How Grounded is Wikipedia?
A Study on Structured Evidential Support

Anonymous ACL submission

Abstract

Wikipedia is a critical resource for modern NLP, serving as a rich source of current 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 provides a quantitative analysis of the extent to which Wikipedia *is* so grounded and of how readily grounding evidence may be retrieved. To this end, we introduce PEOPLEPROFILES—a large-scale, multi-level dataset of claim support annotations on Wikipedia articles of notable people—and show both that a surprising proportion of Wikipedia claims (20-27%) are in fact *unsupported* by publicly accessible sources and, further, that recovery of complex grounding evidence for claims that *are* supported remains a challenge for standard retrieval methods.[1]

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 *fact* or *claim verification* within NLP (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.[2]

First, we ask to what extent claims in Wikipedia are *grounded*. Acknowledging Wikipedia's distinction between an article's *lead* (i.e. intro) section and its *body*, we are the first to jointly explore both how claims in the lead are grounded in the body (article-**internal** support) and how claims in the body are in turn grounded in cited sources (article-**external** support). Second, we ask how effectively standard retrieval methods can recover evidence for (or against) these claims—either from the body (for claims in the lead) or from source documents (for claims in the body). 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 a surprising proportion of *lead* claims ($\sim$ 20%) are unsupported by the body contents *of the same article*, and an even higher proportion of *body* claims ($\sim$ 27%) are unsupported by scrapable cited sources.
- We show that even in Wikipedia, evidence for these claims is often *complex*, involving multiple premises, and that retrieval of such evidence remains challenging.

2 Data Collection

Methodology We obtain evidence for Wikipedia claims and scalar [-1,1] judgments of the degree of support/refutation for those claims given that evidence.[3] 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

[1] Code and data will be released upon paper acceptance. Data is in the supplementary materials.

[2] Of these, WICE is most similar to our work. Appendix C has a detailed discussion of differences.

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

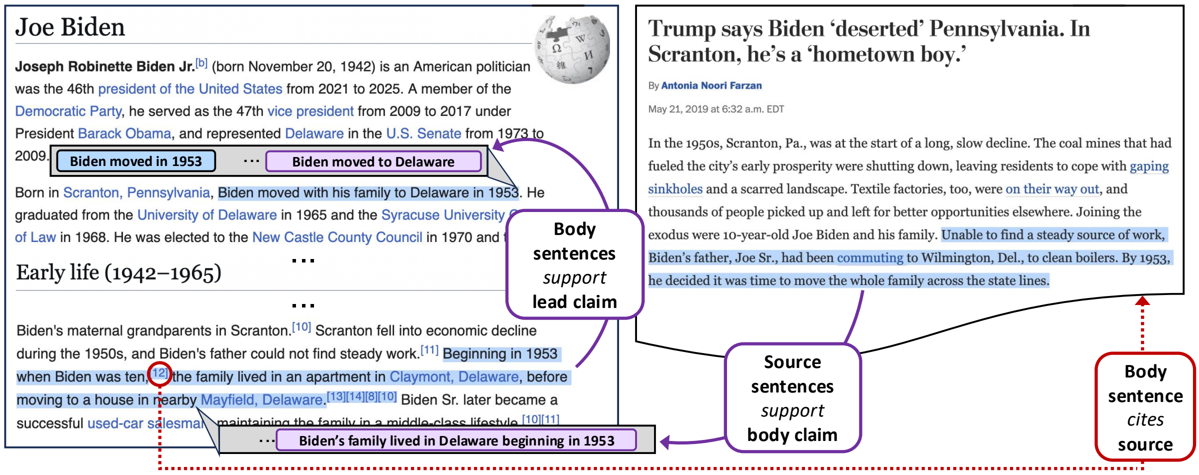

Figure 1: An example of the *multi-level structure* of PEOPLEPROFILES annotations. Claims in the *lead* of a Wikipedia article (top left) are supported by sentences in the *body* (bottom left), whose claims in turn are supported by evidence in cited sources (right). Prior work on Wikipedia claim verification has not attended to this structure.

is motivated by the different guidelines Wikipedia establishes for these two parts of an article: while citations are *required* for key claims in the body (e.g. quotations, statistics),[4] "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."[5] 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 3 sentences that together provide the strongest evidence for or against each target claim.

Claims We adopt the view championed 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.*).[6] 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 (e.g. to resolve pronouns). Following Wanner et al., we use GPT-4o-mini (OpenAI, 2024) to perform the decomposition.[7] Annotators can see both versions of a subclaim when assessing its support.

	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.

Data Source We select for annotation Wikipedia articles of notable people (*entities*) studied in prior work on claim verification, including the full sets from Min et al. (2023) and Jiang et al. (2024), yielding 1,485 entities that represent a range of nationalities and degrees of renown. We rely on data from the MegaWika project (Barham et al., 2023) to obtain the (structured) English articles for each entity, including their in-text citations and the citations' scraped source texts. We annotate claim support for subclaims decomposed from *all* sentences in articles' leads and *all* body sentences that bear citations to *publicly accessible* sources, as we cannot verify paywalled or print sources at scale—nor can Wikipedia users or RAG-enabled search engines. We use the DeBERTa-based (He et al., 2020) text quality classifier from NVIDIA's NeMo Curator to filter low-quality sources.[8] We divide examples roughly 60/20/20 into train/dev/test splits via stratified sampling on the number of lead subclaims.

Pilot Annotation To ensure high-quality automatic annotation on the full entity set, we conduct a pilot human annotation on a set of 160 body claims obtained from 10 entities, divided into 3 batches. Each batch was annotated with

[4] https://en.wikipedia.org/wiki/Wikipedia:When_to_cite
[5] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section
[6] When we refer to *claims* in this work, we mean *subclaims*.
[7] See Appendix A for prompts.

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

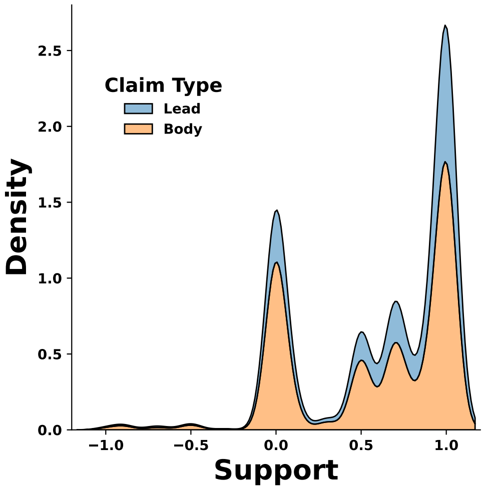

Figure 2: Kernel density estimation plots for Wikipedia lead/body claim support in the PEOPLEPROFILES dev split. We find that many claims are *not* fully grounded.

Task	Model	NDCG@5	R@5	R@10
B → **L**	ColBERTv2	**52.59**	**57.90**	**68.18**
	Stella-1.5B-v5	30.03	38.03	51.35
	BM25	49.92	56.02	66.01
+Rerank	Rank1	60.55	63.25	66.01
S → **B**	ColBERTv2	**70.02**	**76.37**	**87.16**
	Stella-1.5B-v5	49.37	60.89	78.15
	BM25	61.70	68.21	80.24
+Rerank	Rank1	73.02	76.18	80.24
S → **E**	ColBERTv2	**24.53**	**18.91**	**26.66**
	Stella-1.5B-v5	13.56	12.90	18.69
	BM25	14.59	13.29	19.29
+Rerank	Rank1	20.54	15.49	23.84

Table 2: Evidence retrieval results for lead (top) and body claims (bottom). Best first-stage results are bolded. "**+Rerank**" is reranked BM25 results ($k = 10$ for **B** → **L** and **S** → **B**; $k = 100$ for **S** → **E**).

two-way redundancy by three authors, using an interface and instructions we designed for the task (see Appendix A). We assess inter-annotator agreement on support judgments using Krippendorff's α (Krippendorff, 2018) and on the selected evidence sentences using average pairwise F_1, obtaining $\alpha = 54.3$ and $F_1 = 53.8$. We then use these results to guide prompt engineering for the bulk annotation, assessing GPT-4o-mini on the same examples, with annotations from our final prompt yielding $\alpha = 64.6$ and $F_1 = 54.1$ when included with the original human ones, indicating that GPT-4o-mini can achieve inter-human agreement levels.

Bulk Annotation Using GPT-4o-mini with the same prompt, we collect support and evidence annotations on all 1,485 entities. Table 1 shows statistics of the resulting PEOPLEPROFILES dataset.

3 Analysis & Experiments

3.1 Claim Support

A significant fraction of lead and body claims are unsupported. Figure 2 plots lead and body claim support distributions for the PEOPLEPROFILES dev split. We observe strong bimodality in both distributions, with high density around both full support (1.0) and no support (0.0). Indeed, 19.3% of lead claims are judged *unsupported* (≤ 0) by the body text and 26.5% of body claims by their cited source text(s). Inspection reveals that, contrary to guidelines, many leads make assertions attested nowhere else in the article—notably, about birth and death date and location—while other unsupported claims present inherently difficult attribution problems (e.g. nickname origins). Similarly, many body claims assert propositions unattested in publicly available sources: numerous articles extensively cite copyrighted books or paywalled articles, which is clearly legitimate, but which places hard limits on the amount of content that can be readily verified by (human or machine) readers.

Support does not robustly propagate from sources up to lead claims. We directly annotate lead claim support given body evidence, but we can further 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 given its decomposed claims' support scores, either taking the *mean* of the scores or the *product* (clipping scores < 0 to 0 for the latter). We can then compute an *overall* score for an evidence set via the same aggregations applied to the set, yielding 4 possible overall scores. Broadly we find that (1) most lead claims (82%) do not ground out in source evidence because their body evidence sentences lack citations; (2) of those that do, average overall evidence scores are very modest (e.g. 0.41 when using *mean-mean* aggregation).[9]

3.2 Evidence Retrieval

We consider three evidence retrieval tasks:

1. **B** → **L**: Retrieve **B**ody evidence sentences for a given **L**ead claim
2. **S** → **B**: Retrieve evidence sentences from a single cited **S**ource for a given **B**ody claim
3. **S** → **E**: Retrieve *all* evidence sentences from *all* cited **S**ources for a given **E**ntity

[9] Appendix B plots these overall score distributions.

Task	#Sents	NDCG@5	R@5	R@10
B → **L**	1	88.20	88.90	72.39
	2	57.29	68.42	52.77
	3	33.72	46.83	32.13
S → **B**	1	85.71	91.93	75.78
	2	66.13	79.68	59.71
	3	48.28	66.01	45.98

Table 3: Retrieval results for ColBERTv2 broken down by number of evidence sentences. Retrieval performance drops sharply as amount of evidence increases.

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 variably central to an entity's biography. Source sentences that support *more* claims and support them *more strongly* are assigned higher relevance (details in Appendix B). Here, we use the query: *Tell me about the life of ⟨entity⟩, including early life, education, career, and death.*

For all three settings, we report recall@{5,10} and NDCG@5 results on the PEOPLEPROFILES test set using several widely used retrieval models: BM25 (Robertson et al., 1995), ColBERTv2 (Khattab and Zaharia, 2020; Santhanam et al., 2022), and Stella-v5 1.5B (Zhang et al., 2024).

Main Results Table 2 reports the main results for all three models on all three tasks. We consistently obtain our best results with ColBERTv2, which shows 2+ point gains across metrics on **B** → **L** and **S** → **E**, and 6+ point gains on **S** → **B**.

Evidence retrieval difficulty increases with query scope. We observe wide variability in the difficulty of different tasks, with highest scores on **S** → **B**, followed by **B** → **L** and then by **S** → **E**. Intriguingly, this ranking tracks the granularity of claims/queries, where body claims (**S** → **B**) tend to provide the most detailed information, lead claims (**B** → **L**) present key high-level facts, and entity-level queries (**S** → **E**) represent a limiting case—seeking *any* biographical information. Intuitively, highly specific body claims likely bear greater lexical and semantic similarity to their supporting material than the higher-level claims of leads or the entity-level queries do to theirs.

Evidence retrieval difficulty increases with evidence complexity. Table 3 presents retrieval results on **B** → **L** and **S** → **B** broken down by number of gold-annotated evidence sentences. Whereas retrieval performance is strong for single-sentence evidence, we observe double-digit drops in moving to 2- and 3-sentence evidence sets. This may be explained by the fact that evidential support is often *compositional*, requiring integration of independently non- or weakly supporting pieces of evidence via inference rules. Simply indexing larger passages, though tempting, would severely curtail the ability to localize the relevant evidence: the average distance between evidence sentences for dev set body claims with multi-sentence evidence sets is 8.7 sentences, expanding to 14.6 for lead claims. That this occurs even in Wikipedia indicates that complex evidence is not a niche concern.

Reasoning rerankers help. The above observations suggest that effective retrieval of complex evidence demands more sophisticated methods than lexical or semantic similarity match. Recent work shows that *reasoning-based* rerankers achieve substantial gains on other complex retrieval tasks (Weller et al., 2025; Shao et al., 2025; Zhuang et al., 2025). Accordingly, we leverage Rank1-7B, a pointwise reranker based on Qwen 2.5 7B (Qwen et al., 2025) distilled from 635K reasoning traces for MS MARCO relevance judgments produced by R1 (Guo et al., 2025). We use Rank1 to rerank the top 10 evidence sentences from BM25 for **B** → **L** and **S** → **B** and the top 100 for **S** → **E**. Results are in Table 2's "**+Rerank**" rows, where we find large gains over first-stage retrieval across all metrics—pointing to *a vital role for reasoning-based rerankers in complex evidence retrieval.* Table 4 has fine-grained results.

4 Conclusion

We have presented a study of evidential support in Wikipedia and have introduced PEOPLEPROFILES, a large new resource of fine-grained, multi-level support annotations on 1,500 Wikipedia articles and their cited sources. We have shown that: (1) a sizable fraction of Wikipedia claims are unsupported by their body text and by publicly accessible cited sources; (2) evidence retrieval for these claims grows much more challenging as query scope and evidence complexity increases; and (3) new reasoning-based rerankers open the door to much more effective retrieval of complex evidence. We release PEOPLEPROFILES to aid future work on claim verification and on furthering understanding of Wikipedia as a key resource 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 present fairly straightforward, uncontroversial facts relative to other domains (e.g. concepts or events). However, it is possible the support distributions or the difficulty of evidence retrieval for articles in these other domains could differ from what we observe here. Second, as we emphasize throughout the paper, our claims about evidential support extend only to publicly accessible, digital sources—those that a human or machine reader could readily use to verify an article's claims. We therefore cannot make conclusions about support *across all source types* in Wikipedia. Finally, 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 (§2), LLMs have their own response biases and may not be fully calibrated when providing scalar judgments (Lovering et al., 2024).

Ethics

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

References

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, et al. 2023. Megawika: Millions of reports and their sources across 50 diverse languages. *arXiv preprint arXiv:2307.07049*.

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.

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.

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, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2022. Improving wikipedia verifiability with ai.

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*.

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.

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*.

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

Three of the authors, all native English-speaking graduate or professional NLP researchers, conducted the human pilot annotations. These authors also jointly produced the annotation instructions (included in the supplementary materials) 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, often isolating individual, independent claims for downstream applications. A common approach of doing this is using LLMs, which segment a sentence into independent facts, containing one piece of information. However, these subclaims can be ambiguous, with vague references that are uninterpretable without the context of the document. In order to mitigate this issue, *decontextualization* involves 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., which uses an LLM prompt-based method for extracting decompositions and the respective decontextualized subclaims. We decompose and decontextualize sentences from the original Wikipedia page, either from the lead (in the $\mathbf{B} \rightarrow \mathbf{L}$ task) or body (in the $\mathbf{S} \rightarrow \mathbf{B}$ task), and provide the lead paragraph ($\mathbf{B} \rightarrow \mathbf{L}$) or additionally the body paragraph from which the claim originates ($\mathbf{S} \rightarrow \mathbf{B}$) as context for decontextualization. During the pilot annotation, annotators are able to toggle between the subclaim and its decontextualized version to then select evidence sentences supporting (or refuting) the subclaim, and finally determining a support score given that evidence. The bulk annotation provides only the decontextualized subclaim as lead or body claim. We use GPT-4o-mini (OpenAI, 2024) to perform the DnD method, as in Wanner et al.

A.3 Annotation Interface

The annotation interface used for the human annotation is shown in Figure 3. 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 original sentence (**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 5 through Figure 9 (divided over multiple pages due to the length of 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

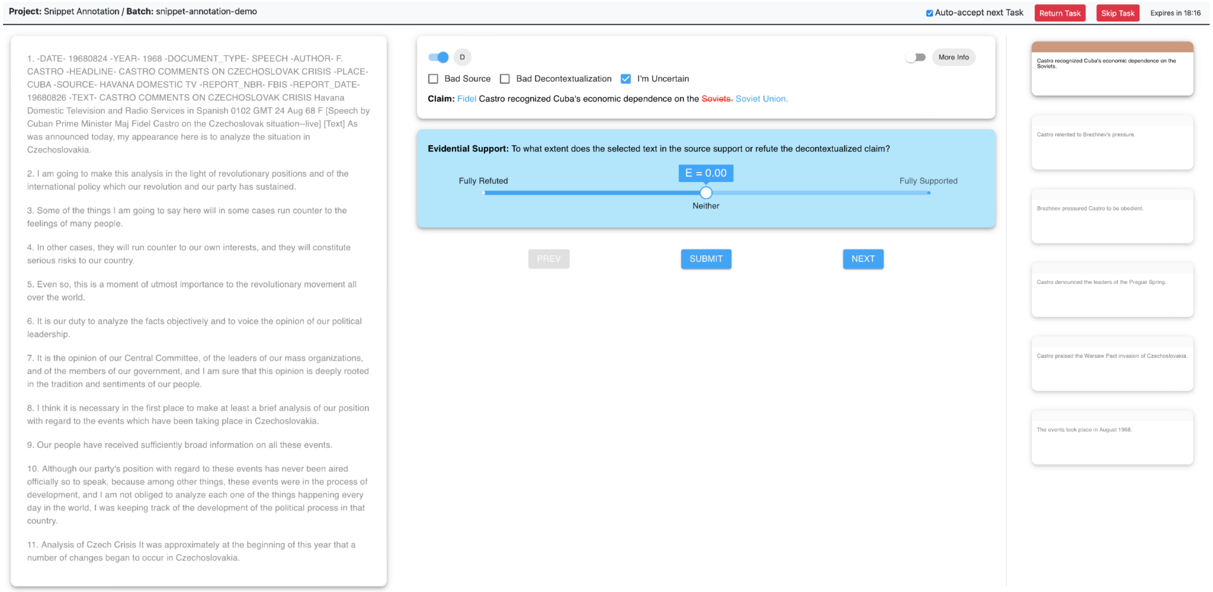

Figure 3: Annotation interface for the human pilot annotation. Detailed description can be found in Appendix A.2.

version of the model available. 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 required rarely in practice.

B Experimental Details and Additional Results

B.1 Qrels for S → E

For the **S** → **E** task in §3, 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 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:

$$\begin{aligned} Rel(Q_E, S_S) &= \textstyle\sum_{C_B \in body_E(S_S)} w_{C_B} \\ &\quad \cdot \mathrm{abs}(support(C_B)) \\ w_{C_B} &= 1 + \textstyle\sum_{C_L \in lead_E(sent(C_B))} \\ &\quad \mathrm{abs}(support(C_L)) \end{aligned}$$

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) decom-

Task	#Sents	Model	NDCG@5	R@5
B → L	1	BM25	76.14	86.64
		Rank1	**84.49**	**90.67**
	2	BM25	53.11	47.28
		Rank1	**61.32**	**63.47**
	3	BM25	25.34	27.40
		Rank1	**35.40**	**35.28**
S → B	1	BM25	75.78	85.71
		Rank1	**85.72**	**90.71**
	2	BM25	59.71	66.13
		Rank1	**71.90**	**75.38**
	3	BM25	45.98	48.28
		Rank1	**58.10**	**58.60**

Table 4: Gains from reranking the top-10 BM25 evidence sentences for **B** → **L** and **S** → **B** using Rank1, broken down by number of gold evidence sentences associated with the query. Rank1 shows major improvements in all cases.

posed 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 Fine-Grained Reranking Results

Table 4 shows the BM25 and Rank1 results from Table 2 broken down by number of evidence sentences in the gold annotations for each query (note: R@10 results are omitted, as they are unchanged by reranking the top-10 sentences). These results convincingly demonstrate that the gains brought by leveraging a reasoning model (Rank1) for reranking are not limited to the "easy" cases of single-sentence contexts but robustly extend to multi-sentence contexts as well.

B.3 Evidence Propagation

§3 briefly presents some analysis on the degree of support *for the body evidence* for a given lead claim. There, we say that we compute an evidence score for a given body sentence by taking either the mean or the product of the annotation support scores for its constituent claims (clipping negative scores to 0 in the latter case). We can then compute an *overall* evidence score for an evidence set by taking the mean or product of the per-sentence scores. Figure 4 plots distributions of overall evidence scores in the PEOPLEPROFILES dev split when applying both mean (blue) and product (orange) aggregation over claims and then (in both cases) applying mean aggregation over sentences. In both cases, we find obtain very middling overall evidence scores—an average of 0.41 for mean and an average of just

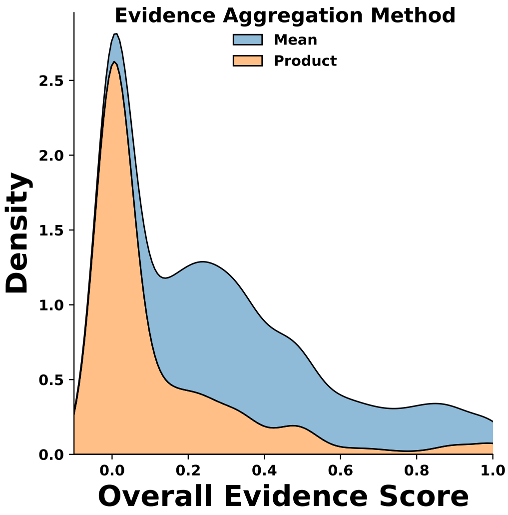

Figure 4: Distribution of overall evidence scores for PEOPLEPROFILES dev split body evidence with mean- (blue) and product-based (orange) aggregation of body claim support scores for each evidence sentence. See Appendix B.3.

0.12 for product.

B.4 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[10], 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.

The prompts for reranking evidence with Rank1-7B are provided in Figure 10 and Figure 11. Outputs were generated with temperature 0. Context size was set to 16K tokens, with a maximum of 8192 output tokens.

B.5 Use of AI Assistants

No AI assistance was used in the ideation or in the writing of this paper. GitHub Copilot was used to

[10] https://github.com/AnswerDotAI/RAGatouille

assist in writing the code for some of the experiments and analysis.

C Further Discussion of Related Work

In §1, we note that the resource most similar to PEOPLEPROFILES is the WiCE dataset from Kamoi et al. (2023), a textual entailment dataset using text-citation pairs from Wikipedia. Here, we discuss some of the key differences between our PEOPLEPROFILES and WiCE, summarized in Table 5. First, support scores in PEOPLEPROFILES are scalar, rather than categorical (`SUPPORTED`, `PARTIALLY-SUPPORTED`, `NOT-SUPPORTED`), as in WiCE, which enables finer-grained analysis of partial support (see §3). Furthermore, PEOPLEPROFILES includes *article-internal* annotations of claim support (**B** $\rightarrow$ **L**) in addition to *article-external* annotations (**S** $\rightarrow$ **B**), whereas WiCE contains only the latter. To our knowledge, ours is the first work to have both types of claim support annotations. We also annotate *all* lead sentences and *all* body sentences with attached citations, with WiCE opting to annotate only the SIDE subset (Petroni et al., 2022), containing citations unlikely to support the claim. Although PEOPLEPROFILES annotations are automated by an LLM instead of human annotation, this allows us to have a dataset over twenty times as large as WiCE.

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 Dev Test	3 human 5 human 5 human	1 LLM 1 LLM 1 LLM
Number of body subclaims	Train Dev Test	3,470 949 958	60,107 19,712 18,712

Table 5: Comparison of dataset characteristics between WiCE and our proposed PEOPLEPROFILES.

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 5

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 6

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 ž013 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 7

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 8

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 9

PEOPLEPROFILES Evidence Reranking Prompt: S $\rightarrow$ B and B $\rightarrow$ L

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

Figure 10

PEOPLEPROFILES Evidence Reranking Prompt: S $\rightarrow$ E

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 11