Original Paper

# Classification of Cochrane Plain Language Summaries by Conclusiveness Using Transformer-Based Models and ChatGPT: Retrospective Observational Study

Antonija Mijatović[1], PhD; Luka Ursić[1], PhD; Nensi Bralić[1], PhD; Ružica Bandić[1], PhD; Barbara Ćaćić[2]; Ivan Buljan[3], PhD; Ana Marušić[1], MD, PhD

[1]Department of Research in Biomedicine and Health, Centre for Evidence-based Medicine, University of Split School of Medicine, Split, Croatia
[2]Faculty of Humanities and Social Sciences, University of Split, Split, Croatia
[3]Department of Psychology, Faculty of Humanities and Social Sciences, University of Split, Split, Croatia

**Corresponding Author:**

Antonija Mijatović, PhD
Department of Research in Biomedicine and Health
Centre for Evidence-based Medicine, University of Split School of Medicine
Šoltanska 2A
Split 21000
Croatia
Phone: 385 21557820
Email: antonija.mijatovic@mefst.hr

## Abstract

**Background:** Cochrane plain language summaries (PLSs) aim to make systematic review findings more accessible to the general public. However, inconsistencies in how conclusions are presented may impact comprehension and decision-making. Classifying PLSs based on conclusiveness can improve clarity and facilitate informed health decisions.

**Objective:** This study aimed to develop and evaluate deep learning language models for the classification of PLSs according to 3 levels of conclusiveness (conclusive, inconclusive, and unclear) and to compare their performance with a general-purpose large language model (GPT-4o).

**Methods:** We used a publicly available dataset containing 4405 Cochrane PLSs of systematic reviews published until 2019, already classified by humans according to 9 categories of conclusiveness regarding the intervention's effectiveness or safety. We merged these categories into 3 classes based on the strength of conclusiveness: conclusive, inconclusive, and unclear. For the fine-tuning, we used Scientific Bidirectional Encoder Representations from Transformers (SciBERT), a pretrained language model trained on 1.14 million papers primarily from the health sciences, and Longformer, a transformer model designed specifically to process long documents. The script was developed using the Python programming language and the PyTorch framework. We computed evaluation metrics using the *scikit-learn* machine learning library and determined the area under the curve of the receiver operating characteristic (AUCROC) to measure the model performance in balancing sensitivity and specificity. We also analyzed a separate set of 213 PLSs and compared the predictions of our pretrained models with both manual verification and outputs generated by ChatGPT.

**Results:** The model based on SciBERT achieved a balanced accuracy of 56.6%. The AUCROC was 0.91 for "conclusive," 0.67 for "inconclusive," and 0.75 for "unclear" conclusiveness classes. The Longformer-based model had a balanced accuracy of 60.9%, with AUCROCs of 0.86 for "conclusive," 0.67 for "inconclusive," and 0.72 for "unclear" conclusiveness classes. Both models underperformed compared with ChatGPT, which demonstrated higher accuracy (74.2%), better precision and recall, and a higher Cohen $\kappa$ (0.57).

**Conclusions:** Fine-tuning 2 transformer-based language models showed mixed results in classifying Cochrane PLSs by conclusiveness, likely due to semantic overlap and subtle linguistic differences. Despite satisfactory internal test metrics, the fine-tuned models failed to generalize to newly published PLSs, where performance dropped to near-chance levels. These findings suggest that general-purpose large language models like GPT-4o may currently offer more reliable results for practical classification tasks in biomedical applications.

*JMIR Med Inform 2026;14:e72657;* doi: 10.2196/72657

**Keywords:** plain language summary; PLS; large language models; Scientific Bidirectional Encoder Representations from Transformers ; SciBERT; Longformer; fine-tuning

# Introduction

A Cochrane plain language summary (PLS) is a stand-alone summary of a Cochrane systematic review used to disseminate health-related evidence to a wider audience with the goal of facilitating evidence-based decision-making about health care, particularly for medical treatments [1]. A well-written PLS should be comprehensible to readers without a background in research or health care, including patients, caregivers, and policymakers [2], and should be presented at or below a sixth-grade reading level to ensure accessibility for all readers [3]. It should also allow readers to comprehend the certainty of evidence and to correctly interpret the results, which is why the authors should not offer specific recommendations but rather present their findings clearly and guide the readers toward independent conclusions [4,5]. However, a PLS necessarily has a conclusion section conveying the main message, where the level of certainty of the evidence is presented using narrative statements [2]. For example, "Intervention causes a large reduction/increase in outcome" is a suggested narrative for large effect size and high certainty of the evidence, whereas "It is unclear if intervention has an effect on outcome" should indicate very low certainty [6,7].

Conclusiveness is an important concept in research and health care, indicating a degree of confidence in the findings and facilitating decision-making [8,9], while also ensuring that the current evidence is easily understood [10]. When provided with conclusive health information, patients rely less on health care professionals to decide on diagnosis and therapy [11,12]. Unfortunately, they do not always succeed in finding relevant information for their health condition, as shown in a study where a quarter of respondents did not find answers to the health-related inquiries they raised on the internet [13]. Patients also often find information in unreliable sources, providing misleading or false data on diagnostics and treatment [14]. In addition, numerous studies have reported higher levels of anxiety and cyberchondria with increases in online health information searches [15-17]. Given that Cochrane Reviews represent the best available knowledge in the field [1], the accurate classification of their PLSs could significantly improve patients' comprehension of these conclusions and allow them to make well-informed decisions about health care interventions.

Several studies and reviews on the conclusiveness of Cochrane PLSs found that 50% to 80% of the reviews enabled readers to reach a relevant conclusion, while the readability of PLSs was generally poor, with conclusions often unclear or missing [18-22]. However, in some cases, conclusive statements were made even when the quality of evidence was low or moderate [23]. Our previous research showed that most PLSs lacked or had unclear conclusions regarding an intervention's efficacy and safety [24].

In all of the studies on the conclusiveness of systematic reviews, the process of classifying the reviews and PLSs was carried out manually, usually by at least 2 independent assessors, which is a demanding and time-consuming task. In this study, we explored whether the classification of PLSs according to their level of conclusiveness could be conducted with the help of artificial intelligence (AI) and natural language processing (NLP). NLP algorithms, particularly deep learning models like neural networks, can automatically learn and extract patterns from language data [25], enabling them to understand context and semantics [26,27]. Their multilingual variants continue to expand and become accessible to speakers of less-represented languages [28] while domain-specific models lead to greater accuracy [29].

# Methods

## Overview

In this retrospective observational study with a supervised machine learning approach, we used a pretrained deep large language model (LLM) for PLS classification according to 3 levels of conclusiveness: conclusive, inconclusive, and unclear. We fine-tuned 2 transformer-based models—Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer—for our task. SciBERT is a variant of the Bidirectional Encoder Representations from Transformers (BERT) model specifically designed for scientific and biomedical text processing that is pretrained on a vast corpus of scientific literature consisting of 18% of papers from computer science and 82% from the biomedical field [30]. Longformer is a transformer architecture optimized for processing long documents through sparse attention mechanisms [31]. SciBERT was selected to leverage the domain-specific language of PLSs, whereas Longformer was selected to accommodate PLSs with extended length. Specifically, the median number of words in a PLS is 345 [32], which corresponds to approximately 500 tokens, just under SciBERT's 512-token limit. However, the Longformer model's extended token capacity of up to 4096 tokens allows for the processing of all PLSs without truncation.

## Data Source

We used the dataset from our previous study [24], which contains 4405 Cochrane PLSs of systematic reviews on intervention studies published until 2019, already classified by 2 independent experts into 9 categories based on the conclusiveness regarding an intervention's effectiveness or safety. We combined these categories into 3 distinct classes: conclusive, inconclusive, and unclear, allowing for a more manageable and interpretable classification task (Textbox 1).

**Textbox 1.** Classification of conclusiveness categories.

**Conclusive**

- Positive: signifies the existence of moderate- or high-quality evidence supporting the effectiveness or safety
- Negative: indicates the presence of moderate- or high-quality evidence of intervention's ineffectiveness or harm
- Equal: denotes that the analyzed interventions were of equal effectiveness or safety

**Inconclusive**

- Positive inconclusive: implies the existence of evidence supporting effectiveness or safety, yet the evidence is low quality or inconclusive. The authors suggest that more research is needed.
- Negative inconclusive: suggests there is evidence of ineffectiveness or harm (indicating that the observed effect or the intervention was unsafe), yet the evidence is low quality or inconclusive. Authors may advise against the intervention or comparison and state that more research is required.
- Equal inconclusive: indicates that the interventions exhibit comparable levels of effectiveness or safety, yet the evidence is low quality or inconclusive. The authors emphasize that more research is required.

**Unclear**

- No opinion: the authors provided no opinion.
- No evidence: there is no evidence from randomized controlled trials because the literature search did not result in any eligible studies (ie, empty reviews).
- Unclear: the authors did not present clear conclusions.

With the classification, the evidence in the "conclusive" class is strong and clear, irrespective of the direction of the effect, as opposed to the "inconclusive" class, where it is uncertain or of lower quality. In the "unclear" class, conclusions are absent, either because authors have not provided a clear opinion or due to a lack of available evidence. This lack of conclusiveness is not indicative of a poorly written PLS, as long as the PLS accurately represents the findings from the systematic review. This is why we must differentiate between PLSs that conclude that there is "no evidence" and those that offer no opinion or present unclear conclusions [24].

## Language Processing Models

LLMs are important components of NLP designed to understand and generate human language. LLMs such as GPT-3 and BERT are pretrained on massive datasets containing text from the internet [33,34]. Most importantly, they are highly adaptable, meaning they can be fine-tuned for specific tasks [35]. For example, Beltagy et al [30] fine-tuned BERT, an LLM that had been pretrained on a wide range of text on the internet [33], to develop SciBERT, an LLM trained on a vast corpus of scientific literature, primarily from biomedical and life sciences, making it suitable for NLP tasks in the scientific and medical research domains. Similarly, Longformer was developed to address the limitations of handling long documents; it uses a sparse attention mechanism, where each token focuses on a limited local context rather than the entire sentence [31].

We achieved transfer learning by further fine-tuning SciBERT and Longformer on our specific PLS classification task. In transfer learning, the LLM adapts its learned features to the nuances of the new task while retaining the knowledge it acquired during pretraining [36]. This approach is intended to mirror how humans learn, as we often apply knowledge and skills acquired in one context to solve new, related problems [37]. Transfer learning not only speeds up the training process but also leads to better performance compared with training from scratch [38].

## Experimental Setup and Fine-Tuning

We wrote the script using the Python programming language (version 3.12.3; Python Software Foundation) with the help of the PyTorch framework [39] and executed it within the Jupyter Notebook environment [40] using the NVIDIA GeForce RTX 3080 GPU (version 8200).

Both of our models came from the Hugging Face library [41]. For SciBERT, we used its associated tokenizer, setting the maximum token length to 512. For Longformer, we extended the maximum token length to 2048 to accommodate the full content of the PLSs without truncation. Both models included a dropout layer with a rate of 0.5, a regularization technique that reduces the risk of overfitting.

For both models, we used a 768-dimensional pooled embedding vector as input to our classifier. This representation was passed through a dropout layer and a linear layer that produced a 3D output corresponding to our target classes. SciBERT used its built-in pooled ([CLS] token) representation, whereas Longformer used mean pooling across all token embeddings due to the absence of a pooler layer. A standard attention mask was applied during encoding, ensuring that padding tokens were fully excluded from self-attention computations. We used AdamW as the optimizer to update network weights [42] and categorical cross entropy as our loss function [43]. We set the maximum number of training epochs (where 1 epoch represents a complete pass through the training dataset) to 15 for SciBERT and 10 for Longformer, with early stopping based on validation loss. In practice, SciBERT training stopped after 7 epochs due to early stopping. The learning rate was set to $2\times10^{-5}$ for both the models. Since these hyperparameters cannot be determined a priori, they were selected via trial and error [44]. For this reason, we monitored training and validation performance and used early stopping based on validation loss to prevent overfitting. The best-performing models (lowest validation loss) were saved and later used for evaluation.

Baseline models (SciBERT and Longformer) were implemented using a frozen feature-extraction transfer-learning approach. All pretrained transformer encoder weights were frozen, and only the newly added linear classification layer was trainable. The encoder itself did not undergo any gradient updates. Therefore, this baseline represents a lightweight transfer-learning model.

### Data Splitting and Handling of Class Imbalance

We divided the dataset into training (80%), testing (10%), and validation (10%) subsets. Random undersampling was applied only to the training set, where the smallest class contained 343 PLSs. All classes were downsampled to this size to create a balanced training set.

### Model Validation

To assess performance, we used functions from *scikit-learn* [45], including *balanced_accuracy_score*, which calculates balanced accuracy for addressing imbalances in multiclass datasets, and *precision_recall_fscore_support*, which provides precision, recall, and F-beta scores for each class. Precision is the proportion of true positives relative to the total of true positives and false positives. Recall is the proportion of true positives relative to the total of true positives and false negatives. The $F_1$-score is the harmonic mean of the precision and recall. Additionally, we evaluated the model's ability to balance sensitivity and specificity by measuring the area under the curve of the receiver operating characteristic (AUCROC), which is the proportion of area below the receiver operating characteristic curve, which in turn is the plot of the true positive rate against the false positive rate.

### Effect of Training and Validation Split on Model Performance

To evaluate the impact of different training and validation splits on model performance, we evaluated SciBERT's performance with 10%, 20%, and 30% of the data reserved for validation. This analysis was conducted to assess the relationship between training set size and classification accuracy, based on the assumption that larger training sets may improve model learning. By progressively reducing the number of training samples, we tested the extent to which performance would degrade with less training data.

### Manual Validation and Comparison With GPT-4o Performance

To evaluate model performance, we used a separate verification dataset consisting of 213 Cochrane PLSs published between September 2024 and May 2025, which was not a part of the original training or evaluation datasets. Each PLS was independently classified by 2 domain experts, with a third expert resolving any discrepancies (NB, a metascientist with expertise in Cochrane PLSs and conclusiveness assessment; RB, a research assistant at the Department of Research in Biomedicine and Health; and BĆ, a psychology student with experience using AI tools to assess compliance with reporting guidelines).

The GPT-4o model was prompted using a zero-shot classification approach, which means no example classifications were provided in the prompt. The prompt instructed the model to classify each PLS into one of 3 predefined categories and included detailed definitions for each class, identical to those used by human annotators. The complete prompt can be found in Multimedia Appendix 1.

The performance of the 2 trained BERT-based models—SciBERT and Longformer—was compared with the baseline GPT-4o model (subscription-based) and evaluated against labels assigned by human experts. Model outputs were compared using standard classification metrics: accuracy, precision, recall, $F_1$-score, and Cohen $\kappa$, with the lattermost being used to assess the level of agreement between model predictions and the expert consensus categorization.

### Calibration Analysis

To evaluate the reliability of predicted probabilities, we conducted a calibration analysis on the fine-tuned SciBERT model. Predicted probabilities were compared with observed outcome frequencies across the 3 target classes using calibration plots and quantitative metrics, including expected calibration error and Brier score. Since Longformer achieved similar classification performance, calibration was performed only for SciBERT.

### Ethical Considerations

This study did not involve human participants or the collection of private data. A publicly available dataset of Cochrane PLSs was used, which can be accessed via the Open Science Framework [46].

The use of publicly available data is exempt from ethics review in accordance with the University of Split School of Medicine guidelines and the Croatian Science Foundation project Professionalism in Health – Decision making in practice and research (IP-2019-04-4882) [47]. Therefore, institutional review board approval and informed consent were not required. The dataset used in this study was collected and shared under conditions that permit secondary analysis without additional consent requirements. The data contain no personally identifiable information. All analyzed PLSs are publicly accessible textual documents.

## Results

### Overview

Among the 4405 PLSs from our dataset, 429 (9.7%) had been manually categorized as conclusive, 1203 (27.3%) as inconclusive, and 2773 (63%) as unclear [24]. These PLSs served as input data for our model. To address class imbalance, we applied random undersampling to the training set by selecting an equal number of PLSs from each class (n=429). This ensured that we had a balanced dataset and reduced the risk of biased model learning.

In classifying the PLSs, the SciBERT model achieved a balanced accuracy of 56.6%, with AUCROC scores of 0.91 for the conclusive, 0.67 for the inconclusive, and 0.75 for the

unclear class. The Longformer model, meanwhile, demonstrated a balanced accuracy of 60.9%, with AUCROC scores of 0.86, 0.67, and 0.72 for the same classes, respectively. The receiver operating characteristic curves and confusion matrices are visualized in Figure 1 and Figure 2, while Table 1 presents a side-by-side comparison of the performance for both models across all classes.

**Figure 1.** Receiver operating characteristic (ROC) curves and the corresponding area under the curve of the receiver operating characteristic scores for each class (0=conclusive, 1=inconclusive, and 2=unclear); (A) SciBERT model and (B) Longformer model. Calculated and visualized using scikit-learn. AUC: area under the curve.

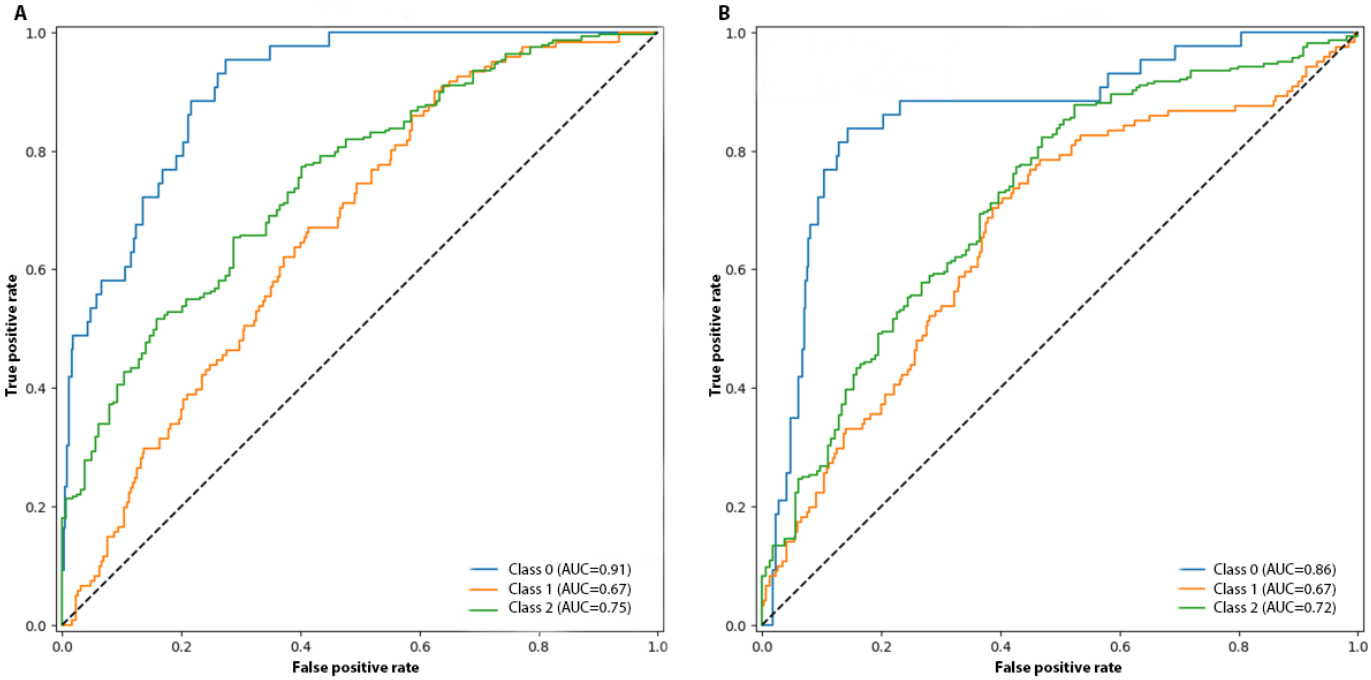


**Figure 2.** Confusion matrix for the fine-tuned (A) SciBERT and (B) Longformer multiclass classification (0=conclusive, 1=inconclusive, and 2=unclear) used to visualize the performance of the conclusiveness classification algorithm (created using the scikit-learn). The confusion matrix evaluates multiclass performance by comparing the predicted classes with the actual classes. The diagonal elements represent the correct predictions.

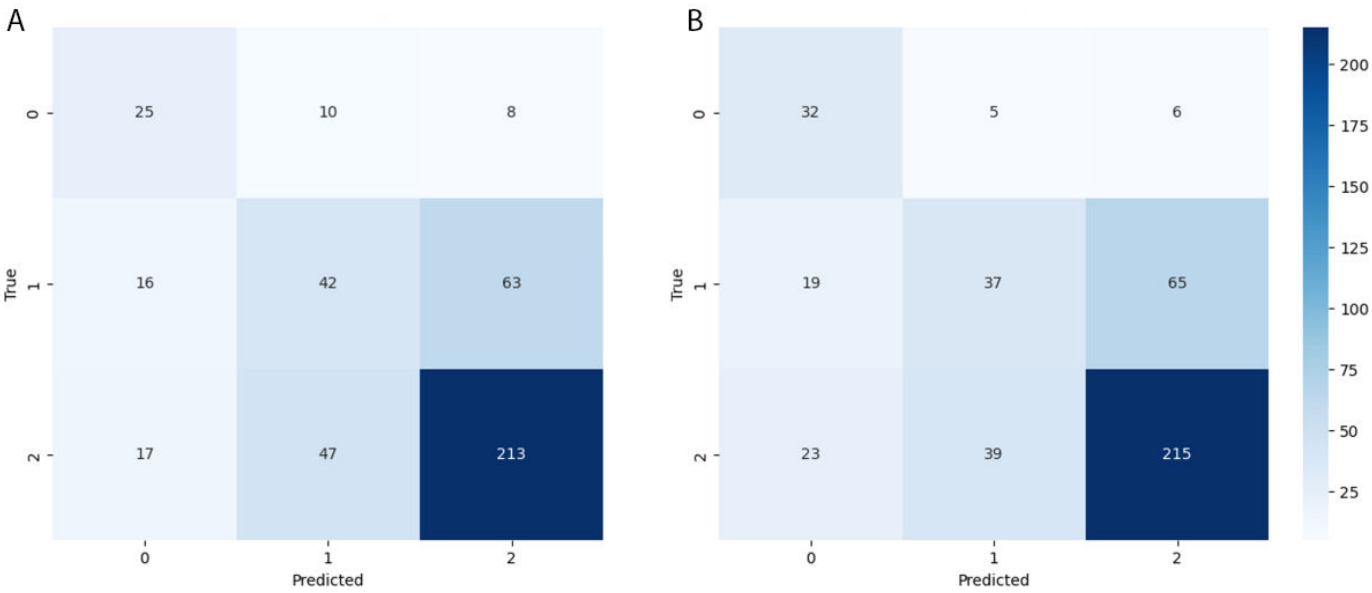

**Table 1.** Per-class performance of the conclusiveness classification for Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer models[a].

| Model | Precision | Recall | $F_1$-score | AUCROC[b] score |
|---|---|---|---|---|
| Conclusive | | | | |
| SciBERT | 0.43 | 0.58 | 0.50 | 0.91 |
| Longformer | 0.43 | 0.74 | 0.55 | 0.87 |
| Inconclusive | | | | |
| SciBERT | 0.42 | 0.35 | 0.38 | 0.67 |
| Longformer | 0.46 | 0.31 | 0.37 | 0.67 |
| Unclear | | | | |
| SciBERT | 0.75 | 0.77 | 0.76 | 0.75 |
| Longformer | 0.75 | 0.78 | 0.76 | 0.72 |

[a]Precision is the proportion of true positives relative to the total of true positives and false positives. Recall is the proportion of true positives relative to the total of true positives and false negatives. $F_1$-score is the harmonic mean of the precision and recall. Area under the curve of the receiver operating characteristic is the proportion of area below the receiver operating characteristic curve, which is the plot of the true positive rate against the false positive rate. The scores were obtained using the scikit-learn functions.
[b]AUCROC: area under the curve of the receiver operating characteristic.

For comparison, the baseline SciBERT model achieved a balanced accuracy of only 42.2%, with AUCROC scores of 0.68 for the conclusive, 0.53 for the inconclusive, and 0.53 for the unclear class. The baseline Longformer's balanced accuracy was 39.0%, with AUCROC scores of 0.69, 0.56, and 0.54 for the same classes, respectively. However, the baseline Longformer was unstable across repeated runs and sometimes predicted only one class. This indicates that the frozen encoder was not able to provide features that separate the 3 categories well. The performance of all models is documented in Multimedia Appendix 1.

## Effect of the Training Set Size and Validation Split on Model Performance

The highest performance was observed with a 10% validation split, whereas 20% and 30% validation splits resulted in similarly reduced accuracies (Table 2).

**Table 2.** Performance of Scientific Bidirectional Encoder Representations from Transformers with varying training and validation splits.

| Validation split | Training samples (per class), n | Balanced accuracy (%) |
|---|---|---|
| 10% | 343 | 56.6 |
| 20% | 257 | 53.1 |
| 30% | 172 | 53.3 |

## Manual Validation and Comparison With ChatGPT Performance

Among the 213 PLSs from our additional verification dataset, 44 (20.7%) had been manually categorized as conclusive, 110 (51.6%) as inconclusive, and 59 (27.7%) as unclear. The Cohen $\kappa$ value between the experts was 0.57, indicating moderate agreement. The baseline GPT-4o model outperformed the trained BERT-based models (Table 3). SciBERT had the poorest performance overall, while ChatGPT had the strongest, with most of its conclusive predictions being correct. ChatGPT also achieved moderate agreement with the human classifications (Cohen $\kappa$=0.57). This indicates that ChatGPT's classifications were as consistent with the expert consensus as an individual expert's were with one another. In contrast, SciBERT and Longformer demonstrated poor alignment with the reference classifications, with a Cohen $\kappa$ value of 0.03 and 0.19, respectively, suggesting worse-than-random agreement (Table 3).

**Table 3.** Comparative performance of fine-tuned Scientific Bidirectional Encoder Representations from Transformers (SciBERT) and Longformer models and baseline GPT-4o model on conclusiveness classification task.

| Metric | Fine-tuned SciBERT | Fine-tuned Longformer | Baseline GPT-4o |
|---|---|---|---|
| Precision | 0.34 | 0.57 | 0.74 |
| Recall | 0.34 | 0.44 | 0.74 |
| $F_1$-score | 0.27 | 0.40 | 0.74 |
| Accuracy (%) | 34.3 | 44.1 | 74.2 |
| Cohen $\kappa$[a] | 0.03 | 0.19 | 0.57 |

[a]Predictions were compared against expert annotations, where each plain language summary was manually classified by 2 experts, with a third expert resolving any discrepancies.

### Calibration Analysis

The calibration analysis was performed on the fine-tuned SciBERT model, which was overconfident for the "conclusive" class, poorly calibrated for the "inconclusive" class, and initially underconfident for the "unclear" class, although highly accurate when assigning high probabilities. Calibration plots and quantitative metrics, including expected calibration error and Brier scores (a measure of the average squared difference between predicted probability and actual outcome), are presented in Multimedia Appendix 1.

## Discussion

### Principal Results

Our results showed that transformer-based language models such as SciBERT and Longformer achieved modest performance in classifying Cochrane PLSs based on their level of conclusiveness. Both models were fine-tuned on a balanced dataset and evaluated using standard classification metrics, with Longformer achieving a balanced accuracy of 60.9%, compared with 56.6% for SciBERT. Both models performed best on the conclusive class, achieving relatively high AUCROC and $F_1$-scores. For the unclear class, SciBERT demonstrated stronger precision and recall. However, both models struggled to distinguish the inconclusive class, with the lowest $F_1$-scores and overlapping errors with the unclear category. Both models underperformed on the inconclusive class, with poor AUCROC and low precision and recall scores. These findings may indicate that conclusiveness is expressed in linguistically nuanced ways that the models are unable to detect and that there could be a semantic overlap between inconclusive and unclear PLSs. Both models were outperformed by ChatGPT, which achieved better accuracy and interrater agreement. This suggests that general-purpose LLMs may offer more reliable performance for this classification task, even without domain-specific fine-tuning. Notably, GPT-4o achieved the same Cohen $\kappa$ value (0.57) as the agreement between human experts, suggesting that it mirrors expert-level judgment and nuanced reasoning. This finding highlights the potential of general-purpose LLMs to approximate human evaluation in semantically complex classification tasks.

### Qualitative Insights From Manual Classification

During manual labeling of PLSs, we identified several challenges that may explain why the experts, the fine-tuned models, and ChatGPT all struggled with the classifications. First, there was some ambiguity between the "inconclusive" and "unclear" classes. For example, some PLSs did not clearly state whether the evidence was insufficient, which might be why both human assessors and models were uncertain when assigning these labels. Furthermore, the interpretation of the criteria for the "conclusive" class was occasionally ambiguous, particularly in cases where PLSs included recommendations but lacked clear statements about intervention effectiveness. This ambiguity likely made it difficult for human annotators to determine whether the conclusion was strong enough to classify a PLS as conclusive or inconclusive. Consequently, models trained on these labels may have inherited this ambiguity. We also observed that PLSs included expressions such as "may help" or "probably works," which are common in scientific writing but can signal uncertainty. This nuance might have been difficult for models to detect, explaining their lower performance in differentiating between inconclusive and unclear statements. These findings suggest that better model instruction, such as through advanced prompt engineering, might help improve future performance of the GPT-4o model. They also highlight the need to incorporate linguistic features of uncertainty more explicitly into the training process.

### Comparison With Prior Work

To our knowledge, there have been no studies on the automatic classification of Cochrane PLSs or full reviews based on the level of conclusiveness, although some studies examined machine learning techniques for making systematic review processes more efficient. For example, one study developed a randomized controlled trial classifier for Cochrane Reviews, a tool that discerns whether a selected study qualifies as a randomized controlled trial [48]. In another, ChatGPT showed strong performance when used for abstractive summarization of longer texts, including news articles and public speeches [49]. However, given the lack of specialized expertise in the field of medicine, ChatGPT does not always grasp the nuances of its terminology and sometimes struggles to recognize important information [50]. In addition, one study found that LLMs sometimes generate factually inconsistent summaries, which could potentially harm readers [51]. Yet, these challenges and the related legal and ethical issues should not discourage the use of LLMs but rather encourage further research and refinement of the technology.

The fine-tuned BERT-based models did not perform well in our classification task, indicating limitations in generalizing to nuanced language in PLSs. In contrast, general-purpose language models can perform better than fine-tuned models in some classification tasks, achieving Cohen $\kappa$ scores comparable to those of human experts. This is likely because they have been trained on much larger and more diverse text corpora and have more complex architectures, allowing them to better understand context and differentiate linguistic nuances [52]. This is also in line with findings by Davidson and Chae [53] that LLMs, particularly when fine-tuned on prompts that include explicit instructions, can outperform traditional supervised models in a variety of classification settings without task-specific training.

### Limitations

This study has several limitations. First, although we began with a relatively large dataset of 4405 PLSs, the "conclusive" class comprised only about 9.73% (n=429) of the total dataset. To address this issue, we applied random undersampling, which reduced the number of PLSs in the "inconclusive" and "unclear" classes. While this approach ensured balanced class representation, it also removed a substantial

amount of data (2344/2773, 84.53% of the "unclear" class) that could have supported more robust model learning. This likely limited the models' ability to learn the linguistic variability of the majority class. In contrast, GPT-4o was used in zero-shot inference settings and was not trained on our dataset, meaning that its performance could not be affected by the undersampling procedure that constrained the fine-tuned models. Alternative approaches, such as applying class-weighted loss functions or oversampling minority classes, may yield improved performance in future work.

Second, the dataset was based exclusively on PLSs of Cochrane Reviews, representing a single domain within evidence-based health literature. This may also limit the model's generalizability to other types of health communication.

Third, there were often very subtle linguistic differences between the inconclusive and unclear PLSs, which introduced noise in model classification. Some PLSs lacked clear phrasing or used expressions such as "may help," and "probably works" that were difficult to interpret consistently, even for human annotators. This ambiguity likely contributed to the models' difficulty in separating these two classes.

Fourth, although one researcher (IB) participated in and provided instruction for both the original 2019 annotation and the current one, the full annotator teams differed between the two studies. It is possible that there were subtle differences in how annotators interpreted or applied the criteria, raising the possibility of annotator drift between the original labels used for model training and the new labels used for verification. Such drift may partly account for the observed decline in performance of the fine-tuned models on the newer dataset.

Fifth, when comparing accuracy across different splits, it is important to note that altering the training and validation proportion also changes the size of the remaining test set. Because of this, the test benchmarks were not identical across these experiments. While this does not affect the qualitative pattern we observed, the varying test baseline may contribute to numerical differences in accuracy.

Additionally, the fine-tuned SciBERT model showed high volatility in validation loss across training epochs, including abrupt spikes prior to early stopping. This suggests that the model may not have reached a fully stable convergence point, possibly due to the limited size of the training set and the complexity of the task. Such fluctuations may have constrained the model's performance.

Finally, although we compared our fine-tuned models with GPT-4o, we did not use advanced prompt engineering or task-specific tuning. This likely underestimated ChatGPT's performance on this classification task.

## Future Work and Recommendations

Future studies should expand the dataset to include PLSs from non-Cochrane sources and from different health domains, which could improve model generalizability. First, although there was no meaningful difference in the performance of different training and validation splits, future work should also explore the impact of larger datasets on model performance. Second, adding task-specific fine-tuning and advanced prompt engineering of LLMs such as GPT-4o could improve classification accuracy even more. Third, models may be able to more successfully differentiate between inconclusive and unclear classes by incorporating linguistic features that capture uncertainty and conclusiveness cues.

In the long term, implementing a general-purpose LLM, such as ChatGPT, within Cochrane platforms (eg, the RevMan Web dashboard) could assist authors in ensuring that their conclusions are clear and guide readers through reviews according to their conclusiveness level. Additionally, LLMs could support users of the Cochrane Library by offering filters or tags that group PLSs by conclusiveness. However, additional model validation and ethical review should precede these real-world applications.

## Conclusions

We explored the use of 2 fine-tuned transformer-based models—SciBERT and Longformer—for classifying Cochrane PLS according to their level of conclusiveness. Both models demonstrated modest internal performance but poor generalization to newly published PLSs, particularly in distinguishing between inconclusive and unclear categories, likely due to their semantic overlap. An empirical analysis of different training and validation splits confirmed that larger training sets improve model performance, although the gains were modest. Most importantly, both models were outperformed by ChatGPT, which as a general-purpose language model, achieved higher accuracy (74%) and agreement with expert annotations, suggesting that state-of-the-art LLMs hold greater potential for health care information dissemination.

### Funding

This study was funded by the Croatian Science Foundation "Professionalism in Health – Decision making in practice and research" (ProDeM) under grant IP-2019-04-4882. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Data Availability

The datasets used and analyzed during the study are publicly available on the Open Science Framework website [46].

### Authors' Contributions

Conceptualization: A Mijatović, IB, LU
Data curation: A Mijatović
Formal analysis: A Mijatović, BĆ, IB, LU, NB, RB

Funding acquisition: A Marušić
Investigation: A Mijatović
Methodology: A Mijatović
Project administration: A Marušić
Resources: IB
Software: A Mijatović
Supervision: A Marušić
Validation: A Mijatović, BĆ, NB, RB
Visualization: A Mijatović
Writing—original draft: A Mijatović
Writing—review and editing: A Marušić, BĆ, IB, LU, NB, RB
All authors approved the submitted version and take accountability for the work.

## Conflicts of Interest

A Marušić and NB are active Cochrane members. All other authors declare no other conflicts of interest.

## Multimedia Appendix 1

Supplementary material presenting detailed model training procedures, experimental setups, and full performance metrics for SciBERT and Longformer models.
[DOCX File (Microsoft Word File), 765 KB-Multimedia Appendix 1]

## References

1. Whiting P, Davenport C. Writing a plain language summary. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. John Wiley & Sons; 2023.
2. Higgins JP, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Collaboration; 2023. URL: https://www.cochrane.org/authors/handbooks-and-manuals/handbook [Accessed 2026-03-30]
3. Weiss BD. Health Literacy and Patient Safety: Help Patients Understand: Manual for Clinicians. 2nd ed. American Medical Association; 2007. URL: https://books.google.be/books?id=quJaYgEACAAJ [Accessed 2026-03-30]
4. Schünemann HJ, Vist GE, Higgins JP, Santesso N, Deeks JJ, Glasziou P, et al. Chapter 15: Interpreting results and drawing conclusions. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6.5. Cochrane Collaboration; 2024.
5. Glenton C, Nilsen ES, Sporstøl Fønhus M, Goudie S, Noonan E. How to write a plain language summary of a Cochrane intervention review (version 10). Cochrane Norway. URL: https://www.cochrane.no/how-write-plain-language-summary [Accessed 2026-03-22]
6. Lasserson T, Santesso N, Cumpston M, Marshall R, Ógáin ON. Incorporating GRADE in Cochrane reviews: feedback from the CEU screening programme. Cochrane; 2016. URL: https://www.cochrane.org/sites/default/files/uploads/PDFs/MECIR/Incorporating%20GRADE%20in%20Cochrane%20Reviews.pdf [Accessed 2026-03-22]
7. Santesso N, Glenton C, Dahm P, et al. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. J Clin Epidemiol. Mar 2020;119:126-135. [doi: 10.1016/j.jclinepi.2019.10.014] [Medline: 31711912]
8. Changiz T, Yousefy A, Fakhari M. Research utilization process model: a cyclical, spiral, and developmental process to provide conclusive research knowledge in health professions education. Med J Islam Repub Iran. 2020;34:79. [doi: 10.34171/mjiri.34.79] [Medline: 33306047]
9. Kurtzman ET, Greene J. Effective presentation of health care performance information for consumer decision making: a systematic review. Patient Educ Couns. Jan 2016;99(1):36-43. [doi: 10.1016/j.pec.2015.07.030] [Medline: 26277826]
10. Gray JA. Discussion and conclusion. AME Med J. 2019;4:26. [doi: 10.21037/amj.2019.04.05]
11. Bujnowska-Fedak MM, Węgierek P. The impact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health. Jan 31, 2020;17(3):880. [doi: 10.3390/ijerph17030880] [Medline: 32023828]
12. Tan SS, Goonawardene N. Internet health information seeking and the patient-physician relationship: a systematic review. J Med Internet Res. Jan 19, 2017;19(1):e9. [doi: 10.2196/jmir.5729] [Medline: 28104579]
13. Murray E, Lo B, Pollack L, et al. The impact of health information on the internet on the physician-patient relationship: patient perceptions. Arch Intern Med. Jul 28, 2003;163(14):1727-1734. [doi: 10.1001/archinte.163.14.1727] [Medline: 12885689]
14. Tonsaker T, Bartlett G, Trpkov C. Health information on the internet: gold mine or minefield? Can Fam Physician. May 2014;60(5):407-408. [Medline: 24828994]

15. Peng RX. How online searches fuel health anxiety: investigating the link between health-related searches, health anxiety, and future intention. Comput Human Behav. Nov 2022;136:107384. [doi: 10.1016/j.chb.2022.107384]
16. Müller A, Baumann E, Dierks ML. Cyberchondria - a new behavioral syndrome? [Article in German] Psychother Psychosom Med Psychol. Jun 2021;71(6):243-255. [doi: 10.1055/a-1348-8059] [Medline: 34102694]
17. Doherty-Torstrick ER, Walton KE, Fallon BA. Cyberchondria: parsing health anxiety from online behavior. Psychosomatics. 2016;57(4):390-400. [doi: 10.1016/j.psym.2016.02.002] [Medline: 27044514]
18. Šuto J, Marušić A, Buljan I. Linguistic analysis of plain language summaries and corresponding scientific summaries of Cochrane systematic reviews about oncology interventions. Cancer Med. May 2023;12(9):10950-10960. [doi: 10.1002/cam4.5825] [Medline: 36951519]
19. Mimouni M, Mimouni F, Segev F. Conclusiveness of the Cochrane Eye and Vision Group Reviews. BMC Res Notes. Jun 16, 2015;8:242. [doi: 10.1186/s13104-015-1221-x] [Medline: 26076817]
20. Yin S, Chuai Y, Wang A, Zhang L. Conclusiveness of the Cochrane reviews in gynaecological cancer: a systematic analysis. J Int Med Res. Jun 2015;43(3):311-315. [doi: 10.1177/0300060515574922] [Medline: 25870179]
21. Mandel D, Littner Y, Mimouni FB, Lubetzky R. Conclusiveness of the Cochrane Neonatal Reviews: a systematic analysis. Acta Paediatr. Oct 2006;95(10):1209-1212. [doi: 10.1080/08035250600580537] [Medline: 16982491]
22. Cohen S, Lubetzky R, Mimouni FB, Marom R, Mandel D. Conclusiveness of the Cochrane Reviews in pediatric-gastroenterology: a systematic analysis. Eur J Gastroenterol Hepatol. Feb 2013;25(2):252-254. [doi: 10.1097/MEG.0b013e32835a1083] [Medline: 23044810]
23. Conway A, Conway Z, Soalheira K, Sutherland J. High quality of evidence is uncommon in Cochrane systematic reviews in Anaesthesia, Critical Care and Emergency Medicine. Eur J Anaesthesiol. Dec 2017;34(12):808-813. [doi: 10.1097/EJA.0000000000000691] [Medline: 29095726]
24. Banić A, Fidahić M, Šuto J, et al. Conclusiveness, linguistic characteristics and readability of Cochrane plain language summaries of intervention reviews: a cross-sectional study. BMC Med Res Methodol. Sep 10, 2022;22(1):240. [doi: 10.1186/s12874-022-01721-7] [Medline: 36088293]
25. Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly Media; 2009. ISBN: 9780596516499
26. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713-3744. [doi: 10.1007/s11042-022-13428-4] [Medline: 35855771]
27. A complete guide to natural language processing. DeepLearning.AI. URL: https://www.deeplearning.ai/resources/natural-language-processing [Accessed 2024-09-21]
28. Srinivasan A, Sitaram S, Ganu T, Dandapat S, Bali K, Choudhury M. Predicting the performance of multilingual NLP models. arXiv. Preprint posted online on Oct 21, 2021. [doi: 10.48550/arXiv.2110.08875]
29. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. Oct 15, 2021;3(1):1-23. [doi: 10.1145/3458754]
30. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 3-7, 2019; Hong Kong, China. [doi: 10.18653/v1/D19-1371]
31. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on Apr 10, 2020. [doi: 10.48550/arXiv.2004.05150]
32. Bralić N, Buljan I. The association between research design and the perceived treatment effectiveness: a cross-sectional study. Front Med (Lausanne). Dec 22, 2023;10:1220999. [doi: 10.3389/fmed.2023.1220999] [Medline: 38196834]
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on Oct 11, 2018. [doi: 10.48550/arXiv.1810.04805]
34. Causevic S. Evolution of large language models—BERT, GPT3, MUM, and PaLM. Medium. 2022. URL: https://medium.com/data-science/self-supervised-transformer-models-bert-gpt3-mum-and-paml-2b5e29ea0c26 [Accessed 2026-03-23]
35. Fine-tuning. Hugging Face. URL: https://huggingface.co/docs/transformers/training [Accessed 2026-03-23]
36. Winastwan R. Text classification with BERT in PyTorch. Towards Data Science; 2021. URL: https://towardsdatascience.com/text-classification-with-bert-in-pytorch-887965e5820f [Accessed 2024-09-21]
37. What is transfer learning? GeeksforGeeks. URL: https://www.geeksforgeeks.org/ml-introduction-to-transfer-learning [Accessed 2024-09-21]
38. Oztel I, Yolcu G, Oz C. Performance comparison of transfer learning and training from scratch approaches for deep facial expression recognition. Presented at: 2019 4th International Conference on Computer Science and Engineering (UBMK); Sep 11-15, 2019; Samsun, Turkey. [doi: 10.1109/UBMK.2019.8907203]

39. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, Alche'-Buc FD, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc; 2019:8024-8035.
40. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016:87-90. [doi: 10.3233/978-1-61499-649-1-87]
41. Models. Hugging Face. URL: https://huggingface.co/models [Accessed 2024-09-21]
42. AdamW. PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html#adamw [Accessed 2024-09-21]
43. CrossEntropyLoss. PyTorch. URL: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html [Accessed 2024-09-21]
44. Reed R, Marks RJ. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. MIT Press; 1999. ISBN: 9780262181907
45. scikit-learn: machine learning in Python. scikit-learn. URL: https://scikit-learn.org/stable [Accessed 2024-09-21]
46. Readability, linguistic characteristics and conclusiveness of Cochrane plain language summaries of intervention reviews: a cross-sectional study. OSF. URL: https://osf.io/qvu3a/overview [Accessed 2026-03-30]
47. Code of ethics of the Croatian Science Foundation. Hrvatska Zaklada za Znanost. 2020. URL: https://hrzz.hr/wp-content/uploads/2020/08/HRZZ-Code-of-Ethics.pdf [Accessed 2026-03-30]
48. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. May 2021;133:140-151. [doi: 10.1016/j.jclinepi.2020.11.003] [Medline: 33171275]
49. Povey N. ChatGPT: abstractive text summarization. Medium. 2022. URL: https://medium.com/@nadirapovey/chatgpt-text-summarization-44f768222a4c [Accessed 2024-09-21]
50. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
51. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. Aug 24, 2023;6(1):158. [doi: 10.1038/s41746-023-00896-7] [Medline: 37620423]
52. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020. [doi: 10.48550/arXiv.2005.14165]
53. Davidson T, Chae Y. Large language models for text classification: from zero-shot learning to instruction-tuning. Sociol Methods Res. 2025. [doi: 10.1177/00491241251325243]


## Abbreviations

**AI:** artificial intelligence
**AUCROC:** area under the curve of the receiver operating characteristic
**BERT:** Bidirectional Encoder Representations from Transformers
**LLM:** large language model
**NLP:** natural language processing
**PLS:** plain language summary
**SciBERT:** Scientific Bidirectional Encoder Representations from Transformers


*Edited by Alexandre Castonguay, Andrew Coristine; peer-reviewed by Ford Lumban Gaol, Ramin Homayouni; submitted 14.Feb.2025; final revised version received 23.Dec.2025; accepted 23.Dec.2025; published 14.Apr.2026*

*Please cite as:*
*Mijatović A, Ursić L, Bralić N, Bandić R, Ćaćić B, Buljan I, Marušić A*
*Classification of Cochrane Plain Language Summaries by Conclusiveness Using Transformer-Based Models and ChatGPT: Retrospective Observational Study*
*JMIR Med Inform 2026;14:e72657*
*URL: https://medinform.jmir.org/2026/1/e72657*
*doi: 10.2196/72657*