SMAFIRA Shared Task at the BioNLP'2025 Workshop: Assessing the Similarity of the Research Goal

Mariana Neves[1] **Iva Sovadinova**[2] **Susanne Fieberg**[1] **Céline Heinl**[1]
diana Rubel[1] **Gilbert Schönfelder**[1,3] **Bettina Bert**[1]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]RECETOX, Masaryk University, Faculty of Science, Brno, Czech Republic
[3]Institute of Clinical Pharmacology and Toxicology,
Charité - Universitätsmedizin Berlin, Berlin, Germany

Abstract

We organized the SMAFIRA Shared in the scope of the BioNLP'2025 Workshop. Given two articles, our goal was to collect annotations about the similarity of their research goal. The test sets consisted of a list of reference articles and their corresponding top 20 similar articles from PubMed. The task consisted in annotating the similar articles regarding the similarity of their research goal with respect to the one from the corresponding reference article. The assessment of the similarity was based on three labels: "similar", "uncertain", or "not similar". We released two batches of test sets: (a) a first batch of 25 reference articles for five diseases; and (b) a second batch of 80 reference articles for 16 diseases. We collected manual annotations from two teams (RCX and Bf3R) and automatic predictions from two large language models (GPT-4omini and Llama3.3). The preliminary evaluation showed a rather low agreement between the annotators, however, some pairs could potentially be part of a future dataset.

1 Introduction

Many countries require the researchers to ask for a permission before they carry out an animal experiment (Vasbinder and Locke, 2017). Some countries, e.g., Germany, require a through search of the scientific literature in order to certify that no alternative methods are already available.

We recently developed the SMAFIRA tool[1] (Butzke et al., 2024) to support the above task. The input to the tool is a PubMed identifier (PMID) of an animal experiment, hereafter called "reference article". From PubMed, the tool retrieves the similar articles to the reference article, for which it performs two automatic tasks: (a) classification of the methods (Neves et al., 2023a), and (b) re-ranking of the retrieved similar articles.

For the latter, the goal is to rank the similar articles according the similarity of their research goal, i.e., with respect to the research goal of the corresponding reference article. Previously, we created the SMAFIRA-c dataset (Butzke et al., 2020), for which we annotated the top 100 (approximately) for four reference articles (cf. Section 2). Based on this dataset, we recently performed an evaluation of various similarity methods (Neves et al., 2023b). However, the dataset is rather small for training or even for a comprehensive evaluation of various methods.

The SMAFIRA Shared Task[2] is a collaborative effort that aimed to collect additional data for this task. We released a list of various reference articles, grouped according to some pre-selected diseases (MeSH terms). Participants were asked validate the top 20 similar articles for any number of reference articles. The similarity was assessed in terms of three labels, namely, "similar", "uncertain", and "not similar". The annotations could be performed either automatically, with any system of their choice, or manually using the SMAFIRA tool.

We describe the shared task in the next section of this publication, including the test sets, annotation tasks, guidelines, and the available dataset. In Section 3 we list the various teams (manual and automatic annotations), including how we retrieved automatic annotations from two large language models (LLMs). We give and overview of the annotations that we obtained in Section 4, as well as the computation of the agreement. Finally, we present an analysis of the annotations in Section 5.

2 SMAFIRA Shared Task

2.1 Test Sets

We compiled a list of reference articles for various disease categories. We started with a list of 23 diseases from the "Diseases (C)" category in the MeSH terms[3]. For each sub-category (MeSH

[1] https://smafira.bf3r.de/

[2] https://smafira-bf3r.github.io/smafira-st/

[3] https://meshb.nlm.nih.gov/treeView

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 388–395
August 1, 2025 ©2025 Association for Computational Linguistics

batch1					
Infections [C01]	36159784	36577999	32485164	37071015	31689515
Neoplasms [C04]	34233949	33320838	36311701	37429473	35623658
Nervous System Diseases [C10]	35709748	37084732	37339207	37749256	37126714
Cardiovascular Diseases [C14]	33635944	37010266	37380648	37268711	35917178
Immune System Diseases [C20]	34503569	36179018	37079985	37256935	37168850
batch2					
Musculoskeletal Diseases [C05]	37775153	36328744	36209953	36661300	36302840
Digestive System Diseases [C06]	26313006	34089528	36717026	30974318	34774008
Stomatognathic Diseases [C07]	32541832	34190354	33673616	35082168	37143319
Respiratory Tract Diseases [C08]	31694835	33524990	33166988	32707078	37730992
Otorhinolaryngologic Diseases [C09]	38531465	35331657	38608332	31570054	30970038
Eye Diseases [C11]	37345657	32721019	32341164	37429715	37757825
Urogenital Diseases [C12]	36581059	37324943	35264456	38688639	34270549
Hemic and Lymphatic Diseases [C15]	32001657	32494068	33639162	31797883	38713510
Congen., Heredit., and Neonatal Dis. and Abnorm. [C16]	33922602	31476705	34533563	38891999	33729473
Skin and Connective Tissue Diseases [C17]	32440554	33391503	34078596	38361478	31481954
Nutritional and Metabolic Diseases [C18]	33762572	38263084	36463128	37245586	36854163
Endocrine System Diseases [C19]	21211517	1617104	23777580	26517045	37480416
Pathological Conditions, Signs and Symptoms [C23]	33744277	32544087	26667043	38690023	24286894
Occupational Diseases [C24]	34139709	27775689	38669965	33705732	28762870
Chemically-Induced Disorders [C25]	23449255	7236062	28263289	31641018	36162952
Wounds and Injuries [C26]	26123115	31111883	29603350	19841895	16929202

Table 1: List of reference articles (test sets) for batch1 and batch2.

term) from the list, we queried PubMed with the corresponding term and for animal models[4]. Subsequently, we filtered for articles with available abstract and that were published in the last five years.

For each disease, we screened the list of results and selected five reference articles that described an animal experiment. We skipped surveys and review articles and checked that the reference article contained a pre-compiled list of similar articles. We aimed at selecting reference articles that referred to distinct diseases, e.g., distinct cancer types for the category "Neoplasms". From the original list of 23 categories, we ended up with 21 categories. We could not find five interesting animal experiments for two categories, namely, "Disorders of Environmental Origin [C21]" and "Animal Diseases [C22]".

We split the above reference articles into two groups: "batch1" and "batch2". Batch1 was released in February/2025 and contains five preselected disease categories, namely, "Infections [C01]", "Neoplasms [C04]", "Nervous System Diseases [C10]", "Cardiovascular Diseases [C14]", and "Immune System Diseases [C20]". Batch2 contains the remaining 16 disease categories and was released in the end of April/2025. Table 1 shows all reference articles for both batches.

2.2 Annotation Tasks

We proposed two annotation tasks: manual and automatic annotation. For both tasks, for any reference article, the top 20 similar articles should be annotated. The annotation should be based on the similarity of the research goal (cf. Section 2.3), and over three possible values for the similarity: "similar", "uncertain", and "not similar".

For the manual annotation, the task should be carried out in the SMAFIRA tool. Participants should enter one of the reference articles (cf. Table 1) in the input field and the tool retrieves the list of similar articles as available in PubMed. The top 20 similar articles should be annotated based on the SMAFIRA-Rank option (the default option). After the annotation, participants have two possibilities to submit their annotations to us per e-mail: (a) share their session URL, or (b) export the annotations into a file. More details about the annotation with the SMAFIRA tool is available on the web site of the shared task.

For the automatic annotation, we provide the reference articles, their corresponding top 20 similar articles, and all titles and abstracts, which were retrieved using the TeamTat tool (Islamaj et al., 2020). This data is available for download in the JSON format in the GitHub repository[5]. There is one folder for each of the batches, in which we released the following files:

[4] e.g., "(Infections[MeSH Major Topic]) AND (Models, Animal[MeSH Major Topic])"

[5] https://github.com/smafira-bf3r/smafira-st

- (a) “batch1.json” or “batch2.json”: complete JSON file with all reference articles, their respective top 20 similar articles, as well as title and abstracts for all PMIDs;
- (b) (optional) “batch1_teamtat.zip” or “batch2_teamtat.zip”: zip file with all articles as exported by TeamTat;
- “sample_submission.json”: sample submission file that include all reference articles and their similar articles, but not the labels.

2.3 Guidelines

For each pair, i.e., a reference article and one of the similar articles, our goal is to assess their similarity based on three labels: “similar”, “uncertain”, or “not similar”. We decided some simple aspects that should be taken into account during the annotation:

- The assessment should only be based on the title and the abstract, thus, the annotator should not consider the full text of the article.
- The methods should not be considered, since two research goals can be similar even if, for instance, one article describes an *in vivo* experiment and the other an *in vitro* experiment.

The actual decision of the label for a particular pair is very subjective and dependent on the opinion of the annotator. The SMAFIRA-c dataset (cf. below) has some examples that can be used for better understanding the various similarity situations. Further, we give some examples on the web site based on three aspects that were curated in the JRC’s reports (e.g., (Commission et al., 2020)), namely, application, disease, and disease feature.

The application refers to the the main scientific aim of the article or the application of the described model or method, e.g., whether the article describes the mechanism of the disease or the development of a new treatment. When addressing a certain disease, an article usually describe which specific aspects of the latter are under study, e.g., the progression of the tumor into an invasive form.

The assessment of the similarity could be based on these three aspects, though this is not mandatory. Pair of articles in which all these aspects are equal (or very similar) could certainly be tagged as “similar”. Since our annotation is based on the list of similar articles, all articles are somehow similar to the reference article. For instance, the disease is usually the same, and exceptions to this usually constitute a good reason for tagging an article as “not similar”. However, the disease feature is often not the same, or more than one are described, and their similarity (or lack of similarity) is usually the main aspect to be observed when deciding about the label. Finally, the application is also usually the same across the articles and exceptions could also be tagged as “not similar”.

2.4 Available data

Previously, we have annotated (approximately) the top 100 similar articles for four reference articles, namely, the SMAFIRA-c dataset[6] (Butzke et al., 2020). This data could be used for manually checking some annotated examples, e.g., for training purposes. Further, for automatic methods, it could be used for few-shot strategies or for the evaluation. However, given its small size, it might not be appropriate for supervised learning purposes. The mapping between the annotations in SMAFIRA-c (“Equivalence” column) and the three labels used in the shared task is shown below:

- “similar”: equivalent “++”, partially equivalent “+(+)” or “+”, noteworthy “n”
- “uncertain”: limbo “L”
- “not similar”: not equivalent “-”

3 Teams and Systems

In this section we give details of the participants of the shared task. For the sake of simplicity, we will sometimes refer to all of participants, whether manual annotators or automatic systems, as “teams” throughout this publication.

For the manual annotation, we had the participation of two teams:

- “RCX” (RECETOX, Faculty of Science, Masaryk University, Czechia);
- “Bf3R” (German Centre for the Protection of Laboratory Animals, Germany).

The annotations from “Bf3R” were carried out by five experts. Some of them annotated the same reference article in order to compare their results, but they did not try to reach a consensus. In these cases, we selected one of them as the official submission of the team.

[6] https://github.com/SMAFIRA/c_corpus

For the automatic annotations, we relied on a zero-shot approach with two LLMs: (a) the GPT-4o-mini model using the OpenAI API[7]; and (b) Llama3.3 (llama-3.3-70b-versatile) using the Groq API[8]. We provided the two texts (title and abstract) in the prompt, i.e., first the one for the reference article and then the one for one of the similar article, followed by the questions with detailed instruction on how to assess the similarity. We used the following user message:

You are a helpful assistant designed to evaluate the similarity between two texts.

and the following user content:

Text 1: REF_ARTICLE_TEXT
Text 2: SIMILAR_ARTICLE_TEXT
Are the the research goals of the two texts above similar? You should compare the research goal based on four aspects: (1) Are the disease(s) addressed in the texts the same? (2) Do they address the same characteristic symptom/feature of the disease? (3) Do they refer to the same biological endpoints, e.g., the same disease mechanism, gene/protein or chemical coumpounds? (4) Is the scientific aim or the future application of the results the same, e.g., for drug development, model development, disease treatment or diagnosis? Answer with either 'similar', 'uncertain', or 'not similar'. The answer is:

We evaluated our prompts with the cases studies of the SMAFIRA-c dataset (cf. Section 2.4) and show the statistics of the corpus (cf. Table 3) and the results (cf. Table 4) in the Appendix A. The same prompt was used for both LLMs when obtaining annotations for the shared task, as well as for the evaluation of the SMAFIRA-c dataset. We retrieved annotations from the two LLMs for all reference articles in batch1.

4 Results

4.1 Overview of the annotations

We describe the annotations that we obtained from two participants and from two LLMs. In this publication, we present results only for batch1.

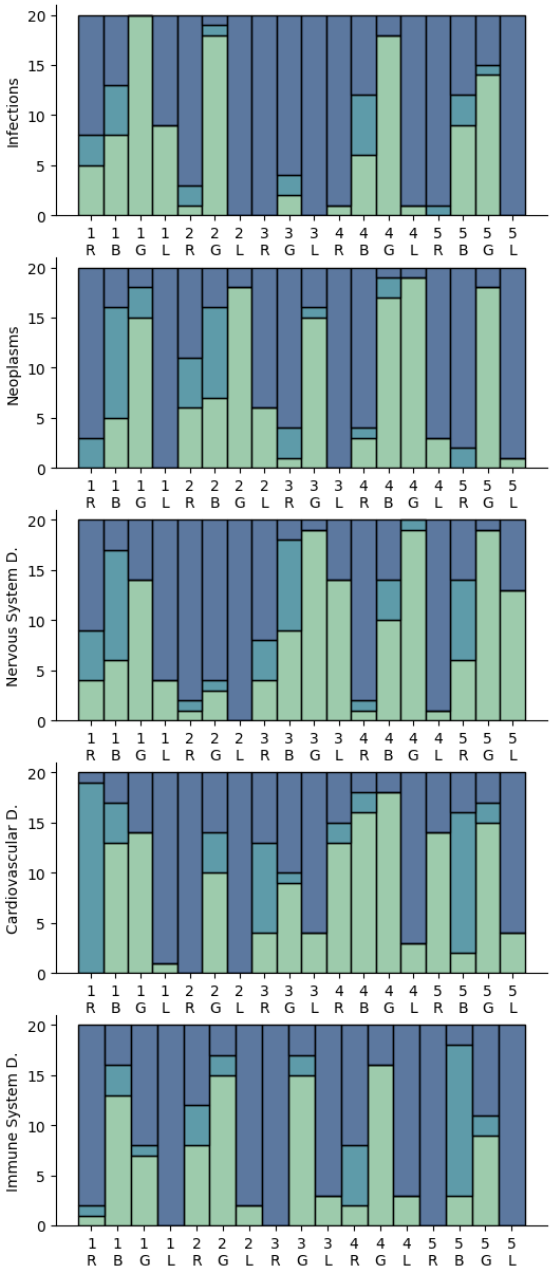

Figure 1: Overview of the annotations per disease in terms of number of annotations (y-axis). The x-axis shows the five reference article (in the order shown in Table 1) and the teams: (R)CX, (B)f3R, (G)PT-4o-mini, and (L)lama3.3. The three-value similarity is the following (from darker to lighter color, from top to bottom in each graph): "similar": dark blue (top color), "uncertain": dark blue/green (middle color), "not similar": light green (bottom color).

We obtained manual annotations for all 25 reference articles from RCX and for 14 reference articles from Bf3R. Further, we collected annotations for all reference articles from the two LLMs. Regarding the three similarity labels, we obtained the following number of annotations (from a total of 1,780):

[7] https://openai.com/

[8] https://console.groq.com/docs/model/llama-3.3-70b-versatile

948 (53%) for “similar”, 202 (11%) for “uncertain”, and 630 (35%) for “not similar”. All annotations are available in our GitHub repository.

We depict the number of annotations for each label in Figure 1, from which we can observe some differences across the teams. On the one hand, RCX and Llama3.3 frequently assigned the “similar” label to all (or most) of the similar articles of some reference articles. On the other hand, GPT-4o-mini frequently assigned the “not similar” label for all (or most) of the similar articles of some reference articles. Further, the “uncertain” label was more frequently assigned by human annotators, but rarely returned by GPT-4o-mini, and never by Llama3.3.

We did not observe considerable differences across the diseases. For all five diseases, the “similar” label was the most frequent one (45% to 62%), followed by “not similar” (28% to 41%) and “uncertain” (7% to 16%).

4.2 Agreement between teams

We analyzed the agreement of the annotations in various ways, e.g., pairwise comparison between two teams, or multiple comparison across all teams. We present the results below.

Agreement for manual annotation. For the RCX and Bf3R teams and for the 14 reference articles annotated by both teams, we observed the following (cf. Figure 1):

- Three cases with good agreement: “36159784” (no. 1) of “Infections”, “33320838” (no. 2) of “Neoplasms” and “37268711” (no. 4) of “Cardiovascular Diseases”.
- Two cases had some agreement: “35709748” (no. 1) and “37339207” (no. 3) of “Nervous System Diseases”.
- Four cases in which one assigned mostly the “uncertain” label, which might overlap with the “similar” or “not similar” labels from the other: “34233949” (no. 1) of “Neoplasms”, “33635944” (no. 1) and “35917178” (no. 5) of “Cardiovascular Diseases”, and “37168850” (no. 5) of “Immune System Diseases”.
- Five cases with very bad agreement: “37071015” (no. 4) and “31689515” (no. 5) of “Infections”, “37429473” (no. 4) of “Neoplasms”, “37749256” (no. 4) of “Nervous System Diseases”, and “34503569” (no. 1) of “Immune System Diseases”.

In general, the agreement for the manual annotation was rather good for the reference articles in the “Cardiovascular Diseases”. However, this comparison did not consider the labels for each particular article, nor agreements that might have occurred by chance.

Pairwise agreement. We computed the kappa score[9] (McHugh, 2012) for all pairwise comparison between the teams and plotted a heatmap in Figure 2. From a total of 114 pairs, 26 of them were negative, (no agreement), 42 between zero and 0.2 (slight agreement), 12 between 0.2 and 0.4 (fair agreement), four between 0.4 and 0.6 (moderate agreement), and none above these values (substantial or perfect agreement). As already observed above, there are less negative scores for the “Cardiovascular Diseases”. From the 14 reference articles annotated by human annotators, seven of them had a negative agreement. The three highest scores, namely, 0.52, 0.51, and 0.49 were obtained between RCX and Llama3.3, followed by a good agreement (0.44) by Bf3R and GPT-4o-mini.

Multiple agreement. For each reference article, we also computed the krippendorff’s alpha score[10] across annotations from all teams (whenever available). We plot the scores on Figure 3. From the 25 reference article, 22 of them were negative, which mean a systematic disagreement. The highest (and positive) scores, i.e., 0.056, 0.051, and 0.035, were obtained by the following reference articles, respectively: “37380648” and “35917178” of “Cardiovascular Diseases”, and “37126714” from “Nervous System Diseases”.

5 Discussion

5.1 Analysis of the annotations

We analyzed the articles that could already be part a future dataset, as well as potential articles that could be included after an additional round of consensus. Further, we also analyzed whether articles tagged as “similar” were usually ranked higher in the top 20 list.

Pairs of articles with high agreement. The aim of this shared task was to build a dataset for pairs of articles with respect to the similarity of their research goal. Our previous effort, i.e., SMAFIRA-c,

[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[10] https://pypi.org/project/krippendorff/

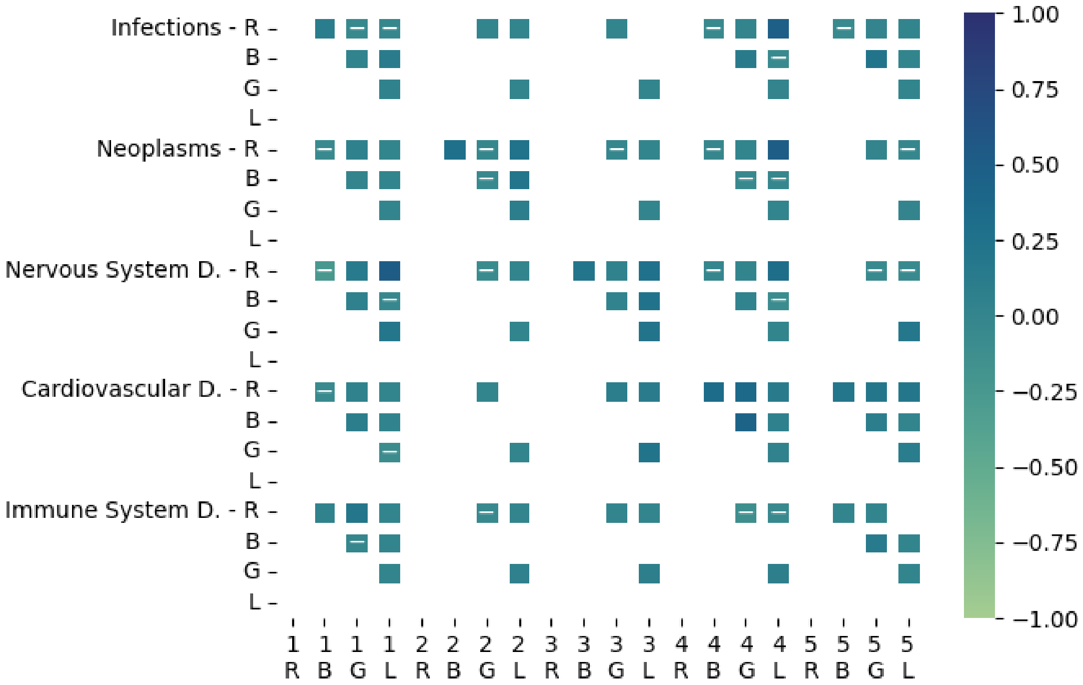

Figure 2: Cohen's kappa scores for each pair of teams. The x-axis shows the five reference article (in the order shown in Table 1) and the teams: (R)CX, (B)f3R, (G)PT-4o-mini, and (L)lama3.3. The y-axis depicts the teams along with the five diseases. Cells with negative scores are depicted with a minus ("-").

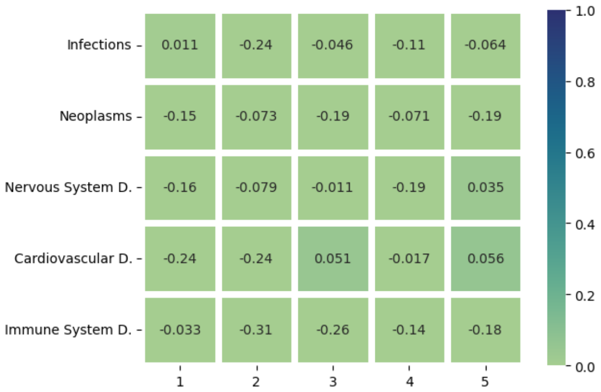

Figure 3: Krippendorff' alpha scores for each reference article across all teams. The x-axis shows the five reference articles (in the order shown in Table 1). The y-axis depicts the five diseases.

is a rather large dataset (around 400 articles), but includes only four reference articles. For all reference articles in batch1, the low kappa and krippendorff's alpha scores showed above indicate that such as dataset should not include all articles from top 20 list. Therefore, we identified the articles (PMIDs) that have high agreement across the teams and that could potentially be included in a dataset. We only considered PMIDs with four equal votes of the same label, i.e., agreement across all teams. We obtained 28 PMIDs from all five diseases with unanimous agreement, 14 "similar" pairs and 14 "not similar" (cf. Table 2).

Pairs of articles with good agreement. Many pairs have a good agreement, even though they have no agreement across the four teams. We identified 55 articles with three unanimous labels, i.e., from the RCX team and the two LLMs. From these, 50 of them were tagged as "similar" and 5 of them as "not similar". These articles come from 11 reference articles, and these are the ones whose annotation from team Bf3R should be prioritized, especially those with already many unanimous labels from the three teams, namely, reference articles "37084732" from "Nervous System Disease" and "32485164" from "Infections". Further, from the 14 reference articles with annotations from the four teams, we identified 77 articles with just one different annotation, e.g., three "similar" annotations and one "not similar". These constitute potential additional 42 "similar" articles and 35 "not similar" articles. A consensus round of annotation could potentially solve these disagreements.

Cases with very low agreement. From the reference articles annotated by all four teams, two of them had no article with an unanimous label,

Diseases	Ref. articles	Articles	Label
Infections	37071015	35605915 36441775	similar
	31689515	28456941 26920550 35798933 26189763	similar
	36159784	34228857	not similar
Neoplasms	34233949	35027827	similar
	33320838	36339405 37376562 35995402 35507699	not similar
	37429473	36740846	not similar
Nervous S.	35709748	25362208	not similar
	37339207	31010153 27174093 27045344 34788059	not similar
Cardiov. S.	37268711	31140393	similar
		36674651 31780864 36990303	not similar
	35917178	23563994	similar
Immune. S.	34503569	35325396 23335001 32693359	similar
	37168850	22673798 25778936	similar

Table 2: Selected unanimous pairs for each disease.

i.e., namely “37749256” of “Nervous System Disease” and “33635944” of “Cardiovascular Diseases”. However, some articles in these reference articles had three votes of the same label (cf. above). Further, the reference article “36179018” from “Immune System Disease” was annotated by three teams and did not obtain any article with unanimous label. Finally, in general, the “uncertain” label had a very low agreement, and no article obtained an unanimous label of this type, not even three unanimous labels (cf. above).

Ranks of the articles. For the articles with full agreement across the four teams (cf. above), we checked whether articles tagged as “similar” were usually on the top of the list, and those tagged as “not similar” were rather at the bottom of the list. For the 14 articles tagged as “similar”, their positions in the list varied from 1 to 14 (average of 5.5). For the 14 articles tagged as “not similar”, their positions in the list varied from 6 to 19 (average of 11.6). On the one hand, and even if the sample is rather small for significant insights, it seems that “similar” articles were actually found in rather higher ranks and “not similar” ones in rather lower ranks. On the other hand, there are some cases of “not similar” ones in the top 10, namely, positions 6, 7, and 9, and “similar” ones below the top 10, namely positions 14 and 15.

6 Conclusion

We proposed the SMAFIRA Shared Task with the aim to collect data for a dataset about the similarity of the research goal between two articles, namely, a reference article and one candidate article from the list of similar articles. We released two batches of references articles: (i) a first one related to five diseases, five reference articles each; (ii) a second one with 16 diseases, also five reference articles each. For any reference article in these batches, we asked the participants to annotate the top 20 similar articles, as available in PubMed. The annotation consisted on assessing the similarity in terms of three labels, namely, “similar”, “uncertain”, and “not similar”.

For the first batch, we collected annotations from two teams that performed manual annotation and two LLMs for automatic annotations. For each reference article, we presented a detailed analysis based on the number of the labels, as well as agreement based on the kappa and the kippendorff’s alpha scores. These scores were very low (often negative) for most reference articles, which means that there is a systematic disagreement across most articles on the top 20.

In spite of the above, there are some articles with high agreement and which could already be part of a dataset. Additionally, some more articles received three equal labels (out of four teams) and could also possibly be included after a consensus round. Finally, some more articles have three unanimous labels and could also potentially be selected after an annotation round from the Bf3R teams. Finally, the RCX team has already agreed to further annotate the second batch, and we could also obtain annotations from the two LLMs, as well as some additional ones from the Bf3R team.

Finally, a preliminary analysis of the ranks of the articles tagged as “similar” or “not similar” confirmed that some articles could have been pushed higher in the top 20 list. Therefore, we need better methods for assessing the similarity of the articles’ research goals. However, a comprehensive dataset is essential for a reliable evaluation of these methods, and for training few-shot approaches.

Acknowledgments

We would like to thank Sebastian Dunst, Daniel Butzke, and Ines Schadock for their feedback on the shared task.

References

Daniel Butzke, Bettina Bert, Konrad Gulich, Gilbert Schönfelder, and Mariana Neves. 2024. SMAFIRA: a literature-based web tool to assist researchers with retrieval of 3R-relevant information. *Laboratory Animals*, 58(4):369–373. PMID: 38872231.

Daniel Butzke, Nadine Dulisch, Sebastian Dunst, Matthias Steinfath, Mariana Neves, Brigitte Mathiak, and Barbara Grune. 2020. Smafira-c: A benchmark text corpus for evaluation of approaches to relevance ranking and knowledge discovery in the biomedical domain.

European Commission, Joint Research Centre, I Adcock, T Novotny, M Nic, K Dibusz, J Hynes, L Marshall, and L Gribaldo. 2020. *Advanced non-animal models in biomedical research : respiratory tract diseases*. Publications Office of the European Union.

Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic Acids Research*, 48(W1):W5–W11.

M. L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Mariana Neves, Antonina Klippert, Fanny Knöspel, Juliane Rudeck, Ailine Stolz, Zofia Ban, Markus Becker, Kai Diederich, Barbara Grune, Pia Kahnau, Nils Ohnesorge, Johannes Pucher, Gilbert Schönfelder, Bettina Bert, and Daniel Butzke. 2023a. Automatic classification of experimental models in biomedical literature to support searching for alternative methods to animal experiments. *Journal of Biomedical Semantics*, under review.

Mariana Neves, Ines Schadock, Beryl Eusemann, Gilbert Schönfelder, Bettina Bert, and Daniel Butzke. 2023b. Is the ranking of PubMed similar articles good enough? an evaluation of text similarity methods for three datasets. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 133–144, Toronto, Canada. Association for Computational Linguistics.

Mary Ann Vasbinder and Paul Locke. 2017. Introduction: Global Laws, Regulations, and Standards for Animals in Research. *ILAR Journal*, 57(3):261–265.

A Evaluation of the LLMs with SMAFIRA-c dataset

We show the statistics of the annotation of the corpus are in Table 3. Further, we evaluated both LLMs in the SMAFIRA-c dataset with the corresponding mapping for the labels (cf. Section 2.4). We show results in Table 4.

ref. PMIDs	similar	uncertain	not similar
16850029	11	14	71
19735549	12	5	81
21494637	5	42	56
24204323	26	0	76

Table 3: Number of annotations in the SMAFIRA-c dataset.

	GPT-4omini			Llama3.3		
16850029	P	R	F1	P	R	F1
similar	0.31	0.90	0.46	0.15	1.00	0.26
not similar	0.84	0.75	0.79	0.89	0.35	0.51
uncertain	0.33	0.07	0.12	0	0	0
overall	0.66	0.66	0.66	0.37	0.37	0.37
19735549	P	R	F1	P	R	F1
similar	0.46	0.55	0.50	0.13	1.00	0.23
not similar	0.87	0.86	0.87	1.00	0.17	0.29
uncertain	0	0	0	0	0	0
overall	0.78	0.78	0.78	0.26	0.26	0.28
21494637	P	R	F1	P	R	F1
similar	0.42	0.56	0.48	0.42	1.00	0.59
not similar	0.58	0.43	0.49	1.00	0.07	0.13
uncertain	0	0	0	0	0	0
overall	0.46	0.46	0.46	0.44	0.44	0.44
24204323	P	R	F1	P	R	F1
similar	0.52	0.48	0.50	0.39	0.92	0.55
not similar	0.84	0.82	0.83	0.95	0.53	0.68
uncertain	0	0	0	0	0	0
overall	0.73	0.73	0.73	0.62	0.62	0.62

Table 4: Evaluation of the LLMs on the SMAFIRA-c dataset.