# Large-scale Biomedical Expert Finding for Health Claim Verification: A PubMed-based Retrieval Framework

Chaoyuan Zuo*
*School of Information and Communication*
*Nankai University*
Tianjin, China
zuocy@nankai.edu.cn

Chenlu Wang
*Department of Computer Science*
*Stony Brook University*
Stony Brook, USA
chenlwang@cs.stonybrook.edu

Ritwik Banerjee
*Department of Computer Science*
*Stony Brook University*
Stony Brook, USA
rbanerjee@cs.stonybrook.edu

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356094

***Abstract*—Verifying health claims amid rampant misinformation requires identifying qualified experts — a manual process that cannot scale. We address this challenge with a computational framework that automatically identifies biomedical researchers to evaluate health claims by analyzing their PubMed publication profiles. We establish the first benchmark for this cross-genre retrieval task, linking 93,404 health claims to 153,147 biomedical experts. Our two-stage neural pipeline addresses the semantic heterogeneity between informal health claims and formal research literature. Systematic evaluation reveals a striking finding: domain-specific models achieve 84.2% Mean Reciprocal Rank, substantially outperforming general-purpose alternatives, including state-of-the-art LLM-based rerankers fine-tuned on domain-specific data. Our findings underscore the necessity of specialized benchmarks for cross-genre information retrieval, and specialized pretraining for biomedical expert identification, while our scalable architecture enables rapid, automated expert matching for evidence-based claim verification in clinical and public health contexts.**

***Index Terms*—Information Retrieval, Medical Expert Identification, Biomedical Corpus, Clinical Misinformation**

## I. Introduction

The proliferation of health-related claims across digital platforms necessitates systematic methods for identifying qualified biomedical experts who can provide authoritative evaluation [1], [2]. As clinical evidence rapidly evolves, including emerging therapeutic approaches and novel epidemiological findings, the need for connecting these claims to domain-specific researchers has become critical [3]. This challenge is particularly acute for preliminary research findings, off-label drug applications, and rapidly developing medical situations.

Traditional computational approaches to claim verification have evolved from content-based analysis [4], [5] to evidence retrieval from structured knowledge bases [6]–[8]. However, these methods face fundamental limitations when established knowledge is insufficient or unavailable. Automated systems lack the clinical expertise necessary for nuanced interpretation of complex biomedical evidence, making expert identification essential for reliable verification [9].

*Corresponding author

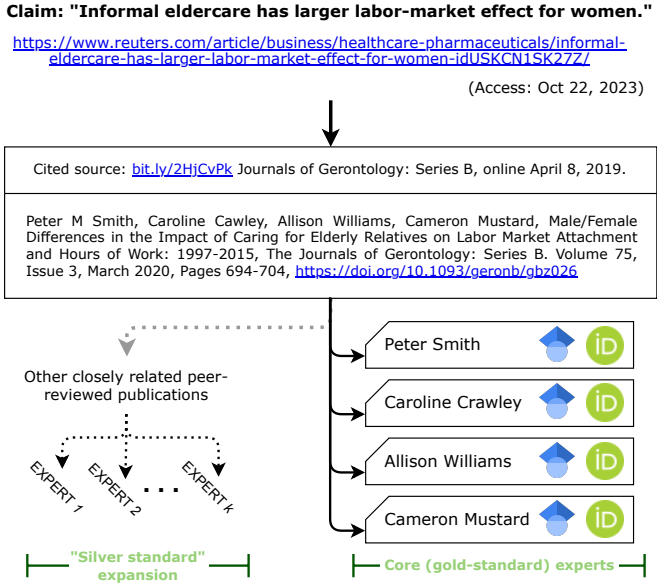


Fig. 1. **Expert identification framework, illustrated with an example:** a claim requiring verification (top) is matched to qualified researchers through their peer-reviewed publications (bottom). The expert pool comprises "gold standard" experts (authors of directly cited studies) and an expended "silver standard" set (authors of topically related research).

We present a computational framework for biomedical expert identification that bridges health claims with qualified researchers through systematic analysis of their scientific publications. Our approach is to frame expert identification (Fig. 1) as an information retrieval task, where health claims serve as queries and expert profiles are ranked by relevance. Our framework leverages PubMed-indexed literature to identify researchers whose publication profiles demonstrate relevant expertise, addressing the fundamental challenge of cross-genre retrieval between the heterogeneous language of health claims and the formal discourse of research literature. This involves representing expertise through multi-publication profiles and developing efficient retrieval methods for large expert pools.

Our framework advances beyond static knowledge bases

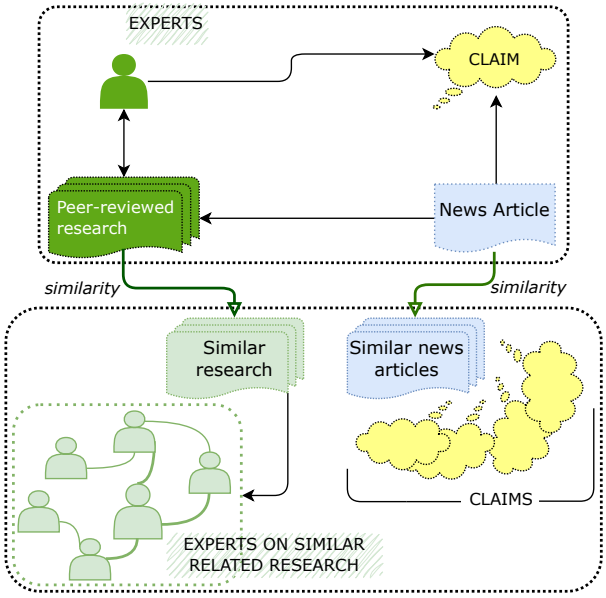


Fig. 2. **Corpus construction:** Health claims are systematically linked to expert researchers through their authored publications, forming the core dataset (top). The expanded corpus (bottom) augments this with semantically similar claims and experts who have authored topically related research, addressing sparse citation patterns in health-related content [10].

by leveraging dynamic expert knowledge encoded in recent publications. This enables identification of authorities on emerging biomedical topics where traditional repositories of evidence lack coverage. Through comprehensive evaluation on a large-scale corpus, we demonstrate that domain-specific neural models are essential for effective cross-genre retrieval in biomedical contexts, substantially outperforming general-purpose architectures. Our contributions are threefold:

1. **Large-scale biomedical benchmark**: We construct a large corpus connecting 93,404 health claims to 153,147 biomedical experts, establishing the first benchmark for cross-genre expert retrieval in clinical domains (§ II).
2. **Scalable retrieval architecture**: We develop a two-stage neural framework optimized for biomedical expert identification, achieving 84.2% MRR and 77.6% Precision@1 through efficient candidate selection and re-ranking (§ III).
3. **Domain-specific model analysis**: We demonstrate that biomedical language models significantly outperform modern general-purpose LLMs, thus empirically supporting domain specialization in clinical expert retrieval (§ IV).

## II. Biomedical Benchmark Dataset

Our corpus features a two-tier architecture, illustrated in Figure 2:[1] (i) The **core corpus** ($\mathtt{D}_{\text{core}}$), featuring health claims with explicit citations to peer-reviewed research, where cited authors constitute gold-standard experts; and (ii) an **expanded corpus** ($\mathtt{D}^{+}$), where semantically related claims and additional domain experts are identified through publication analysis. This design addresses a fundamental challenge in health information retrieval: most health claims lack direct citations to peer-reviewed research [10], [11], requiring semantic matching between heterogeneous text genres. Our expanded corpus thus provides realistic evaluation conditions for expert retrieval systems that must operate without explicit citation signals.

[1]The dataset and code are available at https://zenodo.org/records/15009723.

TABLE I
Corpus expansion ($\mathtt{D}_{\text{core}}$ to $\mathtt{D}^{+}$): impact on scale and diversity.

| | $\mathtt{D}_{\text{core}}$ | $\mathtt{D}^{+}$ | Increase (%) |
|---|---|---|---|
| Health claims | 15,119 | 93,404 | 546.2% |
| Source diversity | 9 | 134 | 1,388.9% |
| Experts per claim | 11 | 18 | 63.6% |
| Tokens per claim | 10.2 | 11.3 | 10.8% |

### A. Construction of the Core Corpus

We collected 150,028 news articles (2018-2024) from 15 RSS feeds spanning general news platforms (`Reuters`, `NYTimes`, `CNN`), and specialized medical news outlets (`MedPageToday`, `News-Medical`). We extracted bibliographic references from these articles, resolved URLs to DOIs, and queried PubMed to retrieve publication metadata (titles, abstracts, author information) [12], retaining only articles successfully linked to PubMed-indexed publications, thus yielding 16,537 articles forming the core corpus, $\mathtt{D}_{\text{core}}$.

We define **health claims** as news headlines from articles citing peer-reviewed research. Headlines serve as effective proxies for health claims as they capture the primary assertions presented to the public [13], constitute the most influential component of news articles, and provide concise text for large-scale processing. Though headlines may oversimplify complex findings, they reflect what readers actually remember and share, making them essential targets for verification.

**Claim-Evidence Validation:** News headlines frequently misrepresent cited research, with studies showing inaccuracies or exaggerations [14]–[16]. We employed GPT-4 to validate that news headlines accurately represent the research they cite, comparing against research titles and abstracts. Human validation on 200 randomly sampled pairs showed 98% agreement with automated assessments. This filtering retained 15,119 high-quality claim-evidence pairs.

We define **gold-standard experts** as authors of the studies cited in news articles. These researchers have direct knowledge of their work's methodology, findings, and limitations, making them uniquely qualified to assess claims about their research. We represent each expert through an **expert profile** comprising titles and abstracts from their recent PubMed publications (2018-24). This mirrors a natural estimation of expertise: by examining a body of relevant work rather than credentials. A small fraction (2.33%) share identical publication records, so we retain only one profile per unique publication set.

This core corpus $\mathtt{D}_{\text{core}}$ comprises 15,119 validated health claims linked to 14,393 peer-reviewed papers and 153,147 unique expert profiles (average: 11 experts per claim). Each entry is appears as a tuple $\langle c, \{e_k\}, \{r_k\}\rangle$ containing a health claim $c$, associated experts $\{e_k\}$, and their profiles $\{r_k\}$.

### B. The Expanded Corpus

$\mathtt{D}_{\text{core}}$, however, is skewed toward specialized medical outlets that consistently cite research, under-representing general news sources that typically omit citations [10], [11]. This motivates our corpus expansion, targeting claims as well as experts, to increase source diversity and expert coverage.

Many outlets report the same research without citations. Following Yang et al. [17], we identify such articles through temporal proximity ($\pm$5 days), headline semantic similarity with Sentence-BERT [18]), and named entity overlap measured by Jaccard index. Human validation of 50 randomly selected clusters showed 94% accuracy (47/50 clusters). This **clustering** added 2,879 claims from articles that did not cite any research, and expanded source diversity from 9 to 134 outlets. Further, we generated **linguistic variations** to capture the stylistic diversity in real-world health reporting. Using ChatGPT Paraphraser [19], we created two semantic paraphrases per claim. We also employed GPT-4 to transform headlines into three journalistic styles (*Breaking News*, *Analytical*, and *Investigative*), validated through ROUGE-L scores and expert review of 200 samples. For each cited publication $p$, we identify additional experts who both (i) cite $p$ in their work and (ii) publish thematically similar papers (identified through PubMed's "Similar Articles"[2], and retaining the top 50). This dual-criteria approach ensures the expanded experts possess relevant domain knowledge, while increasing average experts per claim from 11 to 18.

The expanded corpus $\mathtt{D}^{+}$ contains 93,404 health claims linked to 153,147 experts: a 517.8% increase in claims and 1,388.9% increase in source diversity over $\mathtt{D}_{\text{core}}$ (Table I). We partition data into training (55,971 claims), validation (9,364), and test (28,069) sets, with expanded variants maintaining their original assignments to prevent leakage.

## III. Retrieval Framework

We evaluate our expert identification framework following the established two-stage neural retrieval pipeline: bi-encoder candidate retrieval followed by cross-encoder re-ranking [21], [22]. We then present systematic component ablations to isolate the effects of corpus expansion techniques.

For **candidate selection**, we employ diverse retrieval algorithms to narrow the search space. We evaluate token-based models BM25 [23] and its variants BM25+ [24] and BM25L [25], which provide strong baselines. For semantic matching, we implement Transformer-based bi-encoders DistilBERT [26] and MiniLM [27], both fine-tuned with Sentence-BERT [18] on MS-MARCO [28]. These balance efficiency and performance through vectorized claims and cosine similarity between embeddings. For biomedical specialization, we integrate PubMedBERT [29] (pretrained on 14M PubMed abstracts) and PubMedBERT-MS-MARCO [30], which combines domain expertise with MS-MARCO fine-tuning.

For **candidate refinement**, cross-encoders process query-document pairs jointly and output relevance scores (0-1). We create training pairs with positive labels (relevant) and negatives randomly sampled from BM25L's top candidates [10], [31], using a 1:4 positive-to-negative ratio to reflect the natural imbalance. We train DistilBERT, PubMedBERT, and PubMedBERT-MS-MARCO cross-encoders, re-ranking BM25L's top 100 candidates per claim while manually reintroducing any gold-standard experts missed by BM25L during candidate selection.

We incorporate LLM-based re-ranking to explore recent advances. Specifically, we implement BGE reranker models [32] (Gemma architecture [33]) in both zero-shot and fine-tuned settings, and evaluate Qwen3-8B [34] in zero-shot mode. Unlike smaller Transformer cross-encoders, these LLM-based rerankers leverage richer semantic representations from large-scale pretraining, potentially offering enhanced understanding of complex query-document relationships.

Representing prolific researchers presents unique challenges: **expert profiles** average 2,933 words in our corpus due to concatenating multiple abstracts, exceeding most Transformer capacities. We address this with (1) sliding windows that process overlapping chunks of expert documents and employ max-pooling aggregation to select the most relevant segments [35]; and (2) attention mechanisms to reduce computational complexity while maintaining performance on lengthy texts [36]. We further leverage structured abstracts: 36% in our corpus contain explicit 'Conclusion' sections summarizing primary research claims [37]. We extract these directly when present; and for unstructured abstracts, we fine-tune PubMedBERT to identify conclusions ($F_1$=0.91). This reduces the average size of expert profiles by 71%—from 2,933 to 855 words—while retaining essential information.

## IV. Results and Analysis

We evaluate performance using standard IR metrics: precision at $k$ ($P@k = 1, 10$) to measure the accuracy of top-ranked results, recall at $k$ ($R@k = 10, 100$) to capture relevant items across deeper ranks, mean reciprocal rank (MRR) to reward early relevant results, and mean average precision (MAP) to balance precision and recalls across all relevant items.

### A. Candidate Selection

We evaluate nine models across three groups: (a) lexical models: BM25 and its variants BM25+ and BM25L [23]–[25]; (b) general-purpose Transformers: MiniLM [27] and DistilBERT [26]; and (c) domain-specific Transformers: PubMedBERT [29] and its variants. All models except PubMedBERT$_o$ are trained on MS-MARCO. By default, models utilize 'Conclusion' sections from structured abstracts ($*$ marks the models extracting conclusions from unstructured abstracts).

Fig. 3 shows that BM25 and its variants consistently outperform Transformer models across most metrics, especially in recall, demonstrating effective lexical matching. Among neural approaches, domain-specific pretraining proves to be crucial: PubMedBERT substantially outperforms MiniLM and DistilBERT, rivaling BM25L's precision and MRR on the expanded corpus. Models exhibit complementary strengths: PubMedBERT achieves best precision@1 (41.2%) and MRR (48.7%) on $\mathtt{D}_{\text{core}}$, while BM25L* yields highest recall@100

[2]A probabilistic topic-based model proposed by Lin and Wilbur [20].

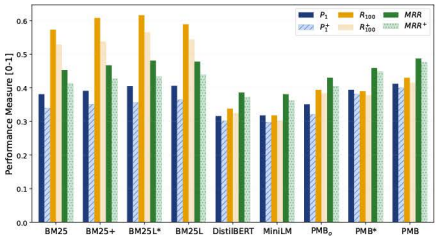


Fig. 3. **Candidate selection** on $\mathtt{D}_{\text{core}}$ and $\mathtt{D}^{+}$ corpora: performance of all nine models on three key ranking measures ($P_1$, $R_{100}$, and *MRR*).

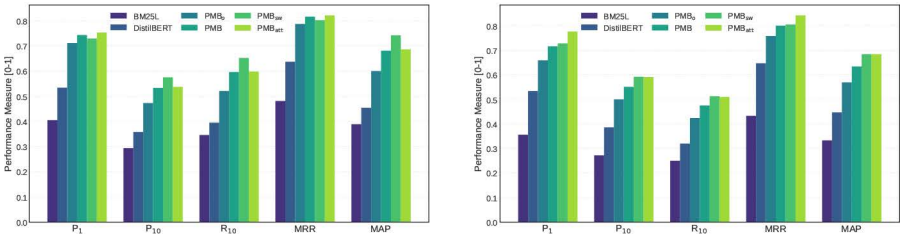


Fig. 4. **Candidate refinement** on $\mathtt{D}_{\text{core}}$ (left) and $\mathtt{D}^{+}$ (right) corpora: performance of all domain-specific models against the best lexical and general-purpose models, BM25L and DistilBERT. Results compare the five ranking measures $P_1$, $P_{10}$, $R_{10}$, $MRR$, and $MAP$.

(61.6%). Structured abstracts significantly impact performance: models extracting conclusions directly show distinct patterns, with BM25L* trading recall for precision. PubMedBERT* achieves highest MAP (0.39 on $\mathtt{D}_{\text{core}}$, 0.36 on $\mathtt{D}^{+}$), suggesting superior ranking through balanced precision-recall trade-offs. These patterns hold across both datasets, with slightly lower performance on the expanded corpus.

### *B. Candidate Refinement*

We compare domain-specific models against the best lexical (BM25L) and general-purpose (DistilBERT) baselines. We evaluate two PubMedBERT variants: PubMedBERT$_{\text{o}}$ is the original pre-trained model, while PubMedBERT is fine-tuned on MS MARCO for improved retrieval. We further introduce PubMedBERT$_{\text{att}}$ (attention mechanism) and PubMedBERT$_{\text{sw}}$ (sliding windows) for long document handling. All models are trained for 2 and 3 epochs with maximum sequence lengths of 256 and 512 tokens. Hyperparameters are selected based on development set MRR.

While BM25L excels in candidate selection, its reliance on lexical cues is a clear impediment in refinement and re-ranking. As shown in Fig. 4, it has the worst performance across all metrics due to lexical mismatch between claims and research documents. Domain-specific models, particularly PubMedBERT variants, significantly outperform all other approaches, reflecting effective capture of biomedical terminology and clinical relationships. DistilBERT achieves intermediate performance, closer to BM25L than specialized models.

Furthermore, we benchmark against modern LLM rerankers, revealing surprising limitations (Fig. 6). The BGE-reranker (based on Gemma 2B) dramatically underperforms even DistilBERT despite state-of-the-art results on standard IR benchmarks. Qwen3 (8B) outperforms BGE but substantially trails domain-specific models. However, BGE shows dramatic improvements with fine-tuning (denoted by $_{\text{ft}}$), and achieves competitive results, though still trailing PubMedBERT$_{\text{att}}$. Among domain-specific models, the use of attention mechanism yields the highest performance across most metrics, robust across both corpora. Employing sliding windows yields excellent MAP on $\mathtt{D}_{\text{core}}$ but remains at par with the attention-based model on $\mathtt{D}^{+}$. Comparing each model across corpora reveals contrasting patterns: domain-specific models perform better on $\mathtt{D}_{\text{core}}$, while general-purpose models (DistilBERT, BGE, and Qwen3) achieve higher scores on $\mathtt{D}^{+}$. This suggests the expanded corpus' linguistic diversity benefits general language understanding, partially compensating for limited domain-knowledge.

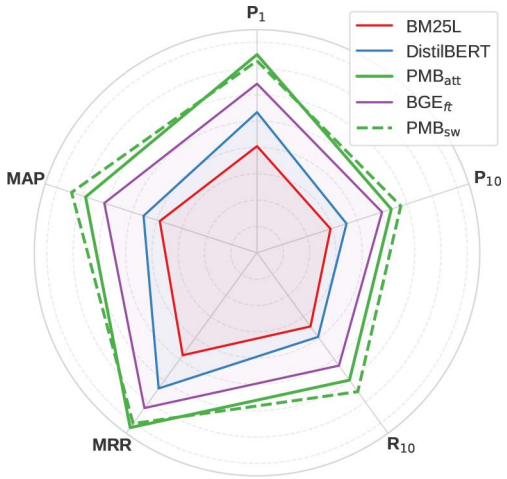


Fig. 5. **Performance comparison across all five IR metrics:** the best domain-specific models based on attention (PMB$_{\text{att}}$) and sliding windows (PMB$_{\text{sw}}$) significantly outperform the best lexical (BM25L), general-purpose cross-encoder (DistilBERT), and LLM-based reranker (BGE$_{ft}$) approaches.

**Insight 1:** These findings underscore the necessity of our cross-genre benchmark for health information retrieval. Models achieving state-of-the-art results on general IR benchmarks, including BGE-reranker and Qwen3, perform poorly on our task, revealing that standard benchmarks fail to capture the complexities of retrieving health-related expertise. While fine-tuning dramatically improves BGE-reranker, domain-specific models retain their significant advantage, demonstrating the value of purpose-built biomedical pretraining over adapting general-purpose LLMs.

**Insight 2:** Our study carries critical implications for cross-genre information retrieval. The substantial performance gap between general-purpose and domain-specific models highlights the essential role of domain-specific pretraining. The poor results from vanilla LLM rerankers demonstrate that even state-of-the-art models require significant domain adaptation for specialized cross-genre tasks. The dramatic improvement observed from fine-tuning BGE-reranker exemplifies this necessity, yet the

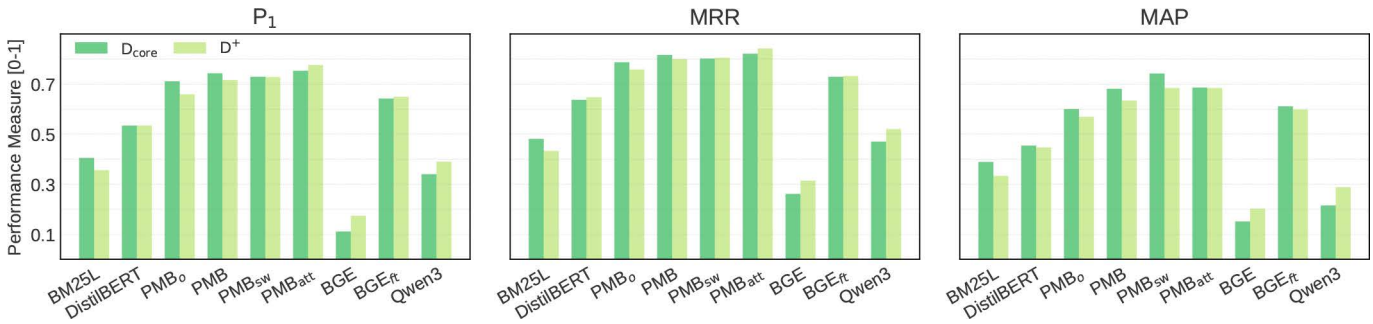


Fig. 6. Performance of all nine models across $\mathtt{D}_{\text{core}}$ and $\mathtt{D}^{+}$, compared on the basis of $P_1$, $MRR$, and $MAP$. The domain-specific models based on PubMedBERT are abbreviated as PMB, and the BGE-reranker fine-tuned on domain data is denoted by $\text{BGE}_{ft}$.

TABLE II
ABLATION STUDY RESULTS

| Model | P@1 | P@10 | R@10 | MRR | MAP |
|---|---|---|---|---|---|
| Baseline ($\text{PMB}_{\text{att}}$) | 77.6 | 59.1 | 51.0 | 84.2 | 68.4 |
| $\mathcal{B}^{\text{cl}}$ | 74.4 | 59.8 | 43.9 | 82.0 | 65.7 |
| $\mathcal{B}^{\text{para}}$ | 77.2 | 59.8 | 49.1 | 84.0 | 67.7 |
| $\mathcal{B}^{\text{st}}$ | **78.7** | **60.5** | 49.8 | **85.0** | **68.9** |

persistent advantage of purpose-built biomedical models (Fig. 5) indicates that domain-specific pretraining combined with architectural optimization remains the most effective approach for identification of biomedical experts.

### C. Ablation Study

We quantify the effect of each corpus expansion component (*news clustering*, *paraphrasing*, and *style transfer*) by training $\text{PubMedBERT}_{\text{att}}$ on $\mathtt{D}^{+}$, and evaluating on each ablated component. Table II shows that our model is robust against semantic variations: paraphrasing ($\mathcal{B}^{\text{para}}$) and style transfer ($\mathcal{B}^{\text{st}}$) maintain near-baseline performance. However, the clustering component ($\mathcal{B}^{\text{cl}}$) shows notable performance decline, particularly in recall (down to 43.9), indicating challenges in generalizing to news headlines with subtle contextual differences.

## V. Related Work

Biomedical claim verification has evolved from rule-based to neural approaches leveraging scientific knowledge bases. Wadden et al. [38] established verification benchmarks, extended to COVID-19 claims by Sarrouti et al. [7]. Studies have examined claim properties including statistical exaggeration [16] and verifiability criteria [9]. While current systems effectively match claims against biomedical repositories [6], they struggle with emerging clinical evidence. Our framework addresses this through direct expert identification from PubMed literature.

Expert finding research provides foundational methods for expertise modeling. Balog et al. [39] developed probabilistic frameworks for researcher retrieval, while Wu et al. [40] addressed multi-domain expertise. Recent work identifies medical experts in social platforms [41], though within single domains. Computational expertise representation remains challenging across document types [42]. We extend these methods to cross-genre biomedical retrieval.

Cross-genre retrieval addresses terminology and conceptual gaps between text types. Jurczyk and Choi [43] aligned conversational and formal medical texts through relation extraction. While cross-domain biomedical retrieval shows promise [44], connecting clinical claims to research literature requires specialized methods [10]. Our framework uniquely combines cross-genre retrieval with expert identification, enabling systematic consultation for emerging clinical topics where knowledge-base approaches are insufficient.

## VI. Conclusion

This work presents a computational framework for biomedical expert identification, establishing a benchmark corpus linking 93,404 health claims to 153,147 researchers through PubMed-indexed publications. Our systematic evaluation demonstrates that domain-specific models are essential for cross-genre retrieval in biomedical contexts, achieving 84.2% MRR and 77.6% precision@1 while substantially outperforming general-purpose architectures, including fine-tuned large language models. These results underscore the critical importance of domain-specific pretraining for specialized information retrieval tasks.

Our framework enables rapid expert identification for health claim verification, supporting clinical research, systematic reviews, and peer reviewer selection. The cross-encoder architecture provides quantitative relevance scores for threshold-based filtering and expertise-weighted evaluation. By automatically incorporating new PubMed publications, the system continually maintains current expertise profiles as biomedical knowledge evolves, and thus providing a foundation for automated expert consultation in clinical and research settings.

## Acknowledgment

Chaoyuan Zuo was supported by the National Natural Science Foundation of China (Grant No. 62406150). Chenlu Wang and Ritwik Banerjee were supported in part by the U.S. National Science Foundation under the award CNS-2335686.

## REFERENCES


[1] S. Medlock, S. Eslami, M. Askari, D. L. Arts, D. Sent, S. E. d. Rooij, and A. Abu-Hanna, "Health information–seeking behavior of seniors who use the internet: a survey," *Journal of Medical Internet Research*, vol. 17, no. 1, p. e3749, Jan. 2015.

[2] L. Sbaffi and J. Rowley, "Trust and credibility in web-based health information: a review and agenda for future research," *Journal of Medical Internet Research*, vol. 19, no. 6, p. e218, Jun. 2017.

[3] A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud, "Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study," *PLOS Medicine*, vol. 9, no. 9, p. e1001308, Sep. 2012.

[4] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, and G. Zuccon, "Overview of the TREC 2020 health misinformation track," in *Proc. TREC*, ser. NIST special publication, vol. 1266, 2020.

[5] W. Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection," in *Proc. ACL (Short Papers)*. ACL, 2017, pp. 422–426.

[6] N. Kotonya and F. Toni, "Explainable automated fact-checking: a survey," in *Proc. COLING*. International Committee on Computational Linguistics, Dec. 2020, pp. 5430–5443.

[7] M. Sarrouti, A. Ben Abacha, Y. Mrabet, and D. Demner-Fushman, "Evidence-based fact-checking of health-related claims," in *Findings of EMNLP*. ACL, 2021, pp. 3499–3512.

[8] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and VERification," in *Proc. NAACL-HLT*. ACL, Jun. 2018, pp. 809–819.

[9] A. Wuehrl, Y. Menchaca Resendiz, L. Grimminger, and R. Klinger, "What makes medical claims (un)verifiable? Analyzing entity and relation properties for fact verification," in *Proc. EACL*. ACL, Mar. 2024, pp. 2046–2058.

[10] C. Zuo, N. Acharya, and R. Banerjee, "Querying across genres for medical claims in news," in *Proc. EMNLP*. ACL, 2020, pp. 1783–1789.

[11] M. Ashoorkhani, J. Gholami, K. Maleki, S. Nedjat, J. Mortazavi, and R. Majdzadeh, "Quality of health news disseminated in the print media in developing countries: a case study in Iran," *BMC Public Health*, vol. 12, no. 1, p. 627, Aug. 2012.

[12] J. Wang and B. Yu, "News2PubMed: A browser extension for linking health news to medical literature," in *Proc. SIGIR*. ACM, Jul. 2021, pp. 2605–2609.

[13] X. Gu, Y. Mao, J. Han, J. Liu, Y. Wu, C. Yu, D. Finnie, H. Yu, J. Zhai, and N. Zukoski, "Generating representative headlines for news stories," in *Proc. WWW*. ACM, Apr. 2020, pp. 1773–1784.

[14] J. Schat, F. G. Bossema, M. E. Numans, I. Smeets, and P. Burger, "Exaggerated health news: association between exaggeration in university press releases and exaggeration in news media coverage," *Nederlands Tijdschrift voor Geneeskunde*, vol. 162, no. 1, pp. D1936–D1936, 2018.

[15] S. Y. Yip, D. Namah, R. Cook, and C. Isles, "It Must be True ...I Read it in the Tabloids," *Journal of the Royal College of Physicians of Edinburgh*, vol. 48, no. 3, pp. 251–256, 2018.

[16] C. Zuo, Q. Zhang, and R. Banerjee, "An empirical assessment of the qualitative aspects of misinformation in health news," in *Proc. Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. ACL, 2021, pp. 76–81.

[17] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. Archibald, and X. Liu, "Learning approaches for detecting and tracking news events," *IEEE Intelligent Systems*, vol. 14, no. 4, pp. 32–43, Jul. 1999.

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proc. EMNLP-IJCNLP*. ACL, 2019, pp. 3980–3990.

[19] V. Vorobev and M. Kuznetsov, "A paraphrasing model based on ChatGPT paraphrases," 2023.

[20] J. Lin and W. J. Wilbur, "PubMed related articles: a probabilistic topic-based model for content similarity," *BMC Bioinformatics*, vol. 8, p. 423, 2007, Also see: pubmed.ncbi.nlm.nih.gov/help/.

[21] Z. Dai and J. Callan, "Deeper text understanding for IR with contextual neural language modeling," in *Proc. SIGIR*. ACM, Jul. 2019, pp. 985–988.

[22] S. MacAvaney, A. Yates, A. Cohan, and N. Goharian, "CEDR: Contextualized embeddings for document ranking," in *Proc. SIGIR*. ACM, Jul. 2019, pp. 1101–1104.

[23] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[24] Y. Lv and C. Zhai, "Lower-bounding term frequency normalization," in *Proc. CIKM*. ACM, Oct. 2011, pp. 7–16.

[25] ——, "When documents are very long, BM25 fails!" in *Proc. SIGIR*. ACM, Jul. 2011, pp. 1103–1104.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Mar. 2020, arXiv:1910.01108 [cs].

[27] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MINILM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Proc. NeurIPS*, ser. NIPS '20. Curran Associates Inc., Dec. 2020, pp. 5776–5788.

[28] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," Nov. 2016.

[29] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 2:1–2:23, Oct. 2021.

[30] P. Deka, A. Jurek-Loughrey, and D. P., "Improved methods to aid unsupervised evidence-based fact checking for online heath news," *Journal of Data Intelligence*, vol. 3, no. 4, pp. 474–504, Nov. 2022.

[31] N. Thakur, N. Reimers, J. Daxenberger, and I. Gurevych, "Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks," in *Proc. NAACL-HLT*. ACL, Jun. 2021, pp. 296–310.

[32] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of ACL*. ACL, Aug. 2024, pp. 2318–2335.

[33] Gemma Team, Google DeepMind, "Gemma: Open models based on gemini research and technology," Apr. 2024, arXiv:2403.08295 [cs].

[34] A. Yang and et al., "Qwen3 technical report," May 2025, arXiv:2505.09388 [cs].

[35] S. Hofstätter, H. Zamani, B. Mitra, N. Craswell, and A. Hanbury, "Local self-attention over long text for efficient document retrieval," in *Proc. SIGIR*. ACM, Jul. 2020, pp. 2021–2024.

[36] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," Dec. 2020, arXiv:2004.05150 [cs].

[37] B. Yu, J. Wang, L. Guo, and Y. Li, "Measuring correlation-to-causation exaggeration in press releases," in *Proc. COLING*. International Committee on Computational Linguistics, 2020, pp. 4860–4872.

[38] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. Van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: verifying scientific claims," in *Proc. EMNLP*. ACL, 2020, pp. 7534–7550.

[39] K. Balog, L. Azzopardi, and M. de Rijke, "A language modeling framework for expert finding," *Information Processing & Management*, vol. 45, no. 1, pp. 1–19, Jan. 2009.

[40] T. Wu, Q. Wang, Z. Zhang, and L. Si, "Determining expert research areas with multi-instance learning of hierarchical multi-label classification model," in *Proc. IJCAI*, ser. IJCAI'15. AAAI Press, Jul. 2015, pp. 2305–2511.

[41] F. Haouari, T. Elsayed, and W. Mansour, "Who can verify this? Finding authorities for rumor verification in Twitter," *Information Processing & Management*, vol. 60, no. 4, p. 103366, Jul. 2023.

[42] D. Wu, S. Fan, and F. Yuan, "Research on pathways of expert finding on academic social networking sites," *Information Processing & Management*, vol. 58, no. 2, p. 102475, Mar. 2021.

[43] T. Jurczyk and J. D. Choi, "Cross-genre document retrieval: matching between conversational and formal writings," in *Proc. Workshop on Building Linguistically Generalizable NLP Systems*. ACL, 2017, pp. 48–53.

[44] T. N. Nguyen, N. L. Hai, N. D. Hieu, D. A. Nguyen, L. N. Van, T. H. Nguyen, and S. Dinh, "Improving vietnamese-english cross-lingual retrieval for legal and general domains," in *Proc. NAACL (Short Papers)*. ACL, Apr. 2025, pp. 142–153.