Biomedical Text Simplification Models Trained on Aligned Abstracts and Lay Summaries

Jan Bakker and **Taiki Papandreou-Lazos** and **Jaap Kamps**
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
j.bakker@uva.nl, taiki.papandreou-lazos@student.uva.nl, kamps@uva.nl

Abstract

This paper documents the University of Amsterdam's participation in the TREC 2024 *Plain Language Adaptation of Biomedical Abstracts* (PLABA) Track. We investigated the effectiveness of text simplification models trained on aligned pairs of sentences in biomedical abstracts and plain language summaries. We participated in Task 2 on *Complete Abstract Adaptation* and conducted post-submission experiments in Task 1 on *Term Replacement*. Our main findings are the following. First, we used text simplification models trained on aligned real-world scientific abstracts and plain language summaries. We observed better performance for the context-aware model relative to the sentence-level model. Second, our experiments show the value of training on external corpora and demonstrate very reasonable out-of-domain performance on the PLABA data. Third, more generally, our models are conservative and cautious in gratuitous edits or information insertions. This approach ensures the fidelity of the generated output and limits the risk of overgeneration or hallucination.

1 Introduction

This paper reports on the University of Amsterdam's participation in the TREC 2024 PLABA track (Ondov et al., 2025). The track aims to adapt biomedical abstracts for the general public using plain language. We participated in Task 2 on *Complete Abstract Adaptation*. The PLABA track addresses the key text simplification challenges in the biomedical domain (Ondov et al., 2022). Extensive corpora and reference simplifications were made available as train data for the track (Attal et al., 2023).

Our main aim at TREC 2024 PLABA was to investigate the value of domain-specific biomedical test simplification models relative to those trained on general text simplification corpora such as Wiki-auto or Newsela-auto (Jiang et al., 2020). For this purpose, we constructed a similarly aligned corpus Cochrane-auto (Bakker and Kamps, 2024), based on document-, paragraph-, and sentence-level alignments of Cochrane abstracts and their corresponding plain language summaries.

The plain language summaries are no direct sentence-level text simplification as provided in the PLABA track's training and text data. Instead, they exhibit far more variation and complex information alignment, considering the entire abstract's discourse structure. It is interesting to investigate how models trained on such data perform on the PLABA *Complete Abstract Adaptation* task, containing only direct simplifications from the original biomedical abstract.

Our experiments are not trained or fine-tuned on the PLABA train data and, hence, are not expected to outperform models specifically trained for the direct text simplification task of the PLABA track. Instead, we hope to understand some differences with models trained on real-world plain language summaries, exhibiting more significant variation and incorporating the discourse structure of the entire abstract. We hope to extend further the scope of text simplification approaches to address all the interesting NLP challenges this presents.

The rest of this paper is structured in the following way. Our experimental setup is described in Section 2 and the results of these experiments are in Section 3. Finally, we end in Section 4 with a discussion of our main findings.

2 Experimental Design

Complete Abstract Adaptation We participated in Task 2 of the PLABA track and

Task	Run	Description
1a	RoBERTa-MedReadMe	Binary span classification model trained on MedReadMe
1a	RoBERTa-PLABA	Binary span classification model trained on PLABA
1a	RoBERTa-MedReadMe+PLABA	MedReadMe model finetuned on PLABA
1b	RoBERTa-PLABA	RoBERTa-large multi- label classifier with abstract text and target jargon term as input
2	UAms-BART-Cochrane	Plan guided BART model instructed to *rephrase* every source sentence
2	UAms-ConBART-Cochrane	Plan guided context-aware BART model instructed to *rephrase* every source sentence

Table 1: TREC 2024 PLABA Task 2 Submissions

submitted two runs shown in Table 1.

Our runs were created by applying two plan-guided simplification models to the test data: $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ and $\hat{O} \rightarrow \text{ConBART}$. Both runs are BART models, which we finetuned to pairs of complex and simple sentences in the biomedical domain. Following the approach of Cripwell et al. (2023b), we prepended a control token to each input sentence in the training data, indicating how it should be simplified: should it be split, copied, rephrased, deleted, or merged? For inference, we then prepended the *rephrase* token to every input sentence in the PLABA test set. Moreover, $\hat{O} \rightarrow \text{ConBART}$ is a context-aware model, as introduced by Cripwell et al. (2023a). This means that it takes not only a control token and complex sentence as input, like $\hat{O} \rightarrow \text{BART}_{\text{sent}}$, but also a high-level representation of the five sentences surrounding the complex input sentence on either side.

Both of our models were trained on Cochrane-auto (Bakker and Kamps, 2024). We created this dataset by automatically aligning the sentences between technical abstracts and lay summaries of biomedical systematic reviews. Our data is derived from the Cochrane Database of Systematic Reviews. We only used the abstract and summary texts from the results and conclusions sections, and we applied various other preprocessing steps to enable the training of document simplification systems on our dataset. We did not further fine-tune our models on the PLABA train set (Attal et al., 2023).

Term Replacement In addition, we conducted post-submission experiments in Task 1 of the PLABA track, and our four runs are also shown in Table 1. We implemented a jargon identification approach using the RoBERTa-large model (Liu et al., 2019).[1]

Task 1a asks for *identifying non-consumer terms*. Our Task 1a models experimented with out-of-domain evaluation of models trained on MedReadMe (Jiang and Xu, 2024). This model was selected based on both its strong performance reported in the original study and our own reproduction results.[2] The RoBERTa-large model trained on MedReadMe data achieved F1 scores of 91.3% for binary classification, 89.0% for 3-class classification, and 83.7% for 7-class classification. We employed RobertaTokenizerFast for tokenization due to its superior processing speed.

To comprehensively evaluate different training strategies, we implemented and compared three model variants:

- *RoBERTa (MedReadMe)* is a RoBERTa-large model trained on the MedReadMe data, not informed by the PLABA data in any way (out of domain evaluation).
- *RoBERTa (PLABA)* is a RoBERTa-large model trained only on the PLABA data (within domain evaluation).
- *RoBERTa (MedReadMe + PLABA)* is the first MedReadMe model further fine-tuned on the PLABA data.

All models were trained for up to 20 epochs with early stopping.

[1] https://huggingface.co/docs/transformers/model_doc/roberta

[2] The codebase can be found in https://github.com/TaikiLazos/reproducibility.

Model	Prec.	Recall	F1
RoBERTa (M)	0.3953	0.2453	0.3027
RoBERTa (P)	0.3998	0.2583	0.3128
RoBERTa (M + P)	0.4504	0.3001	0.3679

Table 2: Performance on PLABA Task 1A

Task 1b asks for *classifying term replacements*. Our Task 1b model, *RoBERTa (PLABA)*, is a multi-label classification version of the RoBERTa-large model trained on the PLABA data. Our model takes both the abstract text and the target jargon term as inputs. The model was trained to classify multiple simplification actions per term, ensuring flexibility in handling different linguistic modifications.

3 Experimental Results

This section contains the main results of our experiments.

3.1 Term Replacement Task

Identifying non-consumer terms Figure 1 illustrates an abstract Q19_A1 from the PLABA test dataset, highlighting the overlap between true jargon annotations (green), predicted jargon terms (yellow), and their overlap (orange). The predictions are from the *RoBERTa (MedReadMe)* model.

Table 2 presents the performance of each model. The best-performing model was finetuned on MedReadMe and PLABA, achieving the highest F1 score. This indicates that leveraging external domain knowledge from MedReadMe enhances jargon identification performance. The relatively low recall scores across all models suggest that further optimization is needed to improve the coverage of identified jargon terms. Analysis of the training and validation loss suggests that the model overfitted when trained on PLABA. Hence, a more careful training regime if needed for PLABA than what we used for MedReadMe.

Classifying replacement For task 1b, we have only one model *RoBERTa (PLABA)*. On the test set, the model achieved a weighted average F1 score across all labels of 54% and a sample-average F1 score of 71%, consistent with results obtained in previous evaluations.

Table 3 presents the detailed classification performance for different simplification actions. We observe an imbalanced class distribution and the model naturally prioritizes the classes with the largest support.

Class	Prec.	Recall	F1	Sup.
SUBSTITUTE	0.85	0.95	0.89	6,484
EXPLAIN	0.49	0.48	0.49	1,822
GENERALIZE	0.27	0.36	0.31	790
OMIT	0.39	0.18	0.25	1,660
EXEMPLIFY	0.03	0.07	0.04	58
micro avg	0.69	0.70	0.70	10,814
macro avg	0.41	0.41	0.40	10,814
weighted avg	0.67	0.70	0.68	10,814
samples avg	0.75	0.76	0.73	10,814

Table 3: Detailed classification performance for different simplification actions.

3.2 Complete Abstract Adaptation

The remainder of this section focuses on our main participation in Task 2 on *Complete Abstract Adapation*.

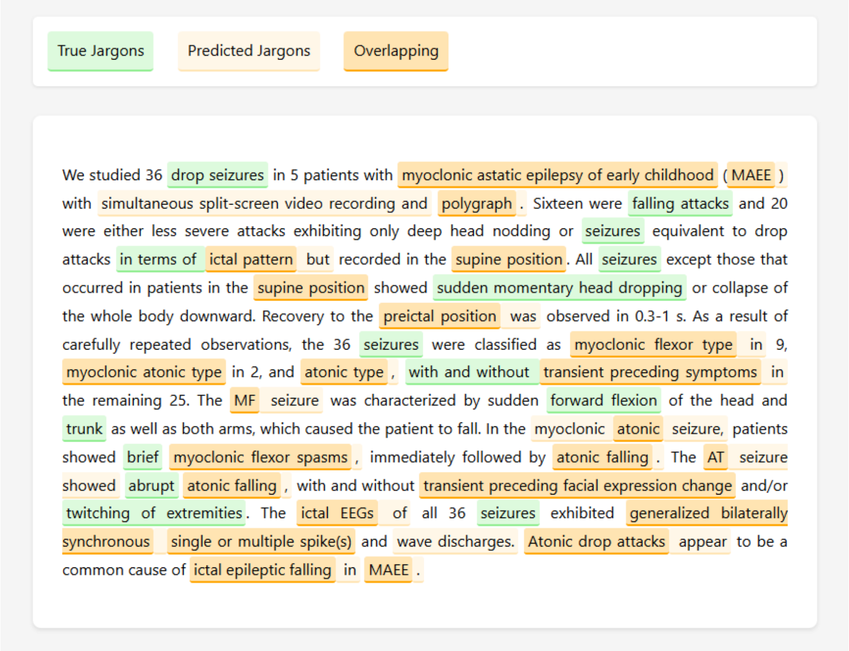
3.2.1 Predictions

Figures 2 and 3 show the source and predictions for one example abstract for the BART and ConBART models, respectively. These are typical examples demonstrating relatively conservative editing, particularly the longer sentences. Both models have a clear preference for narrative text flow and tend to remove detail in parenthesis or headings, breaking the flow of text. The Contextual BART models aware of the context of the entire abstract seems slightly less conservative than the pure sentence-level text simplifications.

3.2.2 Analysis

Table 4 provides various text statistics comparing the source and predictions. There are no references; hence, we calculate BLEU and SARI relative to the source input. The input is relatively long. With the standard EASSE tokenization, it averages 22.15 words per sentence and 224.81 words per abstract.

As our planning model used the “rephrase” instruction throughout, we see no sentence splits or deletions. We see a relatively modest compression ratio of around 90% and relatively light editing with 44-51% of exact sentence copies.



We studied 36 drop seizures in 5 patients with myoclonic astatic epilepsy of early childhood (MAEE) with simultaneous split-screen video recording and polygraph . Sixteen were falling attacks and 20 were either less severe attacks exhibiting only deep head nodding or seizures equivalent to drop attacks in terms of ictal pattern but recorded in the supine position. All seizures except those that occurred in patients in the supine position showed sudden momentary head dropping or collapse of the whole body downward. Recovery to the preictal position was observed in 0.3-1 s. As a result of carefully repeated observations, the 36 seizures were classified as myoclonic flexor type in 9, myoclonic atonic type in 2, and atonic type , with and without transient preceding symptoms in the remaining 25. The MF seizure was characterized by sudden forward flexion of the head and trunk as well as both arms, which caused the patient to fall. In the myoclonic atonic seizure, patients showed brief myoclonic flexor spasms , immediately followed by atonic falling . The AT seizure showed abrupt atonic falling , with and without transient preceding facial expression change and/or twitching of extremities. The ictal EEGs of all 36 seizures exhibited generalized bilaterally synchronous single or multiple spike(s) and wave discharges. Atonic drop attacks appear to be a common cause of ictal epileptic falling in MAEE .

Figure 1: Example *RoBERTa (PLABA)* predictions for the complex abstract A1 for Q19 (PMID 1396420).

run_id	count	BLEU	SARI	FKGL	Compression ratio	Sentence splits	Levenshtein similarity	Exact copies	Additions proportion	Deletions proportion	Lexical complexity score
Source	4,060	100.0	33.33	13.52	1.00	1.00	1.00	1.00	0.00	0.00	8.93
UAms-BART-Cochrane	4,060	83.1	30.26	13.30	0.93	0.99	0.94	0.51	0.02	0.10	8.89
UAms-ConBART-Cochrane	4,060	79.63	29.58	13.07	0.91	0.99	0.92	0.44	0.03	0.12	8.86
Source	400	100.0	33.33	13.53	1.00	1.00	1.00	1.00	0.00	0.00	9.13
UAms-BART-Cochrane	400	82.01	30.16	13.32	0.91	0.99	0.94	0.03	0.01	0.12	9.11
UAms-ConBART-Cochrane	400	78.42	29.47	13.07	0.89	0.98	0.92	0.01	0.02	0.14	9.08

Table 4: Text statistics over source and predictions: sentence level (top) and abstract-level (bottom) over the test set, with BLEU and SARI against source.

At the abstract level, only 1-3% is unchanged. In terms of readability scores, there is a limited reduction of FKGL from 13.5 for the source to 13.1 for the predictions. This is a direct consequence of the lack of sentence splits, as we did not instruct the model to split sentences. When abstracts are compared to human plain language summaries, a relatively low difference in sentence length and, therefore, in classic readability scores is also observed.

3.2.3 Results

Table 5 analyses our submissions in terms of the quality of the text simplification predictions using manual annotation of the output. The following four aspects were assessed (Ondov et al., 2025):

Simplicity Outputs should be easy to under-

Run	Accuracy	Completeness	Simplicity	Brevity	Final avg.
UAms-BART-Cochrane	0.9772	0.9466	0.4881	0.6906	0.7756
UAms-ConBART-Cochrane	0.9546	0.9238	0.5653	0.6781	0.7804

Table 5: TREC 2024 PLABA Task 2 Results (v.3)

~~Background :~~ There is conflicting evidence about the relationship between vitamin D deficiency and depression , and a systematic assessment of the literature has not been available . |Aims : To determine the relationship , if any , between vitamin D deficiency and depression . |~~Method :~~ A systematic review and meta-analysis of observational studies and randomised controlled trials was conducted . |~~Results :~~ One case-control study , ten ~~cross-sectional~~ studies and three cohort studies with a total of 31 424 participants were analysed . |~~Lower vitamin D levels were found in people with depression compared with controls (SMD = 0.60 , 95 % CI 0.23-0.97) and there~~ There was an increased odds ~~ratio~~ of depression for the lowest v. highest vitamin D categories in the cross-sectional studies ~~(OR = 1.31 , 95 % CI 1.0-1.71)~~ . |The cohort studies showed a significantly increased hazard ratio of depression for the lowest v. highest vitamin D categories ~~(HR = 2.21 , 95 % CI 1.40-3.49)~~ . |Conclusions : Our analyses are consistent with the hypothesis that low vitamin D concentration is associated with depression , and highlight the need for randomised controlled trials of vitamin D for the prevention and treatment of depression ~~to determine whether this association is causal~~ . |

Figure 2: Example $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ simplifications for the complex abstract input A7 for Q17 (PMID 33924215): ~~deletions~~ and insertions

~~Background :~~ There is conflicting evidence about the relationship between vitamin D deficiency and depression , and a systematic assessment of the literature has not been available . |Aims : To determine the relationship ~~, if any ,~~ between vitamin D deficiency and depression . |~~Method :~~ A systematic review and meta-analysis of ~~observational~~ studies and randomised controlled trials was conducted . |~~Results :~~ One case-control study , ten ~~cross-sectional~~ studies and three cohort studies with a total of 31 424 participants were analysed . |~~Lower vitamin D levels were found in people with depression compared with controls (SMD = 0.60 , 95 % CI 0.23-0.97) and there~~ There was an increased ~~odds ratio~~ risk of depression for the lowest v. highest vitamin D categories in the cross-sectional studies ~~(OR = 1.31 , 95 % CI 1.0-1.71)~~ . |The ~~cohort~~ studies showed a significantly increased hazard ratio of depression for the lowest v. highest vitamin D categories ~~(HR = 2.21 , 95 % CI 1.40-3.49)~~ . |~~Conclusions : Our analyses are consistent with the hypothesis~~ We conclude that ~~low vitamin D concentration~~ there is ~~associated with depression , and highlight the~~ a need for randomised controlled trials of vitamin D for the prevention and treatment of depression to determine whether this association is causal . |

Figure 3: Example $\hat{O} \rightarrow$ ConBART simplifications for the complex abstract input A7 for Q17 (PMID 33924215): ~~deletions~~ and insertions

stand.

Accuracy Outputs should contain accurate information.

Completeness Outputs should seek to minimize information lost from the original text.

Brevity Outputs should be concise.

We make the following observations. First, the approach scores high on accuracy and completeness but less on simplicity and brevity. Over the 19 judged runs, the average score overall is ranked 12th, but the accuracy and completeness are ranked 1st. Second, the scores signal that our model is conservative in rewriting, prioritizing accuracy and completeness over more drastic text rephrasing. Third, the contextual model obtains slightly higher simplicity scores and outperforms the sentence-level model in terms of the final overall score. Fourth, an informal output inspection confirms the conservative revisions and suggests that the abstract context can benefit simplicity.

3.3 PLABA Train Data Results

As the PLABA tasks did not provide reference simplifications, we also evaluated our models over the train data. Our models are trained on the Cochrane-auto corpus (Bakker and Kamps, 2024). Hence, we report results over the entire set of train data released in the track rather than the separate test split of Attal et al. (2023).

Table 6 shows the familiar text simplification evaluation measures against references. All metrics can account for multiple possible references. We only report scores at the document or abstract level but observed similar performance for sentence-level text simplification.

We observe a limited reduction in readability (FKGL) and a mild reduction in the num-

Run	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	FKGL	SARI	Avg. #Tokens
Source	63.3	40.7	55.7	35.2	14.14	17.8	343.2
BART	62.3	39.2	54.2	32.8	13.49	28.8	296.5
ConBART	62.0	38.6	53.5	32.1	13.22	30.5	289.2

Table 6: TREC 2024 PLABA Train Data Results

ber of tokens compared to the source abstracts (about 15% shorter). The main advantage of the train data is that it has reference simplifications. Here, we observe reasonable SARI scores of 31% and reasonable text overlap with the references (ROUGE-L over 50% and BLEU around 33%).

Our models are tested out-of-domain and trained on real-world plain language summaries without direct sentence-level language adaptations following precise instructions. Hence, performance under the PLABA track's conditions could be improved. However, the models trained on Cochrane data perform well, which suggests that they are widely applicable to the biomedical and health domains.

4 Discussion and Conclusions

This paper documented our participation in the TREC 2024 PLABA Track (Ondov et al., 2025), focusing on the effectiveness of models trained on Cochrane-auto (Bakker and Kamps, 2024). Cochrane-auto consists of aligned pairs of sentences in Cochrane abstracts and plain language summaries.

Our models exhibited conservative revisions of the abstracts and were not among the highest-performing systems regarding prediction simplicity. This indicates that there may be remaining barriers to lay or consumer access to biomedical abstracts, and further simplification steps (such as detecting and explaining jargon, as aimed for by Task 1) may be beneficial. However, conservative models also avoid gratuitous changes and are less likely to introduce information not warranted by the source abstract (informally called “hallucination”). Thus, a more controlled and conservative simplification approach remains attractive for deploying these models in realistic applications.

Our experiments highlight differences between the classic text simplification corpora and naturally occurring plain language summaries. Traditional text simplification is based on direct sentence-level text simplification and prioritizing lexical and grammatical text simplification. However, naturally occurring plain language summaries aren't directly aligned per sentence content but exhibit far more significant variation, considering the entire abstract's discourse structure. This includes more complex relations between sentences at the abstract level, including deletions, order changes, sentence merging, and inserting background knowledge or other explanations. The PLABA track data is an instrumental dataset that, on the one hand, corresponds to the sentence-level text simplification setup of earlier collections yet already includes some of the more significant variation observed in real-world plain language summaries. This is essential to developing effective paragraph-level and document-level text simplification approaches.

Acknowledgments

The experiments in this paper were carried out on the National Supercomputer Snellius, which was supported by SURF and the University of Amsterdam's HPC Board. Jan Bakker is funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, 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.

References

Kush Attal, Brian D. Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(8).

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17293–17319, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Brian Ondov, Bill Xia, Ishita Unde, Hoa T. Dang, and Dina Demner-Fushman. 2025. Overview of the TREC 2024 PLABA track. In *Proceedings of the Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*. NIST Publications.

Brian D. Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *J. Am. Medical Informatics Assoc.*, 29(11):1976–1988.

A Human Evaluation

Rather than providing human reference simplifications, the main evaluation is based on a human assessment of four aspects for each submission sentence. Tables 7 and 8 show the labels for the example abstracts shown before for the $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ and $\hat{O} \rightarrow \text{ConBART}$ models.

The per-sentence labels confirm the model is conservative and prefers cautious edits over more risky generative text insertions. For the use-case of consumer adaptations of authoritative health information, it is desirable to err on the side of caution to ensure the fidelity of the information provided to consumers.

The human labels on the simplicity aspect are a valuable indicator of laypersons' or consumers' readability of the remaining text. However, this also shows the complexity of consistently judging these aspects (similar or near-similar text received different labels), and the direct interpretation of these labels is non-trivial.

It also remains unclear how predictions could have been improved as we have no reference examples of gold standard simplifications indicating what should have been changed in the prediction. Such reference simplifications, however, are available for the train data (see Section 3.3 above).

Sentence	Accuracy	Completeness	Simplicity	Brevity
~~Background :~~ There is conflicting evidence about the relationship between vitamin D deficiency and depression , and a systematic assessment of the literature has not been available .	1	1	-1	0
Aims : To determine the relationship , if any , between vitamin D deficiency and depression .	1	1	0	1
~~Method :~~ A systematic review and meta-analysis of observational studies and randomised controlled trials was conducted .	1	1	-1	1
~~Results :~~ One case-control study , ten ~~cross-sectional~~ studies and three cohort studies with a total of 31 424 participants were analysed .	1	1	-1	0
~~Lower vitamin D levels were found in people with depression compared with controls (SMD = 0.60 , 95 % CI 0.23-0.97) and there~~ There was an increased odds ~~ratio~~ of depression for the lowest v. highest vitamin D categories in the cross-sectional studies ~~(OR = 1.31 , 95 % CI 1.0-1.71)~~ .	1	1	-1	1
The cohort studies showed a significantly increased hazard ratio of depression for the lowest v. highest vitamin D categories ~~(HR = 2.21 , 95 % CI 1.40-3.49)~~ .	1	1	0	0
Conclusions : Our analyses are consistent with the hypothesis that low vitamin D concentration is associated with depression , and highlight the need for randomised controlled trials of vitamin D for the prevention and treatment of depression ~~to determine whether this association is causal~~ .				

Table 7: Example $\hat{O} \rightarrow \text{BART}_{\text{sent}}$ simplifications for the complex abstract input A7 for Q17 (PMID 33924215)

Sentence	Accuracy	Completeness	Simplicity	Brevity
~~Background :~~ There is conflicting evidence about the relationship between vitamin D deficiency and depression , and a systematic assessment of the literature has not been available .	1	0	1	1
Aims : To determine the relationship ~~, if any ,~~ between vitamin D deficiency and depression .	1	1	-1	0
~~Method :~~ A systematic review and meta-analysis of ~~observational~~ studies and randomised controlled trials was conducted .	1	1	-1	1
~~Results :~~ One case-control study , ten ~~cross-sectional~~ studies and three cohort studies with a total of 31 424 participants were analysed .	1	1	-1	0
~~Lower vitamin D levels were found in people with depression compared with controls (SMD = 0.60 , 95 % CI 0.23-0.97) and there~~ There was an increased ~~odds ratio~~ risk of depression for the lowest v. highest vitamin D categories in the cross-sectional studies ~~(OR = 1.31 , 95 % CI 1.0-1.71)~~ .	1	0	1	1
The ~~cohort~~ studies showed a significantly increased hazard ratio of depression for the lowest v. highest vitamin D categories ~~(HR = 2.21 , 95 % CI 1.40-3.49)~~ .	1	1	-1	0
~~Conclusions : Our analyses are consistent with the hypothesis~~ We conclude that ~~low vitamin D concentration~~ there is ~~associated with depression , and highlight the~~ a need for randomised controlled trials of vitamin D for the prevention and treatment of depression to determine whether this association is causal .				

Table 8: Example $\hat{O} \rightarrow$ ConBART simplifications for the complex abstract input A7 for Q17 (PMID 33924215): ~~deletions~~ and insertions