Overview of the CLEF 2026 SimpleText Task 3: Research Area Classification

Gautam Kishore Shahi[1], Liana Ermakova[2] and Jaap Kamps[3]

[1]University of Duisburg-Essen, Essen, Germany
[2]Université de Bretagne Occidentale, HCTI, France
[3]University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper presents an overview of the CLEF 2026 SimpleText Task 3 on Research Area Classification. The task aims to classification of scientific articles by research area. We discuss the data and benchmarks provided for these tasks, along with preliminary insights and anticipated challenges. Our main findings are the following. First, the best approach for discipline achieves an exact-match accuracy of 65%, and the best approach for field area achieves an exact-match accuracy of 85%. Second, the top systems combine the best of both worlds: they still achieve very high classification accuracy on the most prevalent classes, yet also achieve the highest accuracy on minority classes. Third, compared to human indexing and cataloging, the performance may still be too low for automatic deployment, but it already seems valuable for supporting human indexing efforts. More generally, we hope and expect that the constructed corpora and evaluation data will be used by researchers to further advance text simplification approaches, both in general and specifically for the biomedical domain.

Keywords

Scientific text simplification, Biomedical AI, Generative AI, Information access, Natural language processing

1. Introduction

Becoming scientifically literate is increasingly crucial. Objective scientific information helps people navigate a world in which misinformation, disinformation, and fabricated claims are only a click away. Although the importance of such information is widely recognized, the general public rarely engages with primary scientific sources. This raises a central question: how can we make scientific texts more accessible to everyone?

The impact of objective scientific information is especially clear in biomedicine, where research findings directly influence health decisions. Yet, the most reliable and up-to-date sources use specialized language and assume substantial prior knowledge, making them inaccessible to non-experts. Our track therefore centers on a concrete, societally relevant biomedical use case and treats accessibility as a core objective, distinguishing our work from earlier text simplification approaches that considered lexical and grammatical complexity in isolation.

Scientific text simplification offers a way to make these sources more approachable, and recent advances in large language models have accelerated progress in this area. Nonetheless, key challenges remain: dealing with domain-specific jargon and preserving the accuracy of the original information. Addressing these challenges requires high-quality training data and rigorously designed benchmark datasets, informed by detailed human assessments and systematic analyses of the limitations of current summarization and simplification systems.

To address these challenges, the CLEF 2026 SimpleTrack has three main aims: First, to *advance scientific text simplification* by expanding existing corpora with more varied paragraph- and document-level simplifications that respect complex discourse structure. This new biomedical corpus is built from aligned Cochrane abstracts and plain language summaries [1, 2] and is tailored to models such as LLMs that handle long inputs. Second, to *detect and reduce hallucinations* by exploiting aligned sources,

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
gautam.shahi@uni-due.de (G. K. Shahi); liana.ermakova@univ-brest.fr (L. Ermakova); kamps@uva.nl (J. Kamps)
0000-0001-6168-0132 (G. K. Shahi); 0000-0002-7598-7474 (L. Ermakova); 0000-0002-6614-0087 (J. Kamps)

references, and model outputs to identify and quantify spurious information introduced by generative models. This tackles a key limitation of current systems and evaluation measures, which often fail to penalize unwarranted additional content—an especially critical issue for scientific applications and for NLP/IR more broadly. Third, to *classify scientific text* by predicting the field and the discipline of a scientific paper based on a given taxonomy. This accelerates consistent metadata across collections (arXiv, OpenAIRE, MADOC).

Hence, the CLEF 2025 SimpleText track is based on three interrelated tasks:

- ***Task 1: Text Simplification****: simplify scientific text.*
- ***Task 2: Controlled Creativity****: identify and avoid hallucination.*
- ***Task 3: Research Area Classification****: classification of scientific articles by research area.*

This paper gives an overview of the CLEF 2026 SimpleText Task 3 on Research Area Classification, which aims to classification of scientific articles by research area. Further details on the entire track are provided in the CLEF 2026 SimpleText Track Overview [3]. Additional details on Task 1 on *Text Simplification* are in a companion Task 1 overview paper [4], details on Task 2 on *Controlled Creativity* are in a companion Task 2 overview paper [5], and details on Task 3 on *Research Area Classification* are in a companion Task 3 overview paper [6]. We also refer to the respective participants' papers for further details.

The rest of this paper is structured as follows. Section 2 describes the task, the data, the format, and the evaluation measures. Section 3 describes the participants' approaches. Section 4 provides detailed results for the task. Section 5 provides further analysis of the results. We end with a discussion and conclusions in Section 6.

2. Task 3: Classification of scientific articles by research area

This section details *Task 3: Research Area Classification* on classification of scientific articles by research area.

2.1. Description

The *Research Area Classification* task aims to rerun *classification of scientific articles by research area*. The pilot task proposal is on *Identification of research areas from scientific text*. The task aims to develop a novel approach to classify scientific articles using a given taxonomy. Traditional subject classification in libraries is a solvable problem [7], and can accelerate consistent metadata across collections (arXiv, OpenAIRE, MADOC). The task aims to predict the field and the discipline of a scientific paper.

2.2. Data

For datasets, the dataset is crawled from arXiv. arXiv has provided open access to scholarly articles across disciplines such as physics, computer science, mathematics, and economics. The collected dataset is further annotated by DFG subject classification.[1] The dataset includes the *id, title, abstract, field_area, discipline_area*.

Table 1 shows the train data distribution. We note that the taxonomy used is limited, with only a very few distinct disciplines and research areas. We also note that the sample of bibliographic records to be classified is highly imbalanced, with a clear overrepresentation of the sciences and some specific fields within those. The test data contains 3,107 bibliographic records, numbered `test_1` until `test_3367`.

2.3. Formats

This section outlines the format of the data used in the CLEF 2026 SimpleText Task 3.

[1] https://www.dfg.de/resource/blob/331950/fachsystematik-2024-2028-en.pdf

Table 1
CLEF 2026 SimpleText Task 3 Train Data Distribution

Discipline Area		Field Area	
Descriptor	**Number**	**Descriptor**	**Number**
Natural Sciences	9,409	*Physics*	4,995
		Mathematics	4,001
		Chemistry	240
		Geosciences	173
Engineering Sciences	1,601	*Computer Science*	1,254
		Electrical Engineering and Information Technology	207
		Mechanical and Industrial Engineering	69
		Systems Engineering	50
		Civil Engineering	12
		Materials Science	7
		Process Engineering	2
Life Sciences	907	*Biology*	696
		Medicine	211
		Veterinary Medicine	0
Social Sciences	361	*Social and Behavioural Sciences*	194
		Psychology	91
		Education	76
		Social Sciences	0
Humanities	23	*Linguistics*	15
		Archeology	5
		Literary Studies	3
		Anthropology	0
		Humanities	0

Classification Taxonomy The taxonomy was distributed in cvs with the following format:

```
[
  {
    "Field_Area":"Humanities",
    "Discipline_Area":"Archeology"
  },
  . . .
  {
    "Field_Area":"Social Sciences",
    "Discipline_Area":"Education"
  },
  . . .
  {
    "Field_Area":"Life Sciences",
    "Discipline_Area":"Biology"
  },
  . . .
  {
    "Field_Area":"Natural Sciences",
    "Discipline_Area":"Chemistry"
  },
  . . .
  {
    "Field_Area":"Engineering Sciences",
    "Discipline_Area":"Mechanical and Industrial Engineering"
  },
  . . .
]
```

Train Data The train data was distributed in cvs with the following format:

```
[
  {
    "id":"train_1",
    "title":"Finding Complex Biological Relationships in Recent PubMed Articles Using\n  Bio-LDA",
    "abstract":"  The overwhelming amount of available scholarly literature in the life\nsciences
    ↪  poses significant challenges to scientists wishing to keep up with\nimportant developments
    ↪  related to their research, but also provides a useful\nresource for the discovery of recent
    ↪  information concerning genes, diseases,\ncompounds and the interactions between them. In this
    ↪  paper, we describe an\nalgorithm called Bio-LDA that uses extracted biological terminology
    ↪  to\nautomatically identify latent topics, and provides a variety of measures to\nuncover
    ↪  putative relations among topics and bio-terms. Relationships identified\nusing those
    ↪  approaches are combined with existing data in life science datasets\nto provide additional
    ↪  insight. Three case studies demonstrate the utility of\nthe Bio-LDA model, including
    ↪  association predication, association search and\nconnectivity map generation. This combined
    ↪  approach offers new opportunities\nfor knowledge discovery in many areas of biology including
    ↪  target\nidentification, lead hopping and drug repurposing.\n",
    "field_area":"Life Sciences",
    "discipline_area":"Biology"
  },
  . . .
]
```


Test Data The test data was distributed in cvs with the following format:


```
  {
    "id":"test_1",
    "title":"Unified algebraic Bethe ansatz for two-dimensional lattice models",
    "abstract":"  We develop a unified formulation of the quantum inverse scattering method
    ↪  for\nlattice vertex models associated to the non-exceptional
    ↪  $A^{(2)}_{2r}$,\n$A^{(2)}_{2r-1}$, $B^{(1)}_r$, $C^{(1)}_r$, $D^{(1)}_{r+1}$ and
    ↪  $D^{(2)}_{r+1}$\nLie algebras. We recast the Yang-Baxter algebra in terms of novel
    ↪  commutation\nrelations between creation, annihilation and diagonal fields. The solution
    ↪  of\nthe $D^{(2)}_{r+1}$ model is based on an interesting sixteen-vertex model which\nis
    ↪  solvable without recourse to a Bethe ansatz.\n"
  },
  . . .
]
```


2.4. Evaluation

LLMs are well known for somewhat "creative" answers and not complying with all specifications under all circumstances. Moreover, even with good specifications, LLMs might produce some "near misses" with their results when, for instance, a research area like "Computer Science & Engineering" might be predicted as just "Computer Science". To overcome these challenges, we used several approaches to determine the actual matches:

- *Exact Match (EM)*: a binary measure comparing the output of the model directly with the ground truth.
- *String Distance (SD)*: a normalized Levenshtein distance between the prediction and reference.
- *Embedding Distance (ED)*: a semantic similarity measure based on the BERT-embeddings of the prediction and reference.

The evaluation will focus on accuracy as the primary performance metric, as recommended by Chang et al. [8]. In addition to EM accuracy, we can apply a threshold to convert SD and EM scores to Boolean values, and initial experiments suggest a 70% threshold [9]. Overall, ranking is based on the macro-average of all classes. exact matches between the field and discipline, and an entry matches the test data when both fields and disciplines match. The primary ranking is done using Exact Match.

3. Participant's Approaches

A total of 7 teams submitted 12 runs in total. In the detailed results, we include only runs without errors that achieve a non-zero score.

AIIRLab Hawkins et al. [10] submitted 2 runs in total for Task 3. They use ensemble and multi-agent pipelines. For research area classification, they use a fine-tuned SPECTER2 classifier that outperforms a generative LLaMA baseline. Overall, the results demonstrate the effectiveness of ensemble learning and domain-specific language models across all three tasks.

DataDumplings John et al. [11] submitted 1 run in total for Task 3. Their approach uses a two-stage hierarchical classification framework with fine-tuned SciBERT models to classify scientific papers into field and discipline categories using titles and abstracts. The first model predicts the broad field, while the second predicts the discipline, with taxonomy-based constraints ensuring consistency between the two levels. The system is evaluated using Exact Match, Levenshtein String Distance, and Embedding Distance metrics, demonstrating that combining domain-specific language models with hierarchical inference improves classification validity and interpretability.

SRH-ReliableAI Priyanshu and Chandna [12] submitted 2 runs in total for Task 3. They use a hierarchy-constrained multi-task ensemble for research area classification using DeBERTa-v3-large and SciBERT. The models are fine-tuned with stratified cross-validation, adversarial training, multi-sample dropout, and confidence-based pseudo-labeling to improve robustness under severe class imbalance. During inference, ensemble predictions are averaged and constrained by the taxonomy to ensure consistent field–discipline assignments, resulting in improved classification performance and reliable hierarchical predictions.

StuttgartAI Kumari [13] submitted 1 run in total for Task 3.

UBCS Anderson and Thuma [14] submitted 3 runs in total for Task 3. They use a hierarchical classification pipeline based on domain-specific transformer models, SciBERT and SPECTER, for research area classification. To improve performance on imbalanced datasets, it combines rule-based and T5-based data augmentation with oversampling techniques. A hierarchical constraint ensures that predicted broad fields restrict the possible discipline labels, while a Random Forest meta-classifier dynamically selects the better prediction between SciBERT and SPECTER for each document.

UvA Mathas et al. [15] submitted 2 runs in total for Task 3. They submitted a zero-shot LLM baseline using Gemini 3.5 Flash. The zero-shot model classifies each paper independently based on its title, abstract, and the complete research taxonomy, providing a stronger baseline without task-specific fine-tuning.

Writerslogic Condrey [16] submitted 1 run in total for Task 3. This was just a trial run, with no further details in the paper.

4. Results

This section details the task results for the discipline and research area prediction subtasks.

Table 2 shows the main evaluation results of the submissions to SimpleText Task 3. Our main findings are the following. First, due to the late availability of the data, only six teams could participate in the research area classification task. Each team used advanced classifiers that significantly outperformed baselines such as the majority class. Second, the best approach for discipline achieves an exact-match accuracy of 65%, and the best approach for field area achieves an exact-match accuracy of 85%. These

Table 2
Results of SimpleText Task 3: classification of scientific articles by research area. Sorted on exact match of the field area.

Team Name	Discipline Area			Field Area		
	EM	SD	ED	EM	SD	ED
Data Dumplings	0.5983	0.6909	0.7836	0.8448	0.9080	0.9242
UBCS_2	0.5915	0.6877	0.7738	0.7684	0.8538	0.8912
UBCS_1	0.5518	0.6483	0.7533	0.7573	0.8481	0.8804
SRH-Reliable AI	0.6056	0.7003	0.7780	0.7390	0.8284	0.8689
AIIRLab_2	0.6539	0.7356	0.8068	0.7353	0.8270	0.8695
AIIRLab_1	0.5838	0.6746	0.7696	0.7288	0.8244	0.8698
UAmsterdam_2	0.6122	0.7035	0.7759	0.7033	0.8279	0.8472
Writerslogic Inc	0.4439	0.5676	0.6571	0.5561	0.7471	0.7924
UAmsterdam_1	0.0625	0.2281	0.4099	0.2000	0.6225	0.6659

Table 3
Analysis of SimpleText Task 3: classification of scientific articles by research area. Sorted on macro average of the discipline classification.

	Team Name	Natural Sciences	Life Sciences	Engineering Sciences	Social Sciences	Humanities	Macro Average
Exact Match	Data Dumplings	0.91	0.75	0.79	0.84	0.60	0.78
	Stuttgart AI	0.92	0.78	0.85	0.76	0.30	0.72
	UBCS_3	0.92	0.76	0.86	0.80	0.20	0.71
	UBCS_2	0.92	0.75	0.82	0.79	0.20	0.69
	UBCS_1	0.92	0.75	0.83	0.73	0.20	0.69
	SRH Reliable AI_2	0.92	0.81	0.86	0.82	0.00	0.68
	SRH-Reliable AI_1	0.92	0.81	0.86	0.82	0.00	0.68
	AIIRLab_2	0.91	0.82	0.87	0.75	0.00	0.67
	AIIRLab_1	0.92	0.80	0.78	0.82	0.00	0.66
	UAmsterdam_1	0.81	0.54	0.86	0.48	0.40	0.62
	Writerslogic Inc	0.79	0.32	0.11	0.62	0.00	0.37
	UAmsterdam_2	0.54	0.00	0.00	0.00	0.00	0.11
String match	Data Dumplings	0.91	0.75	0.79	0.84	0.60	0.78
	Stuttgart AI	0.92	0.78	0.85	0.76	0.30	0.72
	UBCS_3	0.92	0.76	0.86	0.80	0.20	0.71
	UBCS_2	0.92	0.75	0.82	0.79	0.20	0.69
	UBCS_1	0.92	0.75	0.83	0.73	0.20	0.69
	SRH Reliable AI_2	0.92	0.81	0.86	0.82	0.00	0.68
	SRH-Reliable AI	0.92	0.81	0.86	0.82	0.00	0.68
	AIIRLab_2	0.91	0.82	0.87	0.75	0.00	0.67
	AIIRLab_1	0.92	0.80	0.78	0.82	0.00	0.66
	UAmsterdam_1	0.81	0.54	0.86	0.48	0.40	0.62
	Writerslogic Inc	0.79	0.32	0.11	0.62	0.00	0.37
	UAmsterdam_2	0.54	0.00	0.00	0.00	0.00	0.11

are impressive scores even when factoring in that only a small subset of the DFG taxonomy was used. Third, more generally, compared to human indexing and cataloging, the performance may still be too low for automatic deployment, but it already seems valuable for supporting human indexing efforts.

5. Analysis

We conduct further analysis by looking at the classification accuracy for each of the five disciplines.

Table 3 shows the classification accuracy of the submissions to SimpleText Task 3. Our main findings are the following. First, we see that class prevalence influences many of the trained classifiers. The majority-class baseline for `UAmsterdam_2` shows that the *Natural Sciences* also dominate the test set.

Second, we observe that untrained, zero-shot approaches achieve solid performance across all classes, making them more competitive in the macro-average score. Third, the top systems are able to combine the best of both worlds: they still have very high classification accuracy on the most prevalent classes, yet also obtain the highest classification accuracy on the minority classes.

6. Discussion and Conclusions

This concludes the results for the CLEF 2026 SimpleText Task 3: Research Area Classification on classification of scientific articles by research area. We are encouraged by the submission of four additional papers by participants from teams participating only in this task.

Overall, the participating systems predominantly rely on domain-specific transformer models (e.g., SciBERT, SPECTER, DeBERTa) combined with hierarchical classification frameworks that exploit the taxonomy of research fields and disciplines. Several approaches further improve performance through ensemble learning, data augmentation, adversarial training, pseudo-labeling, and taxonomy-constrained inference, while one system explores zero-shot classification using a large language model without task-specific fine-tuning.

This Task Overview focuses on the CLEF 2026 SimpleText Track's Task 3 on research area classification. Within the CLEF 2026 SimpleText Task 3, we have constructed extensive corpora and new references for evaluation data. These reusable corpora and evaluation resources are available to participants and other researchers who want to work on the important problem of making scientific information open and easily accessible for everyone.

Acknowledgments

We are incredibly thankful to the master's students in translation and technical writing from the University of Brest for participating in data annotation. We also thank each of the individual track participants for their effort in submitting a record number of submissions to CodaBench and documenting these in their papers.

We thank the CLEF 2026 chairs for hosting us, and the CLEF 2026 Labs and Proceedings chairs for their excellent assistance and flexibility. It is heartwarming to be part of such a great CLEF family. We thank CodaBench [17] for hosting the competition. Post-competition experiments are ongoing at https://www.codabench.org/competitions/16108/ (Task 1.1, Task 1.2, and Task 2.3) and https://www.codabench.org/competitions/16065/ (Task 2.1 and Task 2.2). We hope and expect that these "living test collections" remain in active use until the next iteration of the track.

Liana Ermakova is partly funded by the French National Research Agency (ANR-22-CE23-0019-01, *SimpleText: Automatic Simplification of Scientific Texts*). Liana Ermakova is further supported by the CNRS research group MaDICS (https://www.madics.fr/ateliers/simpletext/).

Jaap Kamps is funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.

[2] J. Bakker, J. Kamps, Section-level simplification of biomedical abstracts, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 13819–13833. URL: https://aclanthology.org/2025.emnlp-main.697/. doi:10.18653/v1/2025.emnlp-main.697.

[3] L. Ermakova, H. Azarbonyad, J. Bakker, B. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: [18], 2026.

[4] J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify Scientific Text, in: [19], 2026.

[5] B. Chakar, J. Bakker, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 2: Identify and Avoid Hallucination, in: [19], 2026.

[6] G. K. Shahi, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 3: Research Area Classification, in: [19], 2026.

[7] G. K. Shahi, O. Hummel, On the effectiveness of large language models in automating categorization of scientific texts, in: J. Filipe, M. Smialek, A. Brodsky, S. Hammoudi (Eds.), Proceedings of the 27th International Conference on Enterprise Information Systems, ICEIS 2025, Porto, Portugal, April 4-6, 2025, Volume 1, SCITEPRESS, 2025, pp. 544–554. URL: https://doi.org/10.5220/0013299100003929. doi:10.5220/0013299100003929.

[8] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.

[9] G. K. Shahi, O. Hummel, Automating categorization of scientific texts with in-context learning and prompt-chaining in large language models, arXiv preprint arXiv:2604.23430 (2026).

[10] E. Hawkins, K. Zbinden, E. Fitzgerald, B. Mansouri, AIIRLab Systems at SimpleText CLEF 2026: Ensemble and Multi-Agent Pipelines for Scientific Text Simplification, in: [19], 2026.

[11] S. N. John, S. Rajagopal, Y. Venkatesh, Hierarchical Scientific Domain Classification with Fine-Tuned SciBERT and Taxonomy-Guided Inference, in: [19], 2026.

[12] P. Priyanshu, S. Chandna, SRH-ReliableAI at SimpleText 2026 Task 3: A Hierarchy-Constrained, Adversarially Trained Multi-Task Ensemble for Scientific Research Area Classification, in: [19], 2026.

[13] S. Kumari, Classification of Scientific Text, in: [19], 2026.

[14] G. Anderson, E. Thuma, Improving Scientific Research Area Classification through Transformer Fine-Tuning and Per-Document Model Selection, in: [19], 2026.

[15] P. Mathas, B. Chakar, D. Carranza Navarrete, J. Bakker, J. Kamps, University of Amsterdam at the CLEF 2026 SimpleText Track, in: [19], 2026.

[16] D. Condrey, Writerslogic at the CLEF 2026 SimpleText Track: Multi-Candidate LLM Simplification and Stacked Complexity Spotting, in: [19], 2026.

[17] Z. Xu, S. Escalera, A. Pavão, M. Richard, W. Tu, Q. Yao, H. Zhao, I. Guyon, Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform, Patterns 3 (2022) 100543. URL: https://doi.org/10.1016/j.patter.2022.100543. doi:10.1016/J.PATTER.2022.100543.

[18] M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[19] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.),

Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.