Can NLP Models Detect When One Publication Outweighs Twenty? Predicting Systematic Review Conclusion Changes

Ebrahim Alharbi[1,2] and Mark Stevenson[1]

[1]School of Computer Science, University of Sheffield, UK
[2]Department of Computer Science, King Abdulaziz University, Saudi Arabia
{eaalharbi1, mark.stevenson}@sheffield.ac.uk

Abstract

Systematic reviews underpin evidence-based medicine but can outdate quickly when new evidence appears. We formulate a novel prediction task: given a review and new studies that have appeared since its publication, predict whether the review's conclusions will change. A dataset of 3,326 Cochrane review-update pairs is constructed and a range of approaches explored including feature-based baselines, zero- and few-shot LLMs, in addition to parameter-efficient fine-tuning. Fine-tuning Qwen2.5-14B achieves the highest AUC-ROC (70.4%).

1 Introduction

Systematic reviews synthesise all available evidence on a clinical question and are fundamental to evidence-based medicine (Sackett et al., 1996). They need to be kept up to date to ensure they reflect the latest evidence (Herrera-Perez et al., 2019; Garner et al., 2016; Cumpston and Flemyng, 2023). However, the rate at which new scientific publications appear means that 23% risk being outdated within two years of publication (Shojania et al., 2007). Identifying whether new evidence warrants a conclusion change is a labour-intensive process (Garner et al., 2016).

A growing body of work has sought to reduce the manual burden of systematic review production through automation (Cohen et al., 2006; Marshall et al., 2016; Jonnalagadda et al., 2015; Marshall and Wallace, 2019), and Large Language Models (LLMs) show promise for biomedical reasoning (Jin et al., 2019; Wadden et al., 2020).

However, no prior work has used textual evidence to predict whether a review's conclusions need to change. The only directly relevant work is Bashir et al. (2019, 2021), who predicted conclusion change risk from four metadata features (trial count, participant count, coverage score, time elapsed) using tree-based classifiers without directly considering the text of publications.

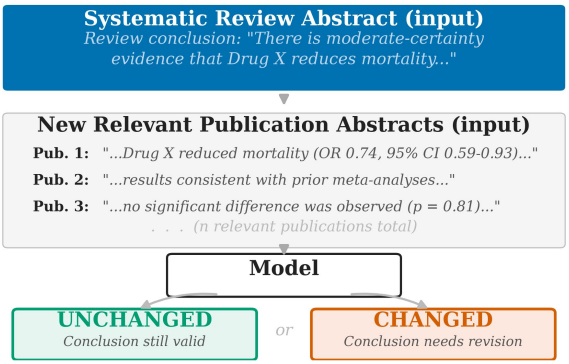

Figure 1: Task overview. Given a systematic review abstract and the abstracts of new relevant publications, the model predicts whether the review's conclusions should be changed. The gold label is derived from comparing conclusions across consecutive Cochrane review versions.

Unlike relevance screening or data extraction, conclusion change prediction requires assessing whether new findings alter an existing review's conclusions (see Figure 1). Metadata may correlate with update need but cannot capture whether a single well-designed trial contradicts twenty confirmatory ones. This motivates a text-based formulation: given a review abstract and the abstracts of its newly included publications, can language models assess whether the conclusions should change?

Beyond the immediate application, this task exposes a broader NLP challenge: reasoning over a variable-size evidence set where document count correlates with but does not determine the label. When documents are concatenated, models can exploit evidence volume as a shortcut rather than reasoning about content.

This paper makes three contributions: (1) formulates conclusion change prediction as a multi-document reasoning task, (2) develops a dataset of 3,326 Cochrane review-update pairs, over ten times larger than in prior work, with naturally derived gold labels, and (3) evaluates a range of ap-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 843–852
July 3-4, 2026

proaches, identifying fine-tuned language models as the highest-discrimination method (AUC-ROC 70.4%).

2 Related Work

Systematic Review Automation Substantial work has applied NLP and machine learning to systematic review production, including citation screening (Cohen et al., 2006; Wallace et al., 2010; Van De Schoot et al., 2021), risk of bias assessment (Marshall et al., 2016), data extraction (Jonnalagadda et al., 2015), and pipeline integration (Marshall and Wallace, 2019). These efforts focus on review creation rather than maintenance.

Review Currency and Updating Shojania et al. (2007) reported a median survival time of 5.5 years for systematic reviews before the appearance of new evidence potentially affecting their conclusions. Despite the emergence of living systematic reviews (Elliott et al., 2017), relatively little computational work has targeted the update process.

Existing computational work on updates includes study identification (Alharbi and Stevenson, 2019), document classification for scheduling (Cohen et al., 2012), and relevant study retrieval (Khabsa et al., 2016). To the best of our knowledge, the only previous work on conclusion change prediction is Bashir et al. (2019, 2021) who trained tree-based classifiers on four metadata features and evaluated on up to 256 systematic review pairs without considering textual features.

Language Models for Biomedical Reasoning LLMs have achieved strong performance on biomedical QA and claim verification (Singhal et al., 2023; Jin et al., 2019; Wadden et al., 2020), and domain-specific pretraining has further improved biomedical NLP (Luo et al., 2022; Gu et al., 2021). Techniques such as LoRA (Hu et al., 2022) enable efficient adaptation.

3 Task Definition and Dataset

Task Formulation Conclusion change prediction is defined as a binary classification task. Each instance pairs the abstract of a Cochrane systematic review with the abstracts of publications newly included in the subsequent version of the same review. Specifically, given r, the abstract of a systematic review, and $P = \{p_1, p_2, \ldots, p_n\}$, a set of n abstracts of new publications relevant to the review, the task is to predict whether the review's conclusions were revised (CHANGED) or not (UNCHANGED).

This approach deliberately isolates the evidence retrieval and reasoning components of the review update task. Identifying relevant new studies is a well-studied problem with established semi-automated tools in routine use (Cohen et al., 2006; Wallace et al., 2010; Marshall and Wallace, 2019). In the standard Cochrane updating workflow, review authors screen and confirm newly included publications before evaluating their impact on conclusions. Our task begins at that decision point: having the studies, the reviewer must decide whether the conclusions should change. The intended use is human-in-the-loop triage, not final determination.

Data Source The dataset is derived from intervention reviews written in English in the Cochrane Database of Systematic Reviews.[1] Inclusion required at least two versions of the review to be available where one is designated as an update of the other and includes at least one additional publication. The dataset was restricted to intervention reviews, excluding other types such as diagnostic test accuracy reviews and overviews, since the conclusion-change task is defined over intervention effects. Protocols were also excluded, following Bashir et al. (2021). These criteria yielded 3,326 review-update pairs from 2,485 unique reviews published between 1995 and 2021.

Label Derivation Gold labels were derived from the structured "What's New" section that Cochrane requires in every review update. Entries marked as *new conclusions* were labelled CHANGED; entries marked as *old conclusions* were labelled UNCHANGED. This binary formulation was not introduced for modelling convenience; rather, it follows the standard conclusion-status categories recorded in Cochrane update records. Ambiguous entries (recorded as UPDATE or AMENDMENT) were resolved using a rule-based procedure (Bashir et al., 2021). Reviews for which no conclusion status could be determined were excluded. Comparison of extracted labels against 300 randomly sampled instances revealed complete agreement.

Identifying New Publications For each review-update pair, newly included publications were identified by comparing reference lists of consecutive systematic review versions. PMIDs were extracted from Cochrane XML where available; other-

[1] https://www.cochranelibrary.com/cdsr/about-cdsr

wise, title-based PubMed queries with Levenshtein matching (Levenshtein et al., 1966) were used. Similarity thresholds (0.75; fallback 0.66) were selected empirically. A 300-publication evaluation sample was checked against gold-standard PMIDs (manual PubMed search by first author; subset verified by second author). Agreement was 95%; all disagreements were "No PMID" returns rather than misassignments, indicating retained publications were correctly mapped.

Dataset Characteristics and Split Table 1 summarises the dataset. The class distribution is moderately imbalanced (61.6% Unchanged).

Statistic	Value
Total review-update pairs	3,326
Unique reviews (DOIs)	2,485
Conclusion changed	1,278 (38.4%)
Conclusion unchanged	2,048 (61.6%)
Training set	2,648
Test set	678
Training DOIs	1,988
Test DOIs	497
New publications per update	
Median (IQR)	5 (2–11)
Range	1–301
Review abstract length, mean ± SD	433 ± 168 words
Time between search dates, median (IQR)	1,936 (1,291–2,848) days

Table 1: Dataset statistics. The train/test split is stratified by review DOI to prevent information leakage.

4 Methods

We evaluate methods spanning multiple paradigms: zero-shot inference, few-shot prompting, and parameter-efficient fine-tuning. All methods are evaluated on the same held-out test set. Prompts are provided in Appendix A.

Model selection was constrained by input length, a single A100 80 GB GPU with 4-bit quantisation, and the need for architectural diversity. A 16,384-token maximum accommodated 95.4% of instances without truncation (see Appendix B). These constraints excluded models above 70B parameters and most biomedical LLMs, whose context windows (typically $\leq$4,096 tokens) are insufficient for multi-document inputs. We fine-tuned Qwen2.5-14B-Instruct (Yang et al., 2025b), Qwen3-14B (Yang et al., 2025a), Llama-3.1-8B (Grattafiori et al., 2024), and OpenBioLLM-8B (Pal and Sankarasubbu, 2024),[2] representing two model families and two scales, with OpenBioLLM-8B providing biomedical pretraining, and included GPT-4.1-mini as a proprietary comparison.

[2]Model checkpoints: `Qwen2.5-14B-Instruct`, `Qwen3-14B`, `Llama-3.1-8B-Instruct`, `Llama3-OpenBioLLM-8B`.

System

```
You are an expert biomedical evidence analyst.
Your job is to decide if the conclusions of
a systematic review (SR) should be updated in
light of new trial evidence.
The publication abstracts provided below have
already been screened by expert reviewers and
confirmed as relevant to this review.  All
publications meet the inclusion criteria.
Reply with EXACTLY ONE TOKEN:
 - N0 = the overall conclusions stay essentially
unchanged
 - N1 = the conclusions would plausibly need to
be changed
```

User

```
SYSTEMATIC REVIEW AND NEW TRIAL ABSTRACTS:
{review_abstract}
{publication_abstracts}
Decision:
```

Assistant

`N0` or `N1` (gold label during training; omitted at inference)

Figure 2: Prompt template used for fine-tuning. Zero-shot prompt variants (P-01, P-02, P-03) use different system messages and are provided in Appendix A.

4.1 Zero-Shot Inference

Zero-shot methods receive the review abstract and concatenated publication abstracts as input, with no task-specific training or in-context examples. We evaluate Qwen2.5-14B-Instruct (Yang et al., 2025b) with three prompt variants: P-01 (standard task description), P-02 (augmented with base-rate information and debiasing instructions), and P-03 (tiered reasoning: first identify conflicting evidence, then decide). Each instructs the model to output a single token.

4.2 Few-Shot Prompting

We evaluate Qwen2.5-14B-Instruct with three in-context demonstrations using P-01, with exemplars randomly sampled from the training set.

4.3 Fine-Tuning

We fine-tuned all four models using QLoRA (Dettmers et al., 2023). Each instance was formatted as a three-turn conversation (see Figure 2). The base model was quantised to 4-bit NormalFloat with double quantisation. LoRA adapters (Hu et al., 2022) ($r = 16$, $\alpha = 32$, dropout 0.05) were applied to all projections. The maximum sequence length was 16,384 tokens. Training ran for 3 epochs with batch size 8,

learning rate 2×10^{-4}, cosine scheduling, on an A100 80GB using SFTTrainer. For Qwen3-14B, thinking mode was disabled because our logit-based probability extraction requires the decision token immediately after the prompt, which is incompatible with interposed reasoning tokens. Hyperparameters follow standard QLoRA defaults (Dettmers et al., 2023).

Probability extraction. The tokeniser encodes possible outputs, `N0` and `N1`, as two-token sequences, `N+0` and `N+1`. We append `N` and extract logits for the `0` and `1` continuation tokens, applying softmax to obtain $P(\text{CHANGED})$. Unparseable responses are treated as UNCHANGED.

Decision threshold. We report results at threshold 0.50 and threshold 0.37 (selected by a 0.01-step sweep on the training set to maximise balanced accuracy). AUC-ROC is reported once.

5 Evaluation

All methods are evaluated on the held-out test set ($n = 678$; 420 Unchanged, 258 Changed). We report accuracy, balanced accuracy, AUC-ROC, and per-class metrics with 95% bootstrap CIs (2,000 resamples). Balanced accuracy is emphasised, given the class imbalance, as is Changed-class recall due to potential negative impacts of missed updates.

Baselines We reimplemented Bashir et al. (2021) on our dataset using their four features (trial count, participant count, coverage score, time elapsed) and three classifiers (logistic regression, decision tree, random forest) with cross-validation grouped by DOI. Missing values were median-imputed on the training set. Random forest achieved the best performance. Two features (coverage score and time elapsed) require information from the completed update, since time elapsed is computed from the update search date, making this baseline a post-hoc upper bound.

Two simple baselines establish lower bounds: majority class (i.e. always predicts UNCHANGED) and random (50/50).

6 Results

Table 2 presents classification performance across all methods.

The majority class baseline achieves 50.0% balanced accuracy but fails to identify any Changed reviews. The reimplemented Bashir et al. (2021) random forest achieves 63.6% AUC-ROC; the lower performance, relative to their reported 80.8% accuracy, reflects our dataset's higher change rate (38.4% vs. 16.8%).

Zero-Shot LLMs No zero-shot Qwen2.5-14B configuration exceeds 56% balanced accuracy (lower CI bounds 50.8–51.8%). P-02 overcorrects: Changed recall rises to 83.3% but Unchanged recall collapses to 24.8%. GPT-4.1-mini under P-02 reaches 59.5% balanced accuracy and 63.4% AUC-ROC, still 7 points below the fine-tuned models, confirming that the task difficulty is not specific to open-source models.

Few-Shot Prompting Three in-context demonstrations yield 54.3% balanced accuracy and 59.0% AUC-ROC, below the best Qwen2.5-14B zero-shot AUC (61.1%), suggesting the difficulty lies in evidence reasoning rather than task-format understanding.

Fine-Tuning Performance Fine-tuning produces the strongest results. Qwen2.5-14B achieves AUC-ROC 70.4%, followed by Qwen3-14B (69.5%), OpenBioLLM-8B (67.5%), and Llama-3.1-8B (66.9%), all exceeding Bashir et al. (2021) (63.6%). The difference between Qwen2.5-14B and the random forest baseline is significant (McNemar's $p = 0.0015$). At threshold 0.37, Qwen2.5-14B achieves 65.7% balanced accuracy with 63.6% Changed recall and 67.9% specificity.

At matched scale, OpenBioLLM-8B (67.5%) exceeds Llama-3.1-8B (66.9%), indicating a modest gain from biomedical pretraining. The gap to Qwen2.5-14B (70.4%) is therefore primarily attributable to scale rather than domain mismatch. Context length is a secondary factor: OpenBioLLM's 8,192-token pretraining range is exceeded by most instances with 21+ publications (Appendix B), whereas the 14B Qwen models natively support our 16,384-token input. Disentangling these effects would require a matched biomedical 14B model with equivalent context, which is not currently available.

7 Influence of Publication Count

At threshold 0.37, the fine-tuned model produces 164 true positives, 285 true negatives, 94 false negatives, and 135 false positives. False negatives involved significantly fewer new publications than true positives (median 3 vs. 16; Mann–Whitney $U = 1{,}794$, $p < 0.001$), while false positives involved more than true negatives (median 12 vs. 2; $U = 35{,}630$, $p < 0.001$): the model over-predicts

Table 2: Test-set performance ($n = 678$; 420 Unchanged, 258 Changed), reported as % with 95% bootstrap CIs (2,000 resamples). †Threshold selected on training data to maximise balanced accuracy; others use the default boundary. **Bold**: best standard fine-tuned model; underline: best zero-shot.

		Overall						Unchanged			Changed		
Method	**Model**	**Acc**	**BalAcc**	**AUC**	**P_M**	**R_M**	**$F1_M$**	**P_U**	**R_U**	**$F1_U$**	**P_C**	**R_C**	**$F1_C$**
Simple Baselines													
Majority class	—	61.95	50.00	50.00	30.98	50.00	38.24	61.95	100.0	76.50	0.00	0.00	0.00
Random (50/50)	—	50.00	50.00	50.00	50.00	50.00	43.22	61.95	50.00	55.37	38.05	50.00	43.22
Prior Work (Bashir et al., 2021; reimplemented on our dataset)													
LR	4 features	61.21 [57.5,64.9]	59.50 [55.6,63.5]	60.43 [55.8,64.9]	59.28 [55.5,63.2]	59.50 [55.6,63.5]	59.35 [55.5,63.2]	69.48 [65.0,73.7]	66.67 [62.3,71.3]	68.04 [64.2,71.7]	49.09 [43.4,55.3]	52.33 [46.3,58.4]	50.66 [45.4,56.1]
DT	4 features	55.75 [51.9,59.6]	59.28 [55.5,62.9]	59.76 [55.6,63.9]	59.33 [55.5,62.9]	59.28 [55.5,62.9]	55.75 [51.9,59.5]	73.62 [67.8,78.9]	44.52 [39.6,49.1]	55.49 [50.7,59.8]	45.05 [40.8,49.9]	74.03 [68.4,79.3]	56.01 [51.8,60.4]
RF	4 features	59.00 [55.3,62.8]	59.65 [55.9,63.6]	63.62 [59.4,68.0]	59.10 [55.5,62.8]	59.65 [55.9,63.6]	58.45 [54.7,62.3]	71.13 [66.2,75.9]	56.90 [52.1,61.7]	63.23 [59.1,67.1]	47.08 [42.2,52.5]	62.40 [56.7,68.2]	53.67 [49.1,58.5]
This Work — Zero-Shot General LLMs													
Standard (P-01)	Qwen2.5-14B	52.51 [48.7,56.2]	55.69 [51.8,59.4]	60.53 [56.2,64.9]	55.69 [51.8,59.4]	55.69 [51.8,59.4]	52.51 [48.6,56.2]	68.99 [63.0,74.8]	42.38 [37.7,47.1]	52.51 [47.6,57.0]	42.38 [37.5,47.2]	68.99 [63.1,74.6]	52.51 [47.7,56.9]
Debiased (P-02)	Qwen2.5-14B	47.05 [43.2,50.7]	54.05 [50.9,57.1]	61.10 [56.7,65.4]	55.62 [51.2,59.8]	54.05 [50.9,57.1]	45.59 [41.7,49.2]	70.75 [62.9,78.2]	24.76 [20.6,29.0]	36.68 [31.5,41.7]	40.49 [36.4,44.6]	83.33 [78.8,87.8]	54.50 [50.2,58.4]
Tiered (P-03)	Qwen2.5-14B	54.42 [50.6,58.3]	54.54 [50.8,58.4]	58.56 [54.3,62.8]	54.28 [50.7,58.0]	54.54 [50.8,58.4]	53.70 [49.9,57.5]	66.18 [61.2,71.1]	54.05 [49.4,58.8]	59.50 [55.3,63.5]	42.39 [36.9,47.8]	55.04 [48.7,61.3]	47.89 [42.8,52.6]
Standard (P-01)	GPT-4.1-mini	55.60 [51.9,59.3]	57.96 [54.2,61.6]	62.86 [58.6,67.1]	57.70 [54.1,61.2]	57.96 [54.2,61.6]	55.53 [51.8,59.2]	70.88 [65.2,75.9]	48.10 [43.4,53.1]	57.30 [52.8,61.7]	44.53 [39.8,49.4]	67.83 [61.9,73.5]	53.76 [49.1,58.1]
Debiased (P-02)	GPT-4.1-mini	52.36 [48.7,56.2]	59.53 [56.7,62.4]	63.35 [59.1,67.6]	62.98 [59.3,66.8]	59.53 [56.7,62.4]	51.14 [47.5,55.0]	82.12 [75.7,88.1]	29.52 [25.5,33.9]	43.43 [38.5,48.4]	43.83 [39.7,48.0]	89.53 [85.8,93.3]	58.85 [54.6,62.9]
Tiered (P-03)	GPT-4.1-mini	60.62 [56.6,64.5]	56.85 [53.2,60.6]	61.83 [57.6,66.2]	57.35 [53.4,61.4]	56.85 [53.2,60.6]	56.91 [53.0,60.8]	66.74 [62.2,71.1]	72.62 [68.0,76.9]	69.56 [65.8,72.9]	47.96 [41.4,54.7]	41.09 [35.3,47.2]	44.26 [38.8,49.8]
This Work — Few-Shot													
3-shot random	Qwen2.5-14B	51.03 [47.3,54.7]	54.27 [50.4,57.9]	58.96 [54.9,63.4]	54.30 [50.4,57.9]	54.27 [50.4,57.9]	51.03 [47.3,54.7]	67.32 [61.4,73.0]	40.71 [36.1,45.4]	50.74 [46.2,55.1]	41.27 [36.6,46.0]	67.83 [61.8,73.5]	51.32 [46.6,55.9]
This Work — Fine-Tuned													
FT	Qwen2.5-14B+QLoRA	**68.44** [64.7,72.1]	63.83 [60.3,67.4]	**70.36** [66.3,74.3]	**66.38** [62.3,70.6]	63.83 [60.3,67.4]	64.17 [60.3,68.0]	70.93 [66.7,75.0]	83.10 [79.4,86.7]	**76.54** [73.3,79.5]	**61.83** [54.4,69.0]	44.57 [38.4,50.6]	51.80 [46.0,57.3]
FT†	Qwen2.5-14B+QLoRA	66.22 [62.5,69.6]	**65.71** [61.9,69.3]	—	65.02 [61.3,68.4]	**65.71** [61.9,69.3]	**65.11** [61.2,68.5]	**75.20** [70.6,79.3]	67.86 [63.5,72.3]	71.34 [67.7,74.7]	54.85 [49.2,60.4]	**63.57** [57.4,69.2]	**58.89** [53.7,63.5]
FT	Qwen3-14B+QLoRA	67.55 [63.9,70.9]	62.67 [59.0,66.1]	69.49 [65.3,73.4]	65.32 [61.0,69.4]	62.67 [59.0,66.1]	62.90 [58.9,66.6]	70.08 [66.1,74.0]	83.10 [79.5,86.7]	76.03 [72.9,79.0]	60.56 [52.8,67.6]	42.25 [36.3,48.4]	49.77 [43.8,55.4]
FT†	Qwen3-14B+QLoRA	65.34 [61.7,68.9]	64.17 [60.3,67.8]	—	63.74 [60.0,67.2]	64.17 [60.3,67.8]	63.86 [60.0,67.4]	73.42 [69.2,77.7]	69.05 [64.9,73.6]	71.17 [67.6,74.6]	54.06 [48.2,59.9]	59.30 [53.4,65.2]	56.56 [51.3,61.1]
FT	Llama-3.1-8B+QLoRA	62.83 [59.3,66.4]	52.13 [50.4,54.0]	66.86 [62.5,70.8]	61.19 [52.2,69.9]	52.13 [50.4,54.0]	44.73 [41.7,48.1]	63.00 [59.3,66.6]	96.90 [95.2,98.5]	76.36 [73.5,79.1]	59.38 [42.3,76.0]	7.36 [4.3,10.8]	13.10 [7.9,18.5]
FT†	Llama-3.1-8B+QLoRA	64.75 [60.9,68.4]	57.79 [54.6,61.1]	—	61.92 [57.2,66.5]	57.79 [54.6,61.1]	56.79 [52.7,60.8]	66.48 [62.5,70.4]	86.90 [83.6,90.0]	75.34 [72.2,78.3]	57.36 [48.6,65.4]	28.68 [23.2,34.3]	38.24 [31.8,44.2]
FT	OpenBioLLM-8B+QLoRA	66.22 [62.5,69.9]	60.33 [56.8,63.9]	67.50 [63.2,71.5]	63.81 [59.2,68.3]	60.33 [56.8,63.9]	60.13 [55.9,64.2]	68.26 [64.3,72.2]	85.00 [81.5,88.4]	75.72 [72.5,78.8]	59.35 [51.2,67.1]	35.66 [29.5,41.8]	44.55 [38.0,50.6]
FT†	OpenBioLLM-8B+QLoRA	65.78 [62.1,69.3]	63.93 [60.1,67.5]	—	63.81 [59.9,67.4]	63.93 [60.1,67.5]	63.87 [60.0,67.4]	72.71 [68.3,76.9]	71.67 [67.4,75.9]	72.18 [68.5,75.6]	54.92 [48.8,60.8]	56.20 [50.2,62.1]	55.56 [50.2,60.4]

change when many publications are present and under-predicts when few are present.

Overall calibration is good (Expected Calibration Error = 0.028, 10 bins), so the probabilities reflect genuine confidence rather than calibration artefacts. However, predictions above 0.80 are poorly calibrated (19 instances; 58% accuracy vs. 91% mean confidence), consistent with overconfidence in the high-publication regime.

Table 3 stratifies Changed-class recall and Unchanged-class specificity by the number of new publications. When only 1–2 new publications are available, the model detects just 18.4% of conclusion changes. Recall rises monotonically with publication count, reaching 95.6% for instances with 21 or more publications, while specificity shows the inverse pattern, dropping from 95.9% to 2.9%. This indicates that the model uses publication count as a proxy for conclusion change, rather than reasoning about evidence content. The predicted probability $P(\text{CHANGED})$ correlates with publication count at Spearman $\rho = 0.80$, far exceeding the true association between publication count and the label ($\rho = 0.30$). Thus, the model has amplified a moderate real association into a dominant decision rule.

Pubs.	*n*	Changed	Recall	Spec.
1–2	219	49	18.4	95.9
3–5	149	50	30.0	87.9
6–10	115	47	72.3	47.1
11–20	92	44	93.2	4.2
21+	103	68	95.6	2.9
All	678	258	63.6	67.9

Table 3: Changed-class recall and Unchanged-class specificity of the fine-tuned model (threshold 0.37), stratified by the number of new publications.

Although publication count is never an explicit feature, the concatenated input exposes it through two channels: positional encoding of the prediction token (which scales with input length) and textual cues such as repeated abstract structures across publications. To isolate the positional channel, we repeated inference with all inputs left-padded to 16,384 tokens, eliminating variation in the prediction token's position. The correlation was virtually unchanged ($\rho = 0.78$ padded vs. $\rho = 0.80$ unpadded; Pearson $r = 0.99$ between probability vectors), confirming that the model infers volume from textual cues rather than from sequence position alone. These cues may be reinforced by attention patterns that concentrate on tokens near sequence boundaries (Liu et al., 2024), though we do not test this directly.

7.1 Publication Count Ablation at Inference

To test whether performance depends on evidence volume or content, we provide only the first k publication abstracts at inference; all other settings are identical. If the model reasons about content, performance should be largely preserved at small k. Table 4 presents the results. At k=1, AUC-ROC drops from 70.4% to 53.4%, barely above chance, and Changed-class recall falls to 5.8%. Performance improves monotonically with k, and the shortcut correlation ρ rises from -0.26 at k=1 to 0.80 at k=All. These results establish a relationship between evidence volume and model performance: this shortcut is the primary driver of the fine-tuned model's discrimination.

k	AUC-ROC	BalAcc	Chg Rec	Spec	ρ
1	53.4 [48.9, 57.8]	51.7	5.8	97.6	−0.26
3	59.3 [54.9, 63.6]	53.7	12.4	95.0	0.05
5	62.3 [58.0, 66.4]	54.5	20.9	88.1	0.36
10	68.9 [64.7, 72.9]	64.7	59.3	70.0	0.73
All	70.4 [66.3, 74.3]	65.7	63.6	67.9	0.80

Table 4: Publication count ablation at inference. The fine-tuned model (Qwen2.5-14B, threshold 0.37) is evaluated with only the first k publication abstracts provided at inference; all other settings are identical.

7.2 Publication Subsampling

To explore the effect of removing publication count information, we trained Qwen2.5-14B with counterfactual publication subsampling at two levels: k=1 and k=3 publications per training instance. For each training instance, k publications were randomly sampled (or all, when fewer than k existed), decoupling publication count from the label during training; at inference all publications were used, leaving the test distribution unchanged. Other hyperparameters were unchanged.

Table 5 presents the results alongside the standard model (k=All). A clear trade-off emerges across three dimensions: overall discrimination (AUC-ROC), shortcut strength (ρ), and low-publication recall.

At k=1, where each training instance contains a single publication abstract, the shortcut is nearly eliminated: ρ drops from 0.80 to 0.15, below even the true label–publication count association

Method	AUC	ρ	1–2 Rec	BalAcc
FT k=1 subsample	55.4	0.15	57.1	52.0
FT k=3 subsample	63.5	0.36	46.9	59.4
FT k=All (standard)	70.4	0.80	18.4	65.7

Table 5: Shortcut–discrimination trade-off. AUC-ROC (%), Spearman ρ between predicted probability and publication count, Changed-class recall (%) at 1–2 publications, and balanced accuracy (%) at threshold 0.37 (selected on the training set).

($\rho = 0.30$). Changed-class recall at 1–2 publications rises from 18.4% to 57.1%, a threefold improvement in the clinically critical low-publication regime. However, overall AUC-ROC falls to 55.4%, barely above chance, indicating the model has lost the volume signal without learning to compensate through content-based reasoning.

At k=3, the trade-off is more balanced: $\rho = 0.36$ (close to the true association), 1–2 publication recall improves to 46.9%, and AUC-ROC remains at 63.5%, comparable to the Bashir et al. random forest baseline (63.6%).

8 Conclusion

We introduced conclusion change prediction as a multi-document evidence reasoning task on 3,326 Cochrane review-update pairs. Fine-tuning achieves the strongest overall discrimination (AUC-ROC 70.4%) but analysis indicates that it relies heavily on an evidence volume shortcut which future work should aim to overcome.

Limitations

The dataset is restricted to Cochrane intervention reviews with binary labels. Models are provided with abstracts, keeping the system broadly deployable given uneven full-text access. Publications were identified retrospectively and retrieval noise in deployment would likely degrade accuracy. Fine-tuning used a single A100 80 GB GPU with 4-bit quantisation, and 154 long-input instances (Appendix B) were right-truncated. Single training run per configuration.

Ethics Statement

All information required to reproduce our work is contained in this paper: dataset construction and label derivation procedures are described in Section 3; model training configurations and hyperparameters in Section 4.3; the full text of all prompts in Appendix A; and input length and truncation details in Appendix B. Researchers with Cochrane Library access can reconstruct the dataset and reproduce the experiments from these specifications.

Acknowledgements

We gratefully acknowledge the Cochrane Library for providing access to the systematic review data used in this work.

References

Amal Alharbi and Mark Stevenson. 2019. A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1257–1260.

Rabia Bashir, Adam G Dunn, and Didi Surian. 2021. A rule-based approach for automatically extracting data from systematic reviews and their updates to model the risk of conclusion change. *Research Synthesis Methods*, 12(2):216–225.

Rabia Bashir, Didi Surian, and Adam G Dunn. 2019. The risk of conclusion change in systematic review updates can be estimated by learning from a database of published examples. *Journal of Clinical Epidemiology*, 110:42–49.

Aaron M Cohen, Kyle Ambert, and Marian McDonagh. 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC Medical Informatics and Decision Making*, 12(1):33.

Aaron M Cohen, William R Hersh, Kim Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Miranda Cumpston and Ella Flemyng. 2023. Chapter IV: Updating a review. In Julian P. T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions, version 6.4*. Cochrane.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. *Advances in Neural Information Processing Systems*, 36.

Julian H Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, and 1 others. 2017. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91:23–30.

Paul Garner, Sally Hopewell, Jackie Chandler, Harriet MacLehose, Elie A Akl, Joseph Beyene, Stephanie Chang, Rachel Churchill, Karin Dearness, Gordon Guyatt, and 1 others. 2016. When and how to update systematic reviews: consensus and checklist. *BMJ*, 354.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Diana Herrera-Perez, Alyson Haslam, Tyler Crain, Jennifer Gill, Catherine Livingston, Victoria Kaestner, Michael Hayes, Dan Morgan, Adam S Cifu, and Vinay Prasad. 2019. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *eLife*, 8:e45183.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic Reviews*, 4(1):78.

Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B. Hugging Face model repository.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How quickly do systematic reviews go out of date? a survival analysis. *Annals of Internal Medicine*, 147(4):224–233.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Rens Van De Schoot, Jonathan De Bruin, Raoul Schram, Parisa Zahedi, Jan De Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, and 1 others. 2021. An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11(1):55.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2025b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

A Prompt Templates

This appendix provides the full text of all prompts used in our experiments. All prompts instruct the model to output a single token: `N0` (unchanged) or `N1` (changed). The user message is identical across all variants and consists of the review abstract followed by the concatenated publication abstracts.

A.1 Zero-Shot Prompts (Section 4.1)

P-01 (Standard). This prompt provides a straightforward task description with no additional guidance.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

P-02 (Debiased). This prompt adds base-rate information and an explicit instruction to counteract majority-class bias.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

IMPORTANT: In this dataset, approximately 38% of systematic reviews have changed conclusions upon updating. Do NOT default to predicting unchanged. Carefully assess whether the new evidence warrants a change.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

P-03 (Tiered). This prompt asks the model to reason in two explicit steps before making a prediction.

```
You are an expert biomedical evidence analyst. You will be given the abstract of a Cochrane systematic review and the abstracts of publications that have been newly included in the review since its last update.

Your task: predict whether the review's conclusions changed in the update.

Reason in two steps:
Step 1: Assess whether any new publication reports findings that conflict with, contradict, or substantially extend the review's existing conclusions.
Step 2: Based on your assessment, decide whether the overall conclusions would need to change.

Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

A.2 Fine-Tuning Prompt (Section 4.3)

The fine-tuning prompt is shown in Figure 2 in the main text. It is reproduced verbatim below for completeness.

```
You are an expert biomedical evidence analyst. Your job is to decide if the conclusions of a systematic review (SR) should be updated in light of new trial evidence. The publication abstracts provided below have already been screened by expert reviewers and confirmed as relevant to this review. All publications meet the inclusion criteria. Reply with EXACTLY ONE TOKEN:
- N0 = the overall conclusions stay essentially unchanged
- N1 = the conclusions would plausibly need to be changed
```

The user message follows the template:

```
SYSTEMATIC REVIEW AND NEW TRIAL ABSTRACTS:
{review_abstract}
{publication_abstracts}
Decision:
```

During training, the assistant message contains the gold label token (`N0` or `N1`). During inference, the assistant message is omitted and the model generates its prediction.

A.3 Few-Shot Prompt (Section 4.2)

The few-shot prompt uses the same system message as P-01 (Standard). Three training examples

are prepended to the user message as demonstrations. Each demonstration consists of a review–publication pair followed by the gold label.

```
Here are three examples of this task:

– Example 1 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_1_text}
Decision: N1

– Example 2 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_2_text}
Decision: N0

– Example 3 –
SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{example_3_text}
Decision: N0

Now predict for this instance:

SYSTEMATIC   REVIEW   AND   NEW   TRIAL
ABSTRACTS:
{test_instance_text}
Decision:
```

For each test instance, the three exemplars are randomly sampled from the training set.

The user message is identical to all other methods. Probability estimates were obtained by extracting logit values for the decision token from the classification call and normalising via softmax.

B Input Length and Truncation

Table 6 reports the distribution of input token counts and truncation at the 16,384-token limit, stratified by the number of new publications per instance. Truncation is concentrated at the high-publication tail.

Pubs.	n	Median tok.	Truncated	% Trunc.
1–2	1,055	1,254	0	0.0
3–5	757	2,269	0	0.0
6–10	625	3,903	0	0.0
11–20	482	6,699	1	0.2
21–50	329	12,819	75	22.8
51+	78	31,320	78	100.0
All	3,326	2,738	154	4.6

Table 6: Input length and truncation at the 16,384-token limit, stratified by the number of new publications. "% Trunc." is the proportion of instances in the row whose tokenised length exceeds 16,384 tokens, at which point later publications are right-truncated. All instances with 51 or more new publications exceed the limit; none with 10 or fewer do. Overall, 95.4% of instances fit within the limit without truncation.