound would thus strengthen findings of a *lack* of evidential support in these articles (§4).

Accordingly, we follow this line of prior work and annotate the Wikipedia pages for all 1,485 entities studied by Min et al. (2023) and Jiang et al. (2024), representing a wide range of nationalities and degrees of fame. We obtain the English articles for each entity from the MegaWika 2 dataset (Barham et al., 2025), which includes (*inter alia*) section boundaries, in-text citations, and citations' scraped source texts for each article. We annotate sets of evidence sentences and scalar support labels for subclaims decomposed from (1) *all* sentences in articles' leads and (2) *all* body sentences that bear citations to *publicly accessible* sources (whose contents are provided in MegaWika 2). We focus on *public* sources because neither we nor Wikipedia users—nor even RAG-enabled search engines—have the resources or licenses required to verify paywalled or print sources at scale.

Barham et al. (2025) acknowledge rare errors in their source scraping process for MegaWika 2, resulting in 404 responses or other incomplete rendering of source content. To mitigate this, we further filter out low-quality sources (e.g. error pages) using the DeBERTa-based (He et al., 2020) text quality classifier from NVIDIA's NeMo Curator.[9]

Finally, we construct train (965 entities), dev (256), and test (264) splits via stratified sampling of entities based on their number of lead subclaims.

3.4 Pilot Annotation

Given the large scale of our (main) bulk annotation (∼150K claims; §3.5), we adopt an automatic annotation process. To verify the quality of this process, we first conduct a pilot human annotation on a collection of 160 body claims obtained from 10 entities, divided into 3 batches. Three authors participated in the pilot, with a unique pair of annotators doubly annotating each batch using an interface and instructions that we developed.[10]

We then use these annotations to guide manual prompt engineering for the bulk annotation, assessing GPT-4o-mini on the same examples and aiming to optimize agreement with the human annotations along two dimensions: (1) average pairwise F_1 with annotated evidence sets (using exact sentence match as the underlying similarity); and (2) Krippendorff's α (with interval difference function; Krippendorff, 2018) on the scalar support labels. Model annotations from our final (best) prompt on the pilot claims yield average pairwise α=66.3 and F_1=62.7 w.r.t. human annotations. This label agreement is only slightly below *human expert* agreement reported for related works, including **F** (α=68.4) and **VC** (α=70.7), despite our use of *scalar* labels and allowance of deviation in the underlying evidence sets.[11] Agreement on evidence selection is also on par with human expert agreement on these same examples (F_1=61.4).

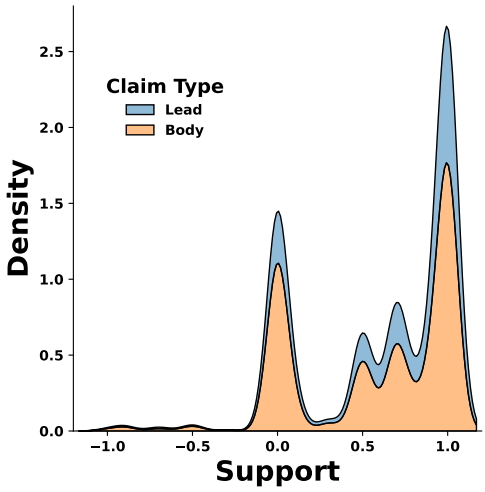

Figure 3: Kernel density estimation plots for Wikipedia lead/body claim support in the PEOPLEPROFILES dev split. We find that many claims are *not* fully grounded.

3.5 Bulk Annotation

Using GPT-4o-mini with the best prompt identified in the pilot (same for lead and body claims), we collect support and evidence annotations on all 1,485 entities. Table 1 shows statistics of the resulting PEOPLEPROFILES dataset. Table 8 further illustrates that our automatic pipeline enables annotation of roughly 20x more body claims than **W** (Kamoi et al., 2023), the most similar prior work.

4 Claim Support

In this section, we present some analysis of Wikipedia claim support enabled by PEOPLEPROFILES, including the degree of support for both

[9] https://github.com/NVIDIA/NeMo-Curator

[10] See Appendix A for interface details, annotation instructions, and annotator demographics.

[11] These other works report agreement as Fleiss's κ, which is a special case of α.

Claim Type	Category	Example Claim
LEAD	Birth/death date/location	*Josh Mansour was born in 1990.*
	Nicknames	*Michael Barakan is known as Shane Fontayne.*
	Nationality	*David Thomas Broughton is English.*
	Career Status	*Richie Dorman is retired.*
BODY	Background/Summary Claims	*Yorktown proved to be the last campaign of the Revolutionary War.*
	Hearsay/Disputed Information	*Some accounts report that Washington opposed flogging.*
	Bleached Claims	*There was a cabinet.*

Table 2: Some common categories of lead and body claims among those with an annotated support score ≤ 0.

body and lead claims, the kinds of claims that tend to be unsupported, and propagation of support from body to lead.

4.1 Support Score Distributions

We find that *significant fractions of lead and body claims are unsupported.* Figure 3 plots (kernel density estimates of) lead and body claim support distributions for the PEOPLEPROFILES dev split. We observe bimodality in both distributions, with high density around both full support (1.0) and no support (0.0). Indeed, 21.7% of lead claims are judged *unsupported* (≤ 0) by the body text and 29.5% of body claims by their cited source text(s). Despite the bimodality, a sizable proportion of claims do receive partial support (i.e. scores in (0,1)), including 32.6% of lead claims and 35.7% of body claims.

4.2 Unsupported Claims

It is natural to wonder what sorts of claims tend to receive support scores ≤ 0. We find that the answer differs between lead and body claims. Table 2 presents several common categories of unsupported claims separately for the lead and body.

For lead claims, we observe that certain key facts—ubiquitous in lead sections—often do not receive any treatment in the body, including birth and death date and place (1st row), nicknames (2nd row), nationality (3rd row), and information about career status (4th row). In some cases (e.g. nicknames, nationality), the evidence itself may be difficult to obtain or arguably may not merit a citation. But in other cases (birth/death details, career status), the observed lack of body evidence seems to represent a departure from Wikipedia guidelines.[12]

Categories of unsupported body claims are more varied, as article bodies themselves are more diverse. Here, some unsupported claims are ones that situate other facts that are more central to the article; others summarize (the import of) those facts (5th row). When the role of such claims is largely to contextualize or to recap ground that has been covered, the absence of supporting source material can perhaps be justified, though not always.

Hearsay or claims implying that some fact is disputed are another important case (6th row). In the example shown, although technically only one account of Washington's opposition to flogging is required to establish that *some accounts report that Washington opposed flogging*, other accounts in the source for this claim say the opposite. Cases like this—where evidence is conflicting—can thus result in support labels of 0 or less.

Finally, claim decomposition (even after decontextualization) can occasionally result in what Jiang et al. (2024) term *bleached claims*—ones that are highly likely to be true, even independent of any evidence (7th row). For this reason, bleached claims are generally not ones that require evidence.

4.3 Claim Support Propagation

We annotate lead claim support *given* body evidence, but we can also consider how strong the support is *for that evidence* based on cited sources. We consider two methods of computing a support score for a body evidence sentence based on the support scores of the claims decomposed from it: taking the *mean* of those scores or taking their *product*.[13] We can then compute an overall score for the *set* of evidence sentences for a given lead claim by applying the same aggregations (mean or product) across sentences.

First, we find that the majority of lead claims (82%) cannot be grounded in external source material via this method—i.e. by attempting to locate that material via its body evidence sentences. Importantly, this does *not* imply that evidence for all

[12] E.g. "Birth and death places, if known, should be mentioned in the body of the article" (Wikipedia Manual of Style).

[13] For product scores, we clip body claim scores < 0 to 0.

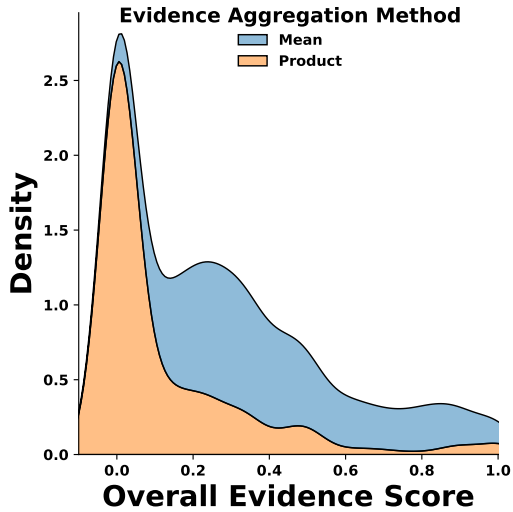

Figure 4: Distribution of overall evidence scores for PEOPLEPROFILES dev split body evidence with mean- (blue) and product-based (orange) aggregation of body claim support scores for each evidence sentence.

these lead claims is *missing* from those sources. Rather, the principal issue is that, for many lead claims, many of the body sentences in the evidence set for that claim do not bear a citation to a (publicly accessible) source. Absent a citation, we thus could not annotate claims in these body sentences for source support.

This finding underscores a key challenge for *all* NLP work focused on evidence attribution: real-world citation practices routinely deviate from the standard model (adopted here and elsewhere) in which *each* citation-worthy sentence is expected to bear supporting citations. In practice, a citation may be provided just once at the *beginning* of a paragraph (if it is clear that the whole paragraph concerns the same source) or lumped together with other citations only at the *end* of a paragraph. While such practices can obscure the mapping between claims and source material, and while they may diverge from documented standards for attribution (whether in Wikipedia or elsewhere), they are nonetheless commonplace and methods should be developed to better accommodate them.

Restricting ourselves to PEOPLEPROFILES lead claims that *can* be grounded in sources, Figure 4 plots the distribution of overall evidence scores for lead claims from the dev split under both the mean (blue) and product (orange) aggregation strategies. We find very modest overall scores in both cases (an average of 0.41 for mean and 0.35 for product).

5 Evidence Retrieval

We now turn our focus from the extent of evidential support for a target claim to retrieval of fine-grained evidence for such a claim. We consider two claim-centric evidence retrieval tasks:

1. **LEAD**: Retrieve all body evidence sentences for a given *lead* claim
2. **BODY**: Retrieve all evidence sentences from a single cited source for a given *body* claim

We also consider a third, *entity-level* retrieval task:

3. **ENTITY**: Retrieve *all* evidence sentences annotated in *all* cited sources for a given *entity*

We treat (1) and (2) as binary relevance tasks, aiming to recover the gold-annotated evidence sentences using the decontextualized claim as the query. For (3), we adopt fine-grained relevance labels, as different source material may be differently important to an entity's biography. Source sentences that support *more* claims and support them *more strongly* are assigned higher relevance.[14] For this task, we use the same query template for all entities: *Tell me about the life of* ⟨*entity*⟩*, including early life, education, career, and death.*

We report first-stage retrieval results on the PEOPLEPROFILES test set with several widely used models: BM25 (sparse, lexical; Robertson et al., 1995), ColBERTv2 (dense, multi-vector; Khattab and Zaharia, 2020; Santhanam et al., 2022), and Stella-v5-1.5B (dense, single-vector; Zhang et al., 2024). We additionally report reranking results on BM25 outputs with four recent rerankers, discussed later in this section. For **LEAD** and **BODY**, we report recall@{5,10} and NDCG@5. For **ENTITY**, we instead report recall@100, as well as NDCG@{5,100}, since there are $\gg 3$ evidence sentences per query.

5.1 First Stage Retrieval

The first three rows of Table 3 report first stage retrieval results for the **LEAD** and **BODY** tasks. On both, we obtain our best results with Stella, which shows 7+ point gains over BM25 across metrics on **LEAD** and 11+ point gains on **BODY**. ColBERT also achieves notable improvements over BM25, albeit not quite as large (2+ on **LEAD** and 8+ on **BODY**).

The first three rows of Table 4 show results on the **ENTITY** task, where both Stella and ColBERT

[14] Relevance label calculation details are in Appendix B.

Task	Model	N@5	R@5	R@10
LEAD	ColBERTv2	52.59	57.90	68.18
	Stella-1.5B-v5	59.67†	64.92†	75.12†
	BM25	49.90	55.98	65.89
	RankZephyr-7B	62.56	62.67	65.89
	Rank1-7B	60.55	63.25	65.89
	Rank-K-7B	63.33	63.84	65.89
	ReasonRank-7B	**64.24**	**64.49**	**65.89**
	ReasonRank-32B	64.24	64.41	65.89
BODY	ColBERTv2	70.02	76.37	87.16
	Stella-1.5B-v5	74.27†	80.69†	90.41†
	BM25	61.52	67.87	78.73
	RankZephyr-7B	73.07	74.40	78.73
	Rank1-7B	73.02	76.18	78.73
	Rank-K-32B	75.01	76.28	78.73
	ReasonRank-7B	75.92	76.96	78.73
	ReasonRank-32B	**76.16**	**77.04**	**78.73**

Table 3: Evidence retrieval results for the **LEAD** and **BODY** tasks. White rows are first-stage retrieval results ("†" indicates best results). Blue rows are reranking results on the top 10 evidence sentences from BM25 (best results are **bolded**). **N**=NDCG; **R**=Recall.

Model	N@5	N@100	R@100
ColBERTv2	24.53†	39.65	62.52
Stella-1.5B-v5	23.79	40.60†	64.37†
BM25	14.59	30.08	56.67
RankZephyr-7B	24.91	35.82	56.67
Rank1-7B	20.54	34.28	56.67
Rank-K-32B	**26.98**	**37.97**	**56.67**
ReasonRank-7B	24.55	36.83	56.67
ReasonRank-32B	26.35	37.60	56.67

Table 4: Evidence retrieval results for the **ENTITY** task. Blue rows are reranking results on the top 100 evidence sentences from BM25. See Table 3 for further details.

again show large improvements over BM25 (up to +10 on NDCG@100; nearly +8 on recall@100), although Stella is no longer the clear winner. ColBERT achieves best results on NDCG@5 (24.53 vs. 23.79), whereas Stella achieves a slight edge on NDCG@100 (40.60 vs. 39.65) and best results on recall@100 (64.37 vs. 62.52).

Collectively, these experiments offer some preliminary evidence that effective evidence retrieval requires more than mere lexical match. While first-stage NDCG scores with dense methods (ColBERT, Stella) are decent—at least for **LEAD** and **BODY**—the next section considers how much these results may be improved by leveraging yet more sophisticated *reasoning* methods for reranking.

5.2 Reranking via Reasoning

Recent work has shown that *reasoning* rerankers, which use reasoning chains to produce relevance judgments, achieve substantial gains on other complex retrieval tasks (Weller et al., 2025; Shao et al., 2025; Zhuang et al., 2025, *i.a.*). Our tasks—particularly the claim-centric ones (**LEAD** and **BODY**)—clearly belong in this category, as evidence may be multi-premise, with premises dispersed across the body of the article or source document. Accordingly, we evaluate several such rerankers on all three tasks:

1. **Rank1** (7B; Weller et al., 2025): A pointwise reranker based on Qwen 2.5 7B (Qwen et al., 2025) distilled from 635K reasoning traces for MS MARCO (Nguyen et al., 2016) relevance judgments produced by DeepSeek R1 (Guo et al., 2025).

2. **Rank-K** (32B; Yang et al., 2025): A listwise reranker based on QwQ 32B (Team, 2025) distilled from 50K reasoning traces from R1 on MS MARCO.

3. **ReasonRank** (7B, 32B; Liu et al., 2025): Listwise rerankers based on Qwen 2.5 instruct models (Qwen et al., 2025) distilled from R1 reasoning traces on relevance judgments across diverse domains (web search, complex QA, coding, math).

We also evaluate **RankZephyr** (7B; Pradeep et al., 2023) as a *non-reasoning*, listwise LLM reranker baseline. RankZephyr is based on Zephyr$_\beta$ (Tunstall et al., 2023) and is distilled from GPT-3.5 rank lists on MS MARCO. For all models, we rerank the top 10 evidence sentences from BM25 for **LEAD** and **BODY** and the top 100 for **ENTITY**.[15]

Results on **LEAD** and **BODY** are shown in the blue rows of Table 3, where we find large gains in NDCG@5 (10+ points) and recall@5 (6+ points) over BM25 on both tasks across all rerankers. Notably, *we achieve our highest scores on* **LEAD** *and* **BODY** *with the listwise reasoning rerankers*—Rank-K and ReasonRank—with the latter achieving best results. Interestingly, however, we do not find substantial improvements in moving from the 7B ReasonRank model to the 32B one.

[15]Thus, R@10 is unchanged between BM25 and reranking results for **LEAD** and **BODY**, as is R@100 for **ENTITY**. Additional reranking results on top of Stella rank lists are in Appendix C.

Broadly, these results echo previous findings that demonstrate superior performance from listwise reasoning rerankers relative to pointwise (Rank1) and non-reasoning models (RankZephyr) on other reasoning-intensive tasks (Yang et al., 2025; Liu et al., 2025). Further, *contra* Weller et al. (2025), Rank1 generally does not improve upon RankZephyr for our tasks, suggesting that *both* using listwise reranking *and* leveraging reasoning are key to strong performance.

5.3 Reasoning Rerankers and Evidence Complexity

A tempting hypothesis is that the effectiveness of listwise reasoning rerankers derives from their comparative advantage in handling more complex evidence—here, examples with multi-premise evidence sets. We explore this in Table 5, which reports results on **LEAD** broken down by the size of the gold evidence set (# premises), including first-stage BM25 results and reranking results with ReasonRank-7B, Rank1 (as a *pointwise* baseline), and RankZephyr (as a *non-reasoning* baseline).

The story these results tell is more complicated than the hypothesis suggests. Although we do find that ReasonRank achieves larger gains over BM25 for 2-premise (+18.3 N@5, +12.4 R@5) and 3-premise (+12.8 N@5, +9.1 R@5) evidence sets than for 1-premise sets (+12.6 N@5, +4.8 R@5), the same is also true of RankZephyr (which does not use reasoning) and Rank1 (which is not listwise). Rather, since ReasonRank achieves the best results among models over *all* evidence set sizes, a better explanation is that this model simply exhibits the strongest ability to reason about evidential support *full stop*, irrespective of evidence complexity.

Lastly, we note the large drops in performance observed across all models as one moves from 1-, to 2-, to 3-premise evidence sets. This suggests that there remains substantial room for better cross-premise reasoning, even among the sophisticated listwise rerankers studied here.

5.4 Retrieval Difficulty by Task

A final notable observation about our retrieval results is that *retrieval difficulty varies widely across tasks*. We obtain highest N@5 scores on **BODY**, followed by **LEAD** and then by **ENTITY**. Intriguingly, this ranking tracks the granularity of the queries, where body claims (**BODY**) tend to provide the most detailed information, lead claims (**LEAD**) present key high-level facts, and entity-level queries (**ENTITY**) are most general, inquiring about the life of the target entity in generic terms. Intuitively, the detailed claims appearing in Wikipedia body sections tend to bear greater lexical and semantic similarity to their supporting source material than the higher-level queries of **LEAD** or **ENTITY** do to theirs. At least for **BODY** and **ENTITY**, an initial *passage*-level retrieval stage could plausibly yield improved recall, but the core challenge of reasoning over combinations of fine-grained (sentence-level) premises would still remain.

#Sents	Model	N@5	R@5	R@10
1	BM25	76.13	86.60	91.78
	RankZephyr-7B	86.53	89.64	91.78
	Rank1-7B	84.49	90.67	91.78
	ReasonRank-7B	**88.68**	**91.38**	**91.78**
2	BM25	47.29	53.18	66.51
	RankZephyr-7B	63.90	63.28	66.51
	Rank1-7B	61.32	63.47	66.51
	ReasonRank-7B	**65.62**	**65.57**	**66.51**
3	BM25	25.25	27.26	39.11
	RankZephyr-7B	37.01	34.88	39.11
	Rank1-7B	35.40	35.28	39.11
	ReasonRank-7B	**38.08**	**36.37**	**39.11**

Table 5: Select retrieval and reranking results on **LEAD** broken down by number of gold evidence sentences.

6 Conclusion

We have presented a study of evidential support and retrieval on Wikipedia and have introduced PEOPLEPROFILES, a large dataset of fine-grained, multi-level evidential support annotations on nearly 1,500 Wikipedia articles and their cited sources. We have shown that: (1) a sizable fraction of Wikipedia *lead* claims are unsupported by their body sections, and Wikipedia *body* claims by their (publicly accessible) cited sources; (2) evidence retrieval for these claims grows much more challenging as the generality of the queries (**LEAD** $\rightarrow$ **BODY** $\rightarrow$ **ENTITY**) and the evidence complexity (1 $\rightarrow$ 2 $\rightarrow$ 3 premises) increases; and lastly (3) new reasoning-based rerankers enable much more effective retrieval of complex evidence relative to traditional (sparse or dense) methods. We release PEOPLEPROFILES to aid future work on claim verification and on advancing understanding of Wikipedia as a knowledge source for modern NLP.

Limitations

We acknowledge several limitations of our work. First, PEOPLEPROFILES focuses only on Wikipedia articles about people. We chose this focus because biographies tend to have a higher proportion of uncontroversial facts relative to other domains (e.g. concepts or events) and because multiple prior works in this area also focus on people (Min et al., 2023; Jiang et al., 2024; Gunjal and Durrett, 2024). However, it is possible support distributions or evidence retrieval difficulty could differ in other domains. (Granted, we believe domain shift would likely *strengthen* many of the claims we make, as discussed in §3.3.)

Second, as we emphasize in the main text, our claims about evidential support extend only to *publicly accessible, digital sources*. We therefore cannot make conclusions about support *across all source types* in Wikipedia. The number of paid licenses and the level of access to print resources required for this are simply infeasible to obtain.

Third, we leverage GPT-4o-mini as an annotator to facilitate our large-scale bulk data collection. While the agreement we observe between this model and our human annotations is strong (§3), LLMs have their own response biases and may not be fully calibrated when providing scalar judgments (Lovering et al., 2024).

Finally, as we note in §4.3, real-world citation practices can differ from what is espoused in documented standards—where citations often do not appear on a sentence that would seem to require one, but where they may appear elsewhere in the same paragraph. Since we restrict our annotations to body sentences that *do* bear citations, we thus cannot draw conclusions about sentences whose source may appear as a citation on some *other* sentence. However, we suspect that our estimate of the proportion of supported body claims decomposed from *citation-bearing* sentences is an upper bound on the proportion we would obtain for all *citation-worthy* sentences.

Ethics

PEOPLEPROFILES's use of sources from MegaWika 2.0 and our release of this data (via a CC-BY-4.0-SA license) is consistent with MegaWika 2's own CC-BY-4.0-SA license. Our principle transformation of the original Wikipedia articles consists in the claim decomposition, which is performed by an LLM (GPT-4o-mini), and which can occasionally result in (sub)claims that misrepresent the article's original content and thus (potentially) facts about the subject. Although our claim decompositions are generally very faithful to the original texts, users of PEOPLEPROFILES should be aware of this possibility.

References

Samuel Barham, Chandler May, and Benjamin Van Durme. 2025. Megawika 2: A more comprehensive multilingual collection of articles and their sources. *arXiv preprint arXiv:2508.03828*.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3426–3437, Online. Association for Computational Linguistics.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024. Core: Robust factual precision with informative sub-claim identification. *arXiv preprint arXiv:2407.03572*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

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.

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*.

Charles Lovering, Michael Krumdick, Viet Dac Lai, Seth Ebner, Nilesh Kumar, Varshini Reddy, Rik Koncel-Kedziorski, and Chris Tanner. 2024. Language model probabilities are not calibrated in numeric contexts. *arXiv preprint arXiv:2410.16007*.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *arXiv preprint arXiv:2407.03618*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-\efficient-intelligence/. Accessed: 2025-05-16.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, et al. 2023. Improving wikipedia verifiability with ai. *Nature Machine Intelligence*, 5(10):1142–1148.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2022. Improving wikipedia verifiability with ai.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih,

Pang Wei Koh, et al. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024a. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024b. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *arXiv preprint arXiv:2412.13175*.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

A Data Collection

A.1 Annotator Demographics

The first three authors of this work—all native English-speaking graduate students in NLP or professional NLP researchers—conducted the human pilot annotations. These authors also jointly produced the annotation instructions beforehand. None was compensated beyond their co-authorship on this work.

A.2 Claim Decomposition

Decomposition is the process of breaking down sentences into simpler, atomic components—aiming to isolate individual, independent claims for downstream applications. A common approach to claim decomposition uses LLMs to segment a sentence into independent facts, each containing one "piece" of information. However, these subclaims can be ambiguous, with vague references that are uninterpretable without the context of the document. The process of *decontextualization* mitigates this issue by rephrasing a subclaim such that it is fully intelligible as a standalone statement, without the original document as context. These two processes are complementary: decomposition divides sentences into smaller parts, whereas decontextualization adds information.

We use the "DnD" decomposition and decontextualization method introduced by Wanner et al. (2024b), which uses an LLM prompt-based method for obtaining subclaim decompositions and the corresponding decontextualized subclaims. We decompose and decontextualize sentences from the original Wikipedia page, either from the lead (in the **`LEAD`** task) or body (in the **`BODY`** task), and provide the lead paragraph (**`LEAD`**) or additionally the body paragraph from which the claim originates (**`BODY`**) as context for decontextualization. During the pilot annotation, annotators are able to toggle between the subclaim and its decontextualized version to select up to three sentences that together provide the *strongest* evidence (either supporting or refuting) for the subclaim. Finally, after identifying evidence, annotators determine a support score for the subclaim given that evidence. The bulk annotation provides only the decontextualized subclaim in the prompt. Following Wanner et al., we use GPT-4o-mini (OpenAI, 2024) to perform DnD.

A.3 Annotation Interface

The annotation interface used for the human annotation is shown in Figure 5. The full, sentence-split text of a cited source article is shown on the far left. All of the subclaims decomposed from a single Wikipedia body sentence citing that source article are shown in a vertical list of tiles on the far right, with the currently selected subclaim displayed in the top middle part of the screen (to the right of "**Claim:**"). Here, annotators can toggle between the original and decontextualized versions of the subclaim using the **D** toggle shown above the subclaim, with differences (additions, deletions) between the decontextualized and original versions shown in blue and red. Annotators can also display the sentence that the current subclaim was decomposed from, along with its full Wikipedia context, by clicking the **More Info** toggle in the top right.

Several checkboxes are also shown above the subclaim to enable annotators to indicate that:

- The source text is uninterpretable or otherwise low quality (**Bad Source**)
- The subclaim is unfaithful to the meaning of the sentence from which it was decomposed (**Bad Decontextualization**)
- It is simply too difficult to determine how the current subclaim relates to the source material—e.g. because the source document is too technical for the annotator to understand (**I'm Uncertain**)

Annotators select up to three sentences from the source text on the left that together provide the strongest evidence (either supporting or refuting) for the target subclaim. We chose a maximum of three sentences because this enabled adequate coverage of the evidence for the vast majority of claims while keeping the task tractable for annotators.

Finally, the blue box (bottom middle) is used to specify the support score for the currently selected subclaim, given the identified evidence. After selecting evidence and providing a support score for all subclaims (toggling between them using the NEXT and BACK buttons on bottom), annotators submit their work via the SUBMIT button.

A.4 Prompts and Hyperparameters

The prompt used for bulk annotation with GPT-4o-mini is shown in Figure 6 through Figure 10 (divided over multiple pages due to the length of

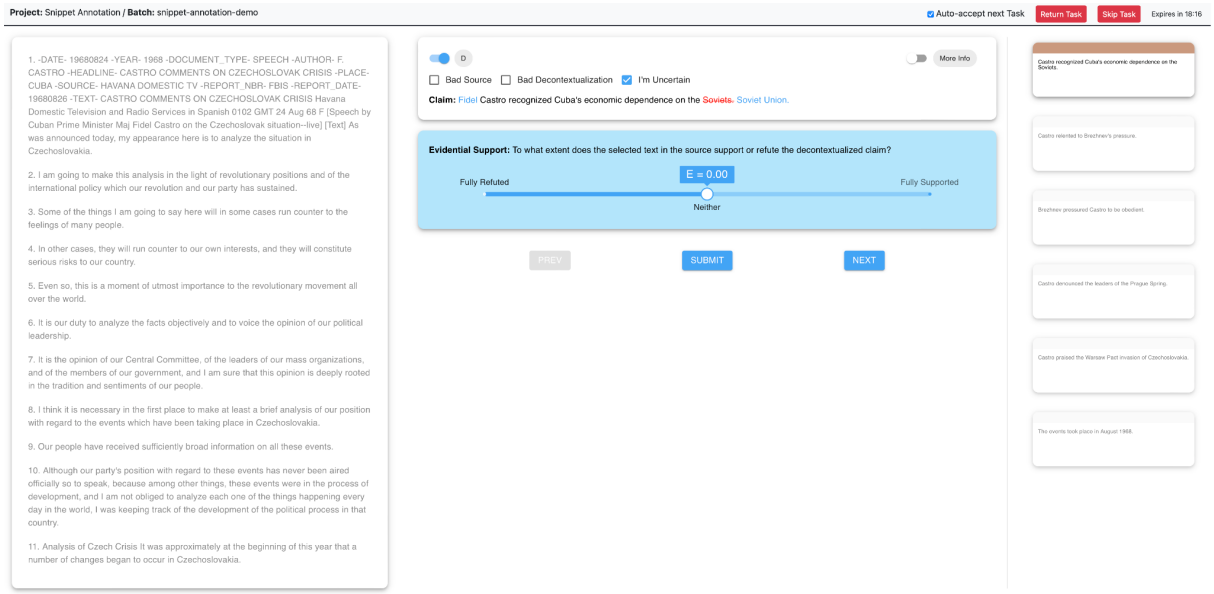

Figure 5: Annotation interface for the human pilot annotation. Detailed description can be found in Appendix A.2.

the instructions). This prompt was selected based on highest agreement with the human pilot annotations after numerous manual iterations on other prompts. We used `gpt-4o-mini-2024-07-18`, the most recent version of the model available at the time. Annotations were generated with temperature 0, with a limit of 2K output tokens to accommodate source texts of up to 126K tokens. Source texts exceeding this limit were truncated, though this was rarely required.

B Experimental Details

B.1 Qrels for ENTITY

For the **ENTITY** task in §5, we assign fine-grained relevance labels to sentences in the source documents for a given entity based on (1) how *strongly* they support a Wikipedia body claim, (2) how many body claims they support, (3) how strongly they support lead claims *via* body claims, and (4) how many lead claims they support.

Given an article for entity E, a sentence S_B in the article's body, a sentence S_S in some cited source, and a claim C, we define the following:

- $lead_E(S_B)$: the set of *lead* claims that have S_B in their (body) evidence set
- $body_E(S_S)$: the set of *body* claims that have S_S in their (source) evidence set
- $support(C)$: the support score for a claim C
- $sent(C)$: the sentence that claim C was decomposed from

Letting C_B be a body claim and C_L be a lead claim, we then define the relevance of a source sentence S_S to a query Q_E about entity E as the following weighted sum:

$$Rel(Q_E, S_S) = \sum_{C_B \in body_E(S_S)} w_{C_B} \cdot \text{abs}(support(C_B))$$

$$w_{C_B} = 1 + \sum_{C_L \in lead_E(sent(C_B))} \text{abs}(support(C_L))$$

Intuitively, $Rel(Q_E, S_S)$ is a weighted sum of the absolute values of the support scores of *all* body claims (C_B's) that S is evidence for ($body_E(S_S)$). We use the absolute value of the support score because S is equally important as evidence regardless of whether it is supporting or refuting evidence.

The weight w_{C_B} associated with each body claim C_B is 1, plus the sum of (absolute values of) support scores of all *lead* claims for which $sent(C_B)$—the sentence C_B was decomposed from—provides evidence. This rewards S for *indirectly* supporting a lead claim C_L *via a body claim* (C_B), proportional to the degree of support for C_L. The motivation here is simply that (1) lead claims typically represent more important facts about an entity than body claims, and thus sentences that (indirectly) provide evidence for them should be rewarded, and (2) that reward should be proportional to the degree of support.

We note that this is a somewhat heuristic weighting scheme, as C_B is given credit merely for being

decomposed from a sentence that supports a lead claim C_L—even if a *different* claim (C'_B) decomposed from the same sentence provides the bulk of the evidence for C_L. Collecting further annotations to enable more precise assignment of relevance scores is a direction we are pursuing for future work.

B.2 Retrieval Model Details

For BM25 (no parameters), we use the implementation provided in the `bm25s` library (Lù, 2024) with default settings. We access Stella-1.5B-v5 (1.5 billion parameters) through the `sentence-transformers` library with default settings (i.e. no hyperparameter search was performed). Finally we access ColBERTv2 (`jinaai/jina-colbert-v2` on HuggingFace; 559M parameters) via the `ragatouille` library[16], leveraging FAISS for indexing (Johnson et al., 2019), and again using default settings. Neither Stella-1.5B-v5 nor ColBERTv2 were fine-tuned on PEOPLEPROFILES. All experiments were carried out on a single NVIDIA A100 GPU except the reranking experiments, for which four A100s were used. All main text results reflect single runs.

Rank1 outputs were generated with temperature 0 and a maximum of 8,192 output tokens. We adopted the generation temperatures for listwise rerankers from the defaults listed in their public repositories and limited their outputs to 8K tokens. For listwise reranking, we used a window size of 20 with stride 10.

B.3 Reranking Prompts

Figure 11, Figure 12, Figure 13, and Figure 14 list the task-specific modifications made to the prompts used in the original reranker implementations.

B.4 Use of AI Assistants

No AI assistance was used in the ideation or in the writing of this paper. GitHub Copilot was used to assist in writing the code for some of the experiments and analysis.

C Additional Results

C.1 Additional Reranking Results

Table 6 shows select reranking results ($k = 10$) on top of Stella-1.5B-v5 outputs (rather than BM25, as in the main text) for the **LEAD** and **BODY** tasks.

[16] `https://github.com/AnswerDotAI/RAGatouille`

Task	Model	N@5	R@5	R@10
LEAD	Stella-1.5B-v5	59.67	64.92	75.12
	RankZephyr-7B	68.96	69.92	75.12
	Rank-K-32B	70.0	71.44	75.12
	ReasonRank-7B	**71.11**	**72.36**	**75.12**
BODY	Stella-1.5B-v5	74.27	80.69	90.41
	RankZephyr-7B	81.28	83.88	90.41
	Rank-K-32B	83.54	86.40	90.41
	ReasonRank-7B	**84.31**	**87.22**	**90.41**

Table 6: Reranking results for select models on outputs from Stella-1.5B-v5 on the **LEAD** and **BODY** tasks (best results are **bolded**). **N**=NDCG; **R**=Recall.

Model	N@5	N@100	R@100
Stella-1.5B-v5	23.79	40.60	64.37
RankZephyr-7B	25.28	41.14	64.37
Rank-K-32B	25.47	42.08	64.37
ReasonRank-7B	**25.99**	**42.54**	**64.37**

Table 7: Reranking results for select models on outputs from Stella-1.5B-v5 on the **ENTITY** task (best results are **bolded**). **N**=NDCG; **R**=Recall.

Similar to the results reported in Table 3, we obtain the highest scores with ReasonRank on both tasks, with comparably large gains over first-stage retrieval.

Table 7 shows the same results for the **ENTITY** task ($k = 100$). Here again, ReasonRank achieves the best scores, although the gains relative to first-stage retrieval are much smaller compared to those based on BM25 outputs (Table 4).

D WiCE Comparison

Table 8 summarizes key differences between our PEOPLEPROFILES and the most similar existing benchmark, WiCE (Kamoi et al., 2023). See §2 for a detailed comparison between PEOPLEPROFILES, WiCE, and other related works.

Dataset Characteristic	Split	WiCE	PEOPLEPROFILES (Ours)
Support Scores	—	Categorical	Scalar
Article-**internal** grounding annotations	—	✗	✓
Article-**external** grounding annotations	—	✓	✓
Subset of article-**external** subclaims annotated	—	SIDE subset (Petroni et al., 2022)	All available
Annotations per subclaim	Train	3 human	1 LLM
	Dev	5 human	1 LLM
	Test	5 human	1 LLM
Number of body subclaims	Train	3,470	60,107
	Dev	949	19,712
	Test	958	18,712

Table 8: Comparison of dataset characteristics between PEOPLEPROFILES and WiCE, the most similar prior work. PEOPLEPROFILES features fine-grained *scalar* support labels, *article-internal* support relations, and ∼20x more examples than WiCE.

PEOPLEPROFILES Annotation Prompt

In this task, you will be shown a claim along with a list of sentences representing a document that might provide evidence for the claim. Given this information, you will perform two steps, described below.

For both steps, rely on the following two definitions of evidence:
Definition 1: "Supporting evidence":
A set of sentences S provides supporting evidence for a claim c if, supposing the contents of S were true, it would give you greater reason to believe that c is true, all else equal.
Definition 2: "Refuting evidence":
A set of sentences S provides refuting evidence for a claim c if, supposing the contents of S were true, it would give you greater reason to believe that c is false, all else equal.

Step 1:

Select 0, 1, 2, or *at maximum* 3 sentence(s) from the document that provide the strongest supporting evidence or refuting evidence for the claim. If no sentences in the document provide evidence, do not select any sentences.

Additional guidelines for Step 1:
(a) You may need to use logic and common sense to *infer* that a sentence provides evidence for the claim. For example, you can use common sense to assume that a person wearing reading glasses struggles with their sight.
(b) Do not assume any parts of the claim are common knowledge. You must find evidence for all parts of the claim. For example, if the claim states that Vidya, the English chef, has poor vision, you would need to find evidence that Vidya is English and a chef, as well.
(c) A sentence might provide evidence for the claim only when combined with other sentences. For example, if Sentence A states Bob is married to Mary, and Sentence B states that Mary is a doctor, Sentences A and B together provide supporting evidence for the claim that Bob has a doctor in his family.
(d) Please make sure the entities and events in your selected sentences match those in the claim. For example, dates and names, as determined by the rest of the document, should match the claim; else, the sentences do not provide evidence.

Figure 6

PEOPLEPROFILES Annotation Prompt, continued

Step 2:

Given your selected set of sentences from Step 1, score the degree to which these sentences (taken together) support or refute the claim. Determine the score according to the following definition of a scale from -1 to 1:

-1: The claim is *fully refuted*: The claim would have to be false, supposing the sentences you selected were true.
Scores between -1 and 0 (-0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1): The claim is *partially refuted*. The claim would have to be false, but some parts are likely true.
0: The claim is neither supported nor refuted. It is equally likely to be true or false.
Scores between 0 and 1 (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9): The claim is *partially supported*. The claim is likely partially true, with missing evidence. No parts of the claim are likely to be false.
1: The claim is *fully supported*: The claim would have to be fully true, supposing the sentences you selected were true.

Additional guidelines for Step 2:
(a) Use only the content of your selected sentences to make your judgment. Do not use any knowledge you may already have about the claim, nor any context from other sentences in the document. For example, even if you know that London is in England, or it is stated elsewhere in the document, you cannot judge that detail of the claim as supported unless it is stated in your selected sentences.
(b) As in Part 1, do not assume any parts of the claim are common knowledge. Assign the score based on all parts of the claim, even if they seem obviously true or false.
(c) The document might only contain evidence for a similar but distinct claim. For example, if the strongest evidence states that the president ate at a restaurant on a Friday, this is not refuting evidence for the claim that the president ate at a restaurant on Tuesday; in fact, there is no evidence to support or refute the claim.

Figure 7

PEOPLEPROFILES Annotation Prompt, continued

Below are 10 examples of scoring sentences that have already been selected from a document as supporting or refuting evidence for a claim:

###Example 1###
Claim: "Methane Momma is a short film directed by Alain Rimbert."
Selected sentences: ["Well, good news ̆2013 last week, in the middle of one of the worst heat waves that New York has seen in recent memory, a pajama-clad (and still ripped) Van Peebles entered ex-Sun Ra bandmember Spaceman's Harlem-based studio and recorded his last takes on the rambling poem he's entitled M̈ethane Momma."̈]
Score: -0.7

###Example 2###
Claim: "Raj Kapoor was hospitalised for about a month."
Selected sentences: ["Suddenly, Kapoor collapsed, and was rushed to the All India Institute of Medical Sciences for treatment.", "The country's top cardiologists tried their best, but could not save him."]
Score: -0.1

###Example 3###
Claim: "Ottawa is a city located in the province of Ontario, Canada, and is where Matthew Perry attended school."
Selected sentences: []
Score: 0

###Example 4###
Claim: "Paul Thomas Anderson registered himself with the Writers Guild of America under the name 'Paul Anderson.'"
Selected sentences: []
Score: 0

###Example 5###
Claim: "There were exile forces opposing Idi Amin's regime."
Selected sentences: ["Since leading his guerrilla forces to Kampala in 1986, his most impressive flexibility has been his capacity to present two concurrent faces: one is that of the democratic reformer, the other is of the fearsome military ruler.", "The former is the saviour of Uganda's post-colonial collapse under presidents Milton Obote and Idi Amin, patron of democracy, and emancipator of woman and ethnic and religious minorities."]
Score: 0.1

Figure 8

PEOPLEPROFILES Annotation Prompt, continued

###Example 6###
Claim: "Margaret Rose Vendryes wrote about Richmond Barth0̆0e9's work further in her 2008 book."
Selected sentences: ["By coincidence, Dr. Vendryes was the Schomburg's scholar-in-residence and was researching her Princeton doctorate thesis on Barthe, which evolved into her landmark book Casting Feral Benga: A Biography of Richmond Barth0̆0e9's Signature Work."]
Score: 0.3

###Example 7###
Claim: "Margaret Rose Vendryes gave a lecture in 2015."
Selected sentences: ["This Thursday, February 5 at the Jepson Center, Dr. Vendryes will give the opening lecture for the exhibition."]
Score: 0.5

###Example 8###
Claim: "The exhibit presented by The New York Public Library for the Performing Arts was extensive."
Selected sentences: ["Curated by Doug Reside, the Lewis B. and Dorothy Cullman curator of the library's Billy Rose Theatre Division, the installation will run through March 31, 2020, and feature original costumes, set models, and archival video tied to Prince's productions, including models for several productions.", "The full display will honor the more than six-decade legacy of Prince.", "An open cabaret stage will allow viewers to perform songs from his shows or record their own stories about their experience with Prince's theatrical work to add to the live nature of the homage."]
Score: 0.7

###Example 9###
Claim: "The location of Matthew Perry's funeral was Forest Lawn Memorial Park (Hollywood Hills), a cemetery."
Selected sentences: ["Photo: David M. Benett/Dave Benett/Getty Matthew Perry's loved ones gathered for the actor's funeral on Friday.", "The service was held at Forest Lawn Memorial Park in Los Angeles near Warner Bros. Studios.,"] Score: 0.9

###Example 10###
Claim: "The promotional video was 60 minutes long."
Selected sentences: ["Microsoft made a c̈yber sitcomẗo promote it.", "The final product [debuted on VHS on August 1, 1995](https://books.google.com/books?id=0QsEAAAAMBAJ&lpg=RA1-PA62&dq=matthew%20perry%20jennifer%20aniston%20windows%2095&pg=RA1-PA62#v=onepage&q&f=false), satisfying everybody who wished Friends were an hour long, had four fewer friends, and involved a guide to file management."]
Score: 1

Figure 9

PEOPLEPROFILES Annotation Prompt, continued

Finally, here are the claim and list of document sentences for your task:

Claim: <subclaim>

Document sentences:
<numbered source sentences>

Write your response in a dictionary in the format shown below. Write the dictionary and nothing else.
Dictionary format:
"sentences": [
"[<sentence number>] <sentence selected from document>",
...,
],
"score": <number between -1 and 1>

###Your Task###
Selected sentences and score in dictionary form:

Figure 10

Task prompt for pointwise evidence reranking: `BODY` and `LEAD`

The following is a claim: <claim>
A relevant passage provides supporting or refuting evidence for the claim.

Figure 11: This task prompt takes the place of the query in the prompt used by the original Rank1 implementation.

Task prompt for pointwise evidence reranking: `ENTITY`

I am writing an encyclopedia article about the following person: <entity>. A relevant passage contains noteworthy biographical facts about this person. For example, a passage containing facts about this person's early life, education, career, or death is relevant.

Figure 12: Like the task prompt in Figure 11, this task prompt takes the place of the query in the prompt used by the original Rank1 implementation.

Task prompt for listwise evidence reranking: `BODY` and `LEAD`

The query given below is a claim. A relevant passage provides supporting or refuting evidence for the claim.

Figure 13: This task prompt is prepended to the user message in the reranker's original implementation, which passes prompts to the model using the conversational format. The original system message, if it exists, is retained.

Task prompt for listwise evidence reranking: `ENTITY`

I am writing an encyclopedia article about the following person given in the query below. A relevant passage contains noteworthy biographical facts about this person. For example, a passage containing facts about this person's early life, education, career, or death is relevant.

Figure 14: Like the task prompt in Figure 13, this task prompt is prepended to the user message in the reranker's original implementation, which passes prompts to the model using the conversational format. The original system message, if it exists, is retained.