# Learning to Combine AI Annotations for Improved Biomedical Relevance Labeling

**Won G Kim, Lana Yeganova, Shubo Tian,**
**Donald C Comeau, W John Wilbur, Zhiyong Lu**
Division of Intramural Research (DIR), NLM, NIH, Bethesda, MD USA 20894
{wonkim,lana.yeganova,shubo.tian,donald.comeau,john.wilbur,zhiyong.lu}@nih.gov

## Abstract

Accurate labeling of relevance between biomedical abstracts is essential for improving information retrieval, semantic similarity modeling, training of ranking systems and other Natural Language Processing tasks. However, manual annotations are time-consuming, labor intensive and costly. Studies show that large language models (LLMs) can facilitate automated annotation, but their performance still falls short of human expert-level accuracy, especially in domain-specific tasks.

It has been shown that combining annotations from multiple non-expert annotators can achieve performance comparable to, or even exceeding, that of trained experts. Based on this evidence, we treat AI-generated annotations as contributions from non-expert annotators and combine them using Learning to Rank framework. Our results show significant improvement in overall annotation quality. The proposed method looks promising to reduce reliance on human annotation while maintaining reliable performance for large-scale biomedical applications.

## 1 Introduction

Recent studies have examined the effectiveness of large language models (LLMs) for annotation and relevance labeling in both general NLP and biomedical domains (Lu et al., 2024; Jin et al., 2023b). In computational linguistics, LLM annotations have been shown to effectively complement human annotation workflows (Gligoric et al., 2025). In biomedical contexts, LLMs have demonstrated strong performance in structured tasks such as clinical information extraction, but still require careful prompt design and iterative refinement to achieve reliable results (Hein et al., 2025). However, multiple studies emphasize that LLM-generated annotations remain imperfect, exhibiting biases and inconsistencies that necessitate human oversight or validation (Xia et al., 2025; Tan et al., 2024). In complex scientific tasks, including systematic review and evidence synthesis, human experts continue to outperform LLMs, highlighting limitations in domain-specific reasoning and reliability (Sollini et al., 2025; Wang et al., 2024). Overall, current evidence suggests that LLMs can approximate non-expert human annotators but fall short of expert-level performance, motivating their use as efficient yet noisy annotators rather than replacements for human expertise.

Wilbur (1996) showed the pooled results of the untrained judges were almost as good as the pooled results for the trained judges and better than any single trained judge. Motivated by this finding, we treat annotations generated by LLMs as noisy and imperfect signals. Analogous to untrained annotators, in this work, we aggregate outputs from multiple LLM judgments using diverse prompts to improve annotation quality.

Wilbur and Kim (2006) studied probabilistic methods for weighting annotators based on global statistical inference from pooled judgments, showing that aggregating multiple annotations improves the prediction of held-out labels and leads to more reliable gold standards. However, their framework treats human annotations as absolute labels rather than subjective or relative assessments. For example, when two annotators assign different relevance scores to the same document pair, this is considered a disagreement, without accounting for potential agreement in ranking.

If two annotators produce perfectly aligned rankings over a set of document pairs, they may be considered fully equivalent from a rank-



*Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, pages 502–507

ing perspective. In this study, we adopt the latter view by emphasizing agreement in ranking. This perspective is particularly aligned with traditional and modern information retrieval methods, such as BM25 (Robertson and Zaragoza, 2009) and embedding-based models such as MedCPT (Jin et al., 2023a). To incorporate such heterogeneous relevance signals, we move beyond absolute agreement and instead employ a pairwise learning-to-rank model based on gradient boosted decision trees using XGBoost (Chen and Guestrin, 2016; Burges et al., 2005), enabling effective training over diverse scoring methods.

In evaluating how well a method aligns with a gold standard, a variety of metrics have been proposed. In traditional information retrieval, Mean Average Precision (MAP) (Buckley and Voorhees, 2000) and Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) are among the most widely used measures. For assessing inter-annotator agreement, metrics such as Cohen's kappa and Spearman's rank correlation are commonly applied. MAP and NDCG place strong emphasis on top-ranked documents, making them inherently biased toward agreement at higher ranks. Cohen's kappa avoids this top-rank bias, but measures absolute agreement and does not account for consistency in ranking. In contrast, Spearman's rank correlation captures agreement in relative ordering and is not affected by bias of the top-rank or differences in absolute annotation scales, making it more suitable for evaluating ranking consistency.

## 2 Data Sets

RELISH is one of the largest expert-annotated dataset for article similarity, containing approximately 196,000 annotated pairs of PubMed abstracts derived from about 3,200 query articles. For each query article, the retrieved documents are assigned relevance scores on a three-point scale (1–3) that correspond to irrelevant, partially relevant and relevant.(Brown et al., 2020).

NCBI Dataset (Wilbur, 1996) consists of 5,000 PubMed abstract pairs derived from 100 query abstracts, each annotated on a 1–4 relevance scale by 13 judges (seven trained judges and six untrained judges).

## 3 Method

We consider seven annotation methods, including three text similarity approaches (BM25, MedCPT, and LitSense2 (Yeganova et al., 2025)) and four AI-based variants derived from different prompting strategies (Prompt 1–4). We denote these methods as F1–F7, corresponding to Prompt 1–4, BM25, MedCPT, and LitSense2, respectively. For AI-based annotation, we use GPT-5.1 and the actual prompts used for generating AI annotations are provided in the appendix. Due to the cost of annotation, we sample 272 queries from the test set defined by (Zhang et al., 2022), a subset of RELISH. This subset provides a practical balance between annotation cost and maintaining a sufficiently large sample for reliable evaluation. Hereafter, "the dataset" denotes the 272 queries together with their retrieved, annotated abstracts. As mentioned in the Introduction, given a query, our goal is to learn a model that aligns the ranking it produces with human judgments as closely as possible. To achieve this, we adopt a pairwise learning-to-rank (LTR) approach (Chen and Guestrin, 2016; Burges et al., 2005) rather than an NDCG-based LTR method commonly used in information retrieval.

## 4 Results

We randomly split the dataset into three folds, using two folds for training and the remaining one for testing. To show the qualities of individual AI rankers in the training and test sets, we compute the Spearman correlation for each individual AI ranker against the corresponding human judgments, as shown in Table 1 and Table 2.

Also, to assess how well individual untrained annotators align with trained experts, we use the NCBI dataset. Let $T$ denote the set of seven trained annotators and $U$ the set of six untrained annotators. For each trained annotator $t \in T$ and untrained annotator $u \in U$, let $\rho_{t,u}$ denote the Spearman correlation between their annotations. First, we average correlations across trained annotators for each

| Feature | Spearman |
|---|---|
| F1 | 0.326 |
| F2 | 0.320 |
| F3 | 0.354 |
| F4 | 0.397 |
| F5 | 0.280 |
| F6 | 0.314 |
| F7 | 0.362 |
| Average | 0.336 |
| Std. Dev. | 0.035 |

Table 1: Spearman correlation scores for individual features for the training set.

| Feature | Spearman |
|---|---|
| F1 | 0.337 |
| F2 | 0.351 |
| F3 | 0.376 |
| F4 | 0.407 |
| F5 | 0.277 |
| F6 | 0.327 |
| F7 | 0.372 |
| Average | 0.350 |
| Std. Dev. | 0.042 |

Table 2: Spearman correlation scores for individual features for the test set.

untrained annotator:

$$\rho_u = \frac{1}{|T|} \sum_{t \in T} \rho_{t,u} \tag{1}$$

We then average these values across all untrained annotators to obtain the overall agreement between untrained and trained annotators, denoted $\bar{\rho}_U$:

$$\bar{\rho}_U = \frac{1}{|U|} \sum_{u \in U} \rho_u \tag{2}$$

The results for $\rho_u$ for $u \in U$ and $\bar{\rho}_U$ are shown in Table 3.

From Tables 1 - Table 3, untrained human annotations may outperform both AI and traditional retrieval methods, while exhibiting similar standard deviations across the three sets. This comparison between NCBI data and RELISH data has two limitations. First, judgments on the RELISH data are on a scale from 1-3, for NCBI dataset they are on a scale from 1-4. Second, instructions to the judges on the RELISH task defined relevance as "topically relevant to the seed article; within the same specific sub-field of research, i.e. an article that would be interesting to read further or could have been cited within the original work" (Brown et al., 2020). Instructions for the NCBI data task were to "judge a document relevant to a query if and only if you would wish to read the full document if you were given the task of writing the query document."(Wilbur, 1996)

Next, we apply Learning-to-Rank (LTR) with a leave-one-out (LOO) strategy to the training set to select an optimal combination of features. We evaluated all possible combinations of the seven features (i.e., $2^7 - 1 = 127$)

| Judges | $\rho_u$ |
|---|---|
| $u_1$ | 0.4371 |
| $u_2$ | 0.3764 |
| $u_3$ | 0.3901 |
| $u_4$ | 0.3972 |
| $u_5$ | 0.4072 |
| $u_6$ | 0.3233 |
| $\bar{\rho}_U$ | 0.389 |
| Std. Dev. | 0.038 |

Table 3: Spearman correlations between trained and untrained annotators.

using the Spearman correlation measure. Table 4 shows the top 7 feature sets ranked by Spearman correlation, with the subset [F1, F2, F3, F4, F7] achieving the highest performance and selected as the optimal combination. With the optimal feature combination [F1, F2, F3, F4, F7], we retrain the model on the training set to obtain the final optimal model $M_{optimal}$. At this stage, we emphasize that the test set remains completely unseen during the model selection and training process. We then apply $M_{optimal}$ to the held-out test set to obtain predicted scores, achieving a Spearman correlation of 0.460, which markedly outperforms all individual features on the test set, with differences that are statistically significant (see Table 5).

Similarly to the evaluation of untrained human annotators against trained experts, we assess how well trained annotators agree with each other, using the same overall agreement measure defined for the untrained annotators. Let $T$ denote the set of seven trained annotators. For any pair of distinct annotators $t, t' \in T$ with $t \neq t'$, let $\rho_{t,t'}$ denote the Spearman correlation between their annotations. For

| Feature Set | Spearman |
|---|---|
| [F1,F2,F3,F4,F7] | 0.4249 |
| [F1,F2,F3,F4,F6,F7] | 0.4248 |
| [F1,F3,F4,F7] | 0.4231 |
| [F1,F2,F3,F4,F5,F7] | 0.4217 |
| [F1,F3,F4,F5,F6,F7] | 0.4207 |
| [F3,F4,F7] | 0.4200 |
| [F1,F2,F3,F4] | 0.4197 |
| [F2,F3,F4,F7] | 0.4197 |

Table 4: Top 7 feature combinations selected from all 127 subsets using leave-one-out (LOO) Learning-to-Rank, ranked by Spearman correlation on the training set.

| Feature | Spearman | p-value |
|---|---|---|
| F1 | 0.337 | 4.57e-08 |
| F2 | 0.351 | 4.29e-07 |
| F3 | 0.376 | 7.37e-07 |
| F4 | 0.407 | 4.05e-04 |
| F5 | 0.277 | 1.15e-12 |
| F6 | 0.327 | 8.62e-10 |
| F7 | 0.372 | 2.13e-09 |
| $\mathbf{M}_{optimal}$ | **0.460** | – |

Table 5: Spearman correlation on the test set and Wilcoxon signed-rank test p-values comparing individual feature-based methods against the final model $M_{optimal}$.

each annotator $t \in T$, we first average the correlations with all other annotators:

$$\rho_t = \frac{1}{|T|-1} \sum_{t' \in T, t' \neq t} \rho_{t,t'}. \quad (3)$$

We then average these values across all trained annotators to obtain the overall agreement between trained annotators, denoted $\bar{\rho}_T$:

$$\bar{\rho}_T = \frac{1}{|T|} \sum_{t \in T} \rho_t. \quad (4)$$

The results for $\rho_t$ for $t \in T$ and $\bar{\rho}_T$ are shown in Table 6. The overall Spearman correlation is $\bar{\rho}_T = 0.453$, which is lower than the Spearman correlation obtained by the model $M_{optimal}$ (0.460). This suggests that the model $M_{optimal}$ may achieve annotation quality comparable to that of human experts.

## 5 Conclusions

We explored seven heterogeneous annotation scores, comprising four LLM-based annotations generated using four different prompts and three scores derived from traditional retrieval methods, and investigate how to best combine them. Based on comparisons with human annotators, we find that each individual feature score used in our study falls short of untrained human performance. An LTR model with pairwise ranking optimization is then employed to select and combine the most effective feature combination, yielding an optimal model. The resulting model significantly outperforms each individual feature scores. Furthermore, the model may achieve annotation quality comparable to that of human experts.

| Judges | $\rho_t$ |
|---|---|
| $t_1$ | 0.4576 |
| $t_2$ | 0.4374 |
| $t_3$ | 0.4552 |
| $t_4$ | 0.5085 |
| $t_5$ | 0.3485 |
| $t_6$ | 0.4850 |
| $t_7$ | 0.4815 |
| $\bar{\rho}_T$ | 0.4534 |
| Std. Dev. | 0.052 |

Table 6: Spearman correlations among trained annotators.

## 6 Limitations and Future Work.

Our evaluation uses 272 RELISH queries due to LLM annotation cost, which may limit generalization. Comparisons with NCBI annotations are also imperfect due to differences in relevance scales, document pairs and annotator instructions. In future work, we will scale evaluation using multiple open-source LLMs instead of a single model to study a larger query set. We also aim to investigate whether aggregating or pooling outputs from multiple weaker-performing LLMs can achieve performance comparable to stronger, more expensive models. We will see how well our optimal model performs on the NCBI data.

## Acknowledgments

This research was supported [in part] by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings

and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

## References


Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2020. A large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2020:baaa039.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40. ACM.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *ICML*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

Kristina Gligoric, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident llm annotations be used for confident conclusions? In *Proceedings of NAACL 2025*.

David Hein, Alana Christie, Michael Holcomb, and 1 others. 2025. Iterative refinement and goal articulation to optimize large language models for clinical information extraction. *npj Digital Medicine*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. In *Proceedings of SIGIR*, pages 41–48.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023a. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2023b. Retrieve, summarize, and verify: How will chatgpt affect information seeking from the medical literature? *Journal of the American Society for Information Science and Technology*.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024. Large language models in biomedicine and health: Current research landscape and future directions. *Journal of the American Medical Informatics Association*, 31(9):1801–1811.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Martina Sollini, Cristiano Pini, Alexandra Lazar, Fabrizia Gelardi, Gaia Ninatti, Matteo Bauckneht, Arturo Chiti, and Margarita Kirienko. 2025. Human researchers are superior to large language models in writing a medical systematic review in a comparative multitask assessment. *Scientific Reports*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Z Wang, L Cao, B Danek, Q Jin, Z Lu, and J Sun. 2024. Accelerating clinical evidence synthesis with large language models. *arXiv preprint arXiv:2406.17755*.

W. John Wilbur. 1996. The knowledge in multiple human relevance judgments. *ACM Transactions on Information Systems*, 14(2):150–163.

W. John Wilbur and Won Kim. 2006. Improving gold standard annotation for biomedical information retrieval. In *Proceedings of the SIGIR Workshop on Biomedical Information Retrieval*.

Meng Xia, Shradha Maharjan, and 1 others. 2025. Syncode: Synergistic human–llm collaboration for enhanced data annotation. *Information*.

Lana Yeganova, Won Kim, Shubo Tian, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2025. Litsense 2.0: Ai-powered biomedical information retrieval with sentence and passage level knowledge discovery. *Nucleic Acids Research*, 53(W1):W361–W368. Open Access.

Li Zhang, Wei Lu, Haihua Chen, Yong Huang, and Qikai Cheng. 2022. A comparative evaluation of biomedical similar article recommendation. *Journal of Biomedical Informatics*, 131:104106.


## A AI Prompts

PROMPT1 = "You are an expert on the biomedical literature. Here I give you two excerpts, labelled passage A and passage B. Tell me if the entities mentioned in passage A are also being discussed in passage B. If the entities from passage A are discussed in passage B tell me if passage B seems to be saying the same thing about them or if it is just too unclear to say. After you have analyzed the passages give a final answer as

an integer: 3 if passage B says all that A says and possibly more; 2 if only part of what A says is also said in B; 1 if passage A and passage B appear to be completely unrelated. Return only a single integer from 1 to 3. Do not include explanations."

PROMPT2 ="You are an expert in biomedical literature analysis. You will receive two text excerpts labeled Passage A and Passage B. For each passage, internally determine: Sentence 1: the primary research problem addressed by the study. Sentence 2: the primary method or approach used to address that problem. Do not output these sentences; use them only for internal reasoning. Then compare Passage A and Passage B based on these internally derived sentences and output exactly one integer according to the rules below: 3: Passage B fully covers the same problem and method as Passage A, and may include additional information. 2: Passage B overlaps with Passage A in either the problem or the method, but not fully. 1: Passage A and Passage B are unrelated in both problem and method. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."

PROMPT3="You are an expert in biomedical literature analysis. You will receive two text excerpts labeled Passage A and Passage B. For each passage, internally determine the following: Sentence 1: Summarize Passage A in one sentence describing the main problem this study is trying to solve. Sentence 2: Summarize Passage B in one sentence describing the main problem this study is trying to solve. Do not output these sentences; use them only for internal reasoning. Then compare Passage A and Passage B based on these internally derived sentences and output exactly one integer according to the rules below: 3: Sentence 2 fully covers the same problem to solve as Sentence 1 and may include additional information. 2: Sentence 2 overlaps with Sentence 1 in the problem to solve, but not fully. 1: Sentence 1 and Sentence 2 are unrelated in the problems they aim to solve. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."

PROMPT4 ="You are an expert in biomedical literature analysis. You will receive two paragraphs labeled Paragraph A and Paragraph B. For each paragraph, internally determine the main conclusions being addressed in one sentence. Do not output these sentences; use them only for internal reasoning. Then compare Paragraph A and Paragraph B based on these internally derived problems and output exactly one integer according to the rules below: 3: Paragraph B addresses the same main problem as Paragraph A and may include additional information. 2: Paragraph B partially overlaps with the main problem addressed in Paragraph A. 1: Paragraph B addresses a different problem and is not relevant to Paragraph A. Output requirements: Output only one digit: 1, 2, or 3. Do not include explanations, labels, or any additional text."