CLARIFY: Concept Level Active-Learning Ranker Interpreter for Systematic Reviews

Jelle Jasper Teijema[1][0000−0001−9282−4311] and Ayoub Bagheri[1][0000−0001−6366−2173]

Methodology and Statistics
Social and Behavioral Sciences
Utrecht University, The Netherlands

Abstract. Transformer-based ranking models have recently advanced active-learning tools for accelerating systematic reviews, but the internal criteria they use to rank documents are opaque, limiting their utility in scientific decision making. We introduce CLARIFY, a post-hoc explainability method for active-learning applications that (i) automatically derives high-level concepts from the model's embedding space, (ii) quantifies each concept's influence on ranking, and (iii) links the most influential concepts to sentences via occlusion and re-projection without retraining or manual intervention. Evaluated on three SYNERGY systematic review datasets, CLARIFY uncovers latent concepts and represents them in a human understandable manner. Similarity metrics show discernible relations between these concepts and elements of the inclusion criteria. By making model reasoning transparent, CLARIFY supports accountable, evidence-based decision-making in systematic-review screening. Our open-source work can be found on GitHub/Zenodo.

Keywords: Explainable Active learning · Systematic review · XAI · Concept activation vectors · ASReview · Interpretability.

1 Introduction

Machine learning, and more specifically Active learning (AL), is rapidly gaining ground in the field of systematic reviews as the key approach for semi-automating the screening phase of systematic reviews [24, 25]. As the volume of published literature continues to grow exponentially, the need for systems that assist with managing this information overload becomes increasingly urgent. Traditionally, these systems have been treated as black boxes, with most available software being closed-source[1]. The software provides a ranking or suggestions but do not explain why a given reference was deemed relevant or not.

Researchers require insight into model decision-making processes in order to understand how predictions are made, build trust in the model, and make

[1] github.com/Rensvandeschoot/software-overview-machine-learning-for-screening-text

informed decisions based on its outputs. Transparency, accountability, and interpretability remain essential, especially in domains such as health, where inclusion decisions must be both defensible and reproducible. Opacity can hinder trust and adoption of AI assistance [18]. Reviewers are understandably cautious about relying on a model's inclusions and exclusions without understanding the model's reasoning.

In the traditional screening phase, human reviewers manually assess each abstract for relevance, selecting studies for inclusion in the next stages of analysis. Machine learning tools like ASReview[2] aim to reduce this burden by predicting which documents are most likely to meet the review's inclusion criteria, using an iterative process called active learning [2]. When simple models are used, for example TF-IDF (term frequency-inverse document frequency) and regression models, interpretability is relatively straightforward: one can trace decisions to specific words or combinations thereof [21]. This helps the reviewer understand and justify the tool's behavior, which is vital in contexts where methodological rigor is non-negotiable and lack of transparency is one of the main barriers to implementation [15].

However, recent developments in the field of active learning for systematic reviews show that more complex natural language processing (NLP) models, such as transformers (large language models, LLMs), can no longer be ignored for their performance [23]. These models better capture the nuances of human decision-making but do so in ways that are inherently opaque. Their logic exists in high-dimensional spaces, often unaligned with symbolic reasoning used by humans. This presents a fundamental challenge: if researchers are to remain accountable for the inclusion and exclusion decisions made with the aid of machine learning, they need a way to understand the basis for those decisions, even if that understanding comes after the fact.

Explainable AI techniques have been shown to complement active learning workflows effectively. A general framework of Explainable Active Learning (XAL) exists [6], in which local explanations are provided during the annotation process to improve annotator understanding and model trust. This framework demonstrates that explanations can enhance user engagement and decision confidence in iterative labeling tasks. However, the work also cautions that explanations may introduce cognitive biases, emphasizing the need for carefully designed, domain-sensitive interpretation methods. While these findings are not situated in the context of systematic reviews, they suggest that integrating explainability into active learning can be beneficial, particularly when model decisions carry scientific or clinical weight. To fill this gap, in the current study, we propose a new method for post-hoc, concept-based explanation of models used in systematic review screening software, namely CLARIFY. Our goal in CLARIFY is to create a tool for decomposing the internal representations of neural networks, so that it can provide insight into the decision-making process of automated active learning-based screening tools such as ASReview. Rather than requiring retraining or architectural changes, CLARIFY operates after the

[2] asreview.ai/

model has completed its predictions, making it well-suited for integration into established review workflows. The contributions in this study are to:

1. introduce CLARIFY, the first Explainable Active Learning pipeline for systematic review screening.
2. provide a pipeline that integrates feature extractors and classification models with concept activation vectors, enabling human-readable concept scores without the need for retraining or architectural changes.
3. reuse the active-learning state directly within the explanation pipeline
4. conduct experiments on multiple review datasets to demonstrate the generalizability of the approach.

The rest of the paper is structured as follows. Section 2 explores related work on active learning for systematic reviews and explainable machine learning. Section 3 details the proposed CLARIFY pipeline and its integration within existing active learning frameworks such as ASReview. Section 4 reports on the results and Section 5 discusses limitations and future work, and section 6 concludes.

2 Related Work

In the domain of systematic reviews, recent work evaluates the performance of various active learning strategies across a number of review datasets. The results of the study show that the difficulty of applying active learning is not confined to a particular research domain. Instead, the work suggests that a possible explanation for difficulty could be attributed to factors such as the complexity of inclusion criteria used to identify relevant publications [4]. Rathbone et al. [17], as cited in Gates et al. [5], observes that the complexity of inclusion criteria can substantially affect the precision of automated screening tools. In their evaluation of Abstrackr, they note that imprecise population definitions (e.g., "young adults") and reviews structured around multiple key questions poses challenges for automated classification. Gates et al. extend this observation by showing that tasks with broad or heterogeneous criteria (e.g., descriptive analyses with no restriction by intervention or outcome) led to poor specificity and minimal workload savings. Ferdinands et al. [4] suggest that variability in active learning performance may also stem from the complexity of the criteria themselves, even across otherwise comparable domains. These results underscore the importance of well-defined and narrowly scoped inclusion criteria in enabling effective automation—and potentially in making the classifier's logic more interpretable.

More generally, Vilone and Longo [26] provide a comprehensive taxonomy of explainable AI methods and their application domains. Their review emphasizes the distinction between global and local explanations, model-agnostic versus model-specific approaches, and the varying interpretability needs across domains. Although the review does not address active learning, it offers a conceptual framework for situating post-hoc explanation techniques, such as the one proposed here, within the broader XAI landscape.

Jourdan et al. [10] introduce COCKATIEL, a post-hoc, concept-based, model-agnostic explainer for neural text classifiers. It finds latent concepts in final-layer representations, ranks their importance, and maps them to text spans via occlusion, requiring only a non-negative embedding and no retraining.

Based on these developments, this study combines elements of XAL, ASReview's active learning approach, and COCKATIEL into a new framework CLARIFY, which is a new method of explainable AI for systematic review screening optimization. To our knowledge, this is the first explainability method specifically designed for active learning in the context of systematic reviews.

While popular explainability methods such as LIME [19], SHAP [14], or attention-based visualizations have been applied to NLP tasks, they are not well-suited for our setting. First, SHAP and LIME focus on local feature attributions, which are often unstable in high-dimensional embedding spaces and do not yield coherent or reusable semantic structures across documents. Attention-based methods, while attractive due to their direct integration in transformer architectures, have been shown to lack fidelity and can mislead users about model causality [8]. Moreover, these techniques typically provide token-level or word-level attributions, which do not align well with the sentence-level, concept-driven reasoning that systematic reviewers use. Our approach instead emphasizes latent concept discovery, enabling higher-level, reusable explanations that are better suited to capturing structured decision criteria such as inclusion rules.

3 Methods

3.1 CLARIFY Architecture

In an active learning cycle, scientific records are iteratively screened and re-prioritized, as implemented in frameworks such as ASReview. Each abstract is first transformed into an embedded representation via a feature extractor $h(x)$, yielding a matrix A that encodes the semantic features of the record collection. For this proof-of-concept, we employ the `mxbai-embed-large-v1` model as the embedding function. This model has shown good performance across a wide range of models in simulations [13].

Once embedded, the active learning system iteratively trains a classifier $c(x)$, updating the model each time new user labels are provided. At each step, the classifier ranks the remaining unlabeled documents by their estimated relevance, reordering the review queue accordingly. When the user classifies additional records, these are added to the labeled set, and the classifier is retrained. This process continues until the screening task is completed.

To provide insight into this classifier's decision process, we apply the CLARIFY explanation method in three stages. First, we factorize the embedding matrix A using Non-negative Matrix Factorization (NMF), yielding two low-rank matrices: a concept alignment matrix U, and a concept base matrix W. The columns of W are interpreted as latent "concepts" learned from the data, while each row of U quantifies the degree to which a given document aligns with those

concepts. Lee and Seung [12] show that imposing non-negativity produces a **parts-based representation**: each column Wk of W captures a latent concept, and any document vector is reconstructed solely by non-negative mixes of those parts [11, 12]. Non-negativity forces negative alignments toward zero, resulting in sparse representations where each document aligns with only a subset of concepts. This sparsity enhances interpretability by allowing us to disregard concepts with (near) zero alignment, effectively identifying which parts are irrelevant for a given document.

In the second step, we compute the global importance of each discovered concept by perturbing concept activations U directly using Sobol sampling, and measuring the resulting variance in classifier output using Total Sobol indices. This uses the latest trained generation of the classifier $c(x)$ in the active learning cycle of ASReview. These indices quantify the variance in the model output attributable to perturbations in each concept's activation, capturing both direct and interaction effects.

The third step involves estimating local contributions: we assess which parts of an abstract contribute to a document's alignment with each concept. We mask individual sentences and create perturbations of the abstract. For each perturbation, we re-embed the perturbed abstract using the same feature extractor $h(x)$ and project it into the concept space using the fixed concept base W using the NMF transformation. This projection results in a new alignment matrix, revealing the new alignment for each perturbed abstract, minus the alignment for the masked sentence to the concept base W. A strong shift in alignment suggests that the masked sentence is strongly associated with the affected concept. While COCKATIEL performs occlusion at the word level, we found that sentence-level masking is more suitable for systematic reviews. In CLARIFY, we therefore apply this approach, since inclusion and exclusion criteria are often satisfied, or violated, within single, self-contained sentences.

For practical deployment, we propose applying CLARIFY in an on-the-fly fashion: explanations are generated for the document currently at the top of the active learning queue—that is, the next document the screener is expected to assess. While sentence-level occlusion explanations (step three) are recomputed for each document, both the NMF decomposition and the Sobol-based global concept importances can be reused across iterations. This makes CLARIFY efficient enough to be integrated into an interactive screening workflow.

Importantly, concept extraction is performed using only the embeddings of positively labeled documents. Since the top-ranked document is selected by the model as most likely to be relevant, it is most informative to explain its alignment with inclusion-related concepts. Attempting to extract concepts from documents predicted to be irrelevant would shift the focus toward exclusion justification, which is not aligned with how active learning operates in ASReview.

3.2 Dataset Selection

This study uses the SYNERGY dataset, a well-established benchmark in the development and evaluation of automated tools for systematic review screen-

ing [3], for the evaluation of CLARIFY. The dataset comprises 26 independent review datasets, each annotated with clearly defined inclusion and exclusion criteria. This availability of inclusion criteria is a key advantage over many other datasets, allowing for a direct comparison between learned model concepts and explicit domain rules.

From the full set, we select a subset of datasets that demonstrate high classification performance in prior simulations, as seen in Table 1[3]. While strong model performance does not necessarily imply the presence of well-defined criteria, we specifically avoid low-performing datasets, following Ferdinands et al. [4], suggestion that poor performance may result from vague or inconsistently applied inclusion and exclusion rules. By focusing on high-performing datasets, we reduce the likelihood of encountering such issues, thereby increasing the chance that the model's decision-making is based on clearer and more stable criteria. Selection is further refined by assessing the quality and conceptual clarity of the inclusion and exclusion rules for each review. Priority is given to datasets where criteria are clearly interpretable, mutually distinct, and plausibly separable in the text.

Name	Relevant records	Total Records	Topic
Hall_2012 [7]	104	8793	Computer science
Jeyaraman_2020 [9]	96	1175	Medicine
Menon_2022 [16]	74	975	Medicine

Table 1: Datasets used for the evaluation of CLARIFY

3.3 Implementation

The CLARIFY explainability method was adapted to operate within the ASReview framework. While the original implementation is based on the PyTorch ecosystem, ASReview incorporates a range of models and utilities implemented in both scikit-learn and TensorFlow, necessitating cross-framework integration. To address this, a hybrid pipeline was developed that extracts final-layer embeddings from ASReview, formats them for compatibility with the decomposition and attribution modules, and returns sentence-level explanations for the top-ranked documents. All code, along with configuration files, results, and documentation, is published openly via GitHub and archived on Zenodo to ensure transparency and reproducibility.

We refactored CLARIFY into a self-contained ASReview plug-in, replacing the PyTorch code with scikit-learn-compatible components and a lightweight NMF-based concept module. The pipeline now (i) extracts transformer embeddings through `MXBAI`, (ii) normalizes them once via a shared min-max scaler, (iii) restricts concept factorization and Sobol attribution to the positively labeled subset, and (iv) returns sentence-level heat-maps through a fast occlusion routine built directly on ASReview's feature interface.

[3] Full Synergy table available online Full Synergy table available

Figure 1 shows the schematic overview of the final CLARIFY architecture embedded in the ASReview active-learning based system, and the pseudo code view of this work can be found in Appendix .2.

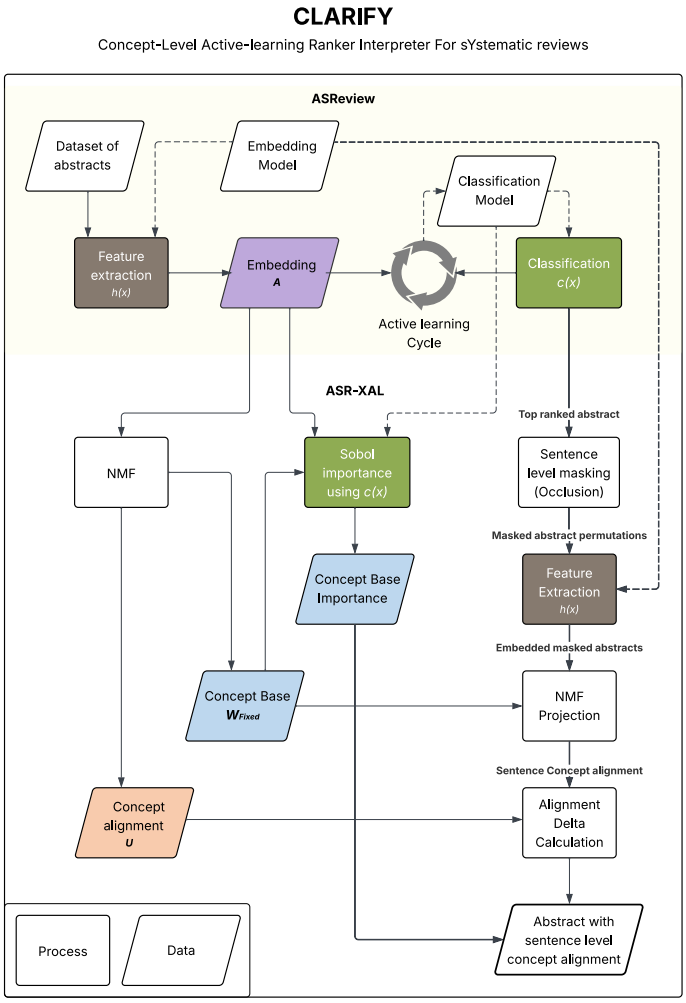
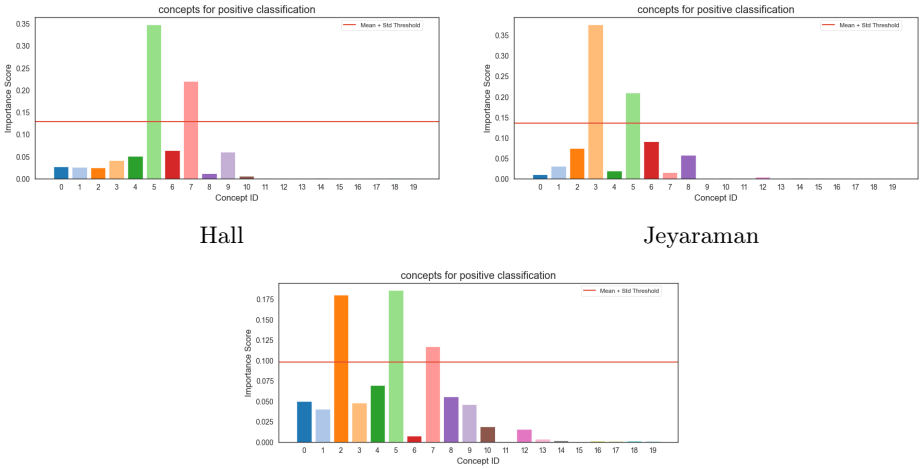


Fig. 1: Overview of CLARIFY explainable active learning pipeline embedded in ASReview. The figure represents the important components of the pipeline using process and data blocks.



Hall Jeyaraman

Menon

Fig. 2: Global concept importance across datasets. Threshold $\mu+\sigma$ shown in red.

4 Results

4.1 Concept Importance and Sentence-Level Highlights

To evaluate whether the CLARIFY yields useful and interpretable outputs in the context of systematic review screening, we applied it to a select subset of high-performing datasets from the SYNERGY benchmark. The central question was whether the discovered concepts extracted from the model's internal representations, can be useful to human reviewers, either by giving insight into the model decision mechanism or by aligning to the inclusion criteria.

We present the results in two stages. First, we present the direct outcomes of the method, including concept importance rankings and highlighted abstracts with sentence-level alignment. These results are shown alongside each dataset's inclusion criteria to support alignment analysis. Second, we present the normalized cosine similarity between the identified concepts and the inclusion criteria, to serve as a quantitative evaluation of the concept usefulness in regards to the inclusion criteria.

Figure 2 shows global concept importance per dataset, computed as total Sobol indices on embeddings of positively labeled records. The red horizontal line marks the selection threshold, defined as the mean plus one standard deviation of the positive-importance distribution ($(\mu+\sigma)$). Concepts above this threshold are used in the later analyses. In these runs, Hall and Jeyaraman each yield two selected concepts; Menon yields three. The threshold is a pragmatic heuristic; other cut-offs (for example a top-quantile rule or a fixed number of important concepts) are also reasonable, providing little difference.

We observe that importance drops sharply after roughly concept 10. We keep all 20 bars visible for transparency, since NMF was run with k=20. In our setup, scikit-learn's NMF initialization uses NNDSVDa, an SVD-based initial-

izer that is energy ordered. On our positive-only data this tends to concentrate mass in early components, yielding lower importance at higher indices [1]. Using `nndsvdar` or `random` spreads variance more evenly and can lift late-index importances, although the set retained after the $(\mu + \sigma)$ threshold is largely uninfluenced.

We set k=20 for NMF to avoid collapse into one dominant factor when k is too small. Overcompleting the basis lets the model express variation, after which Sobol ranking identifies the few concepts that affect the classifier; almost half have near-zero importance. Because NMF is initialization-sensitive, the amount and the exact indices above the threshold can change between runs, but the pattern is consistent: a small set of high-importance concepts and a long tail of negligible ones, consistent with COCKATIEL's outcomes.

Inclusion Criteria Hall 2012

- An empirical study
- Focused on predicting faults in units of a software system
- Faults in code is the main output (dependent variable)

Inclusion Criteria Jeyaraman 2020

- Patients with knee osteoarthritis
- Intervention with MSC therapy
- Comparator: usual care
- Outcomes: VAS for Pain, WOMAC, Lysholm, WORMS, KOOS, and adverse events
- Study design: Randomized controlled trials

Inclusion Criteria Menon 2021

- Explicitly identified as a "systematic review" in the title
- Assessed the effect of a non-acute, non-communicable, environmental exposure on a health outcome
- Included studies in people or mammalian models

Figure 3, Figure 4 and Figure 5 present abstracts the datasets that were included in the study and labeled as relevant. The color indicates concept alignment; values below a set threshold are omitted as the alignment with a concept is deemed too weak to be relevant. The abstracts were selected based on concept occurrence. Not all abstracts contain all concepts, some abstracts have less or no above-threshold sentences. The computations for occlusion, embedding, NFM projection, and delta required for visualizing alignment per abstract take an average of 21 seconds per abstract. Timings were obtained on a 2021 4-core laptop-class CPU. The granularity of the abstracts is sentence based.

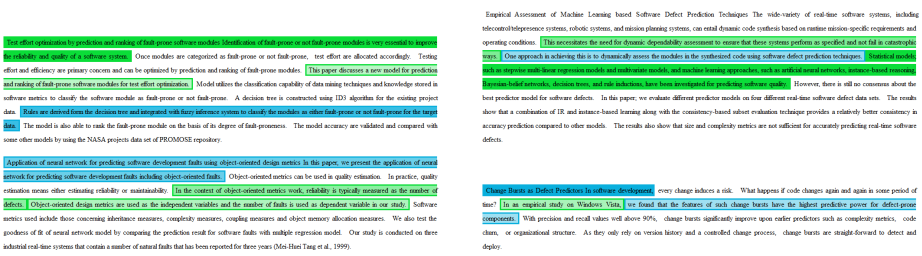


Fig. 3: A selection of highlighted abstracts with all concepts present from the Hall Dataset

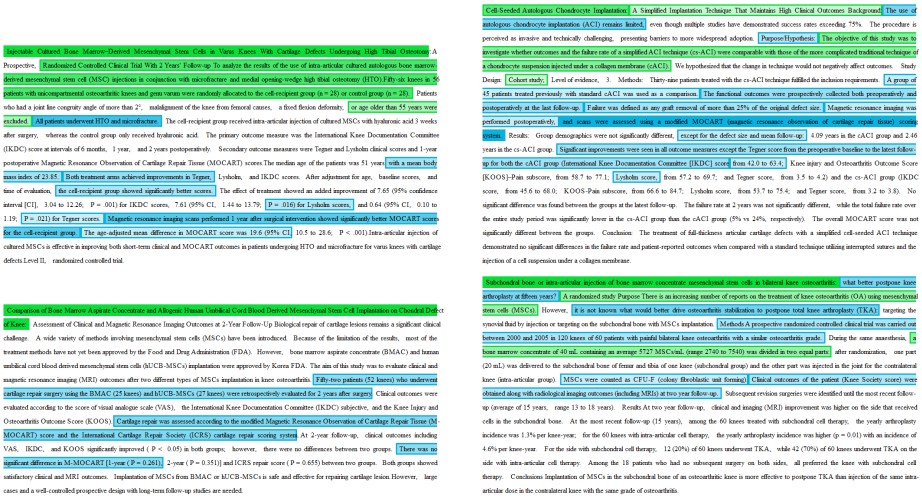


Fig. 4: A selection of highlighted abstracts with all concepts present from the Jeyaraman Dataset

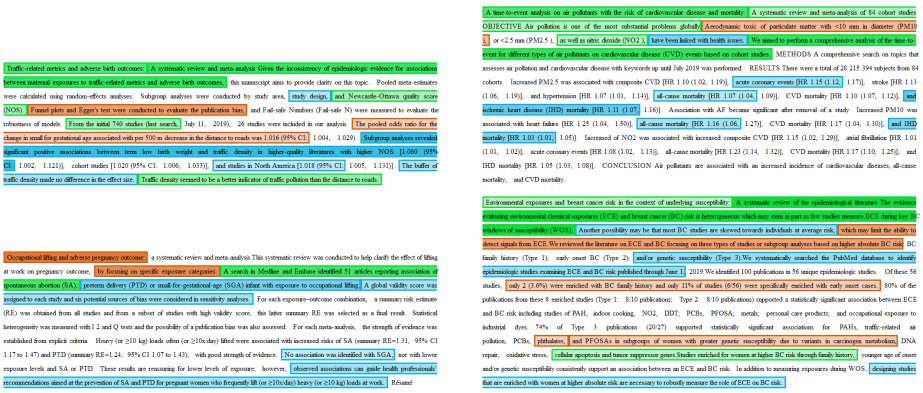


Fig. 5: A selection of highlighted abstracts with all concepts present from the Menon Dataset

4.2 Quantitative Evaluation of Concept Usefulness

As shown in the Inclusion Criteria for Hall 2012, this study applied three inclusion criteria. Using cosine similarity, we calculate the similarity between the embedding of each criterion and each learned concept vector. This allows us to assess whether certain concepts align more strongly with specific criteria. A high similarity score for a given criterion–concept pair suggests that the concept captures semantic information directly related to that criterion, while low scores indicate weak or no alignment.

To calculate similarity, the inclusion criteria are embedded using $h(x)$; the same feature extractor applied during model training. These embeddings are then normalized and compared, via cosine similarity, to the learned concept base W from the NMF decomposition. Along with the inclusion criteria, the similarity scores are computed for unrelated baseline sentences, providing a reference level. Finally, the scores are normalized and visualized in a bar plot, with dashed horizontal lines indicating the mean baseline similarity for each concept. The baseline is the mean similarity to the embeddings of unrelated sentences.

Figure 6 shows the output of this process. We select the concepts identified as important in the CLARIFY process, and compare them to the inclusion criteria for a dataset. For the Hall dataset, 2 important concepts were discovered (concept 5 and concept 7), and 3 inclusion criteria were used for the creation of the dataset (identified as criterion 1, 2 and 3). After calculations, all three inclusion criteria have similarity scores that rise clearly above the baseline for each concept. Inclusion criterion 1 shows a strong alignment with the first important concept, while criterion 2 is more strongly aligned with the second important concept. Criterion 3 exhibits comparable similarity to both concepts. This pattern indicates that the discovered concepts lie within the semantic space of the inclusion criteria. The model's concept structure is organized along dimensions that correspond to the review's decision rules, supporting the hypothesis that

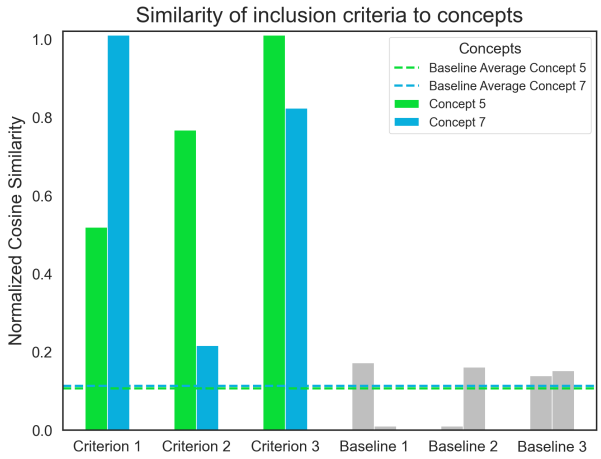


Fig. 6: Normalized cosine similarity for the Hall_2012 dataset between concepts and inclusion criteria, with baseline similarity shown as dashed lines. The plot shows the different similarity between inclusion criteria and concepts.

concept discovery near the classifier (in the final hidden layer) recovers decision-relevant signals.

5 Discussion

5.1 Role of concept positioning

Concept extraction is performed on the embeddings produced by the feature extractor $h(x)$, immediately prior to classification $c(x)$. This location within the pipeline prioritizes the extraction of latent representations that are closely aligned with the classifier's decision function over the interpretability of the concepts, as this is most useful for ASReview.

The hypothesis in this work is that these representations reflect the semantic signals relevant to the inclusion criteria, to the extent that such signals are captured by the model. The goal is not to reconstruct the input or surface linguistically grounded structures, but to identify internal signals that influence classification outcomes. Extracting concepts too early risks overfitting to shallow lexical patterns; extracting them too late risks reducing them to direct encodings of the predicted label.

This is not to say that directly encoding the classification prediction as a concept is useless. By masking each sentence in turn, re-embedding the perturbed text, and comparing the change in predicted relevance to the full abstract, we obtain sentence-level alignment scores visualized in Figure 7. We interpret the final classification output as a single concept and quantify each sentence's effect on the predicted relevance. This provides local accountability for a specific abstract. However, it does not reveal the intermediate semantic factors the model

relies on. It shows the impact sentences have on the classification probability, not which latent dimensions structure the decision.

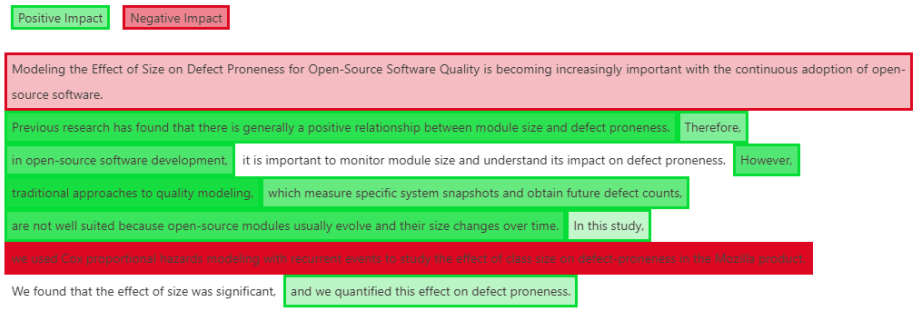


Fig. 7: Final layer accountability highlighted at sentence level for abstract from the Hall 2012 dataset.

Reversing this perspective, CLARIFY does not provide direct accountability for the final classification in the way shown in Figure 7. Instead, it identifies latent concepts that represent intermediate decision signals within the model. These should not be interpreted as explanations for the final prediction. If the goal is to assess the contribution of individual sentences to the final classification, final-layer occlusion as in this example is appropriate. Conversely, occlusion applied to the final hidden layer is suited to revealing which semantic dimensions influence the decision.

In practice, we could use final-layer occlusion for sentence-level accountability, and final hidden-layer occlusion to explain which semantic dimensions drive decisions.

5.2 Cognitive biases and interpretability limitations

The assumption that inclusion criteria are encoded in interpretable units is difficult to support. It is often used to explain the workings of CNN's for images: a complex task like digit recognition is decomposed into simple visual components such as loops, straight lines, and intersections, and recomposed layer by layer into digit identities. Early layers are frequently interpretable; they learn edges, curves, and simple shapes. But move deeper into the network, and the visualizations quickly degrade. Later layers do not resemble meaningful visual parts but instead appear random to the human eye. Neural networks are optimized for task performance, not human interpretability. Often learning performance degrades with an increased explainability [20].

The same goes for lexical challenges. Although earlier layers may offer more interpretable patterns, they carry limited information about the final classification outcome. Highlighting features from these layers may expose the building blocks the model uses and how the lexical input is broken into subproblems, but not how these components are recombined to form a classification. As a result, such representations are not only weakly informative but potentially misleading. They may appear meaningful, yet offer no insight into why a document is

marked relevant. To surface decision-relevant signals, we must operate further towards the end of the pipeline, even if that means forgoing interpretability of steps. Figure 8 visualizes this gradient.

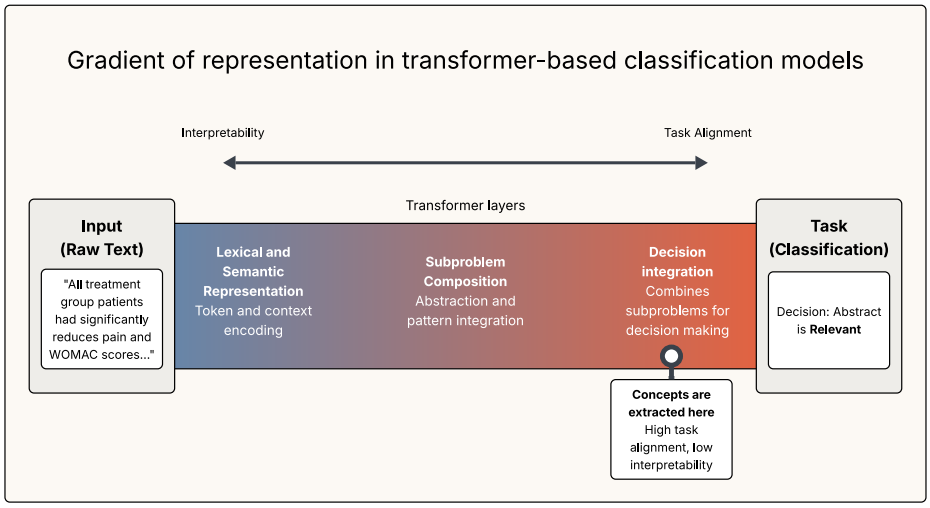


Fig. 8: Abstract visualization of the interpretability-task alignment gradient found in a transformer based classification pipeline.

This pattern holds for ASReview and CLARIFY. While earlier transformer layers may contain interpretable meaningful patterns, the deeper layers, those near the classifier, reflect highly task-specific transformations. There is no reason to assume that latent units in these layers correspond to human-interpretable concepts, even when the training task is structured around inclusion criteria.

This also motivates the decision not to pursue manual labeling of concepts. The pipeline produces alignments between sentences and abstract latent units that influence classification, but does not explain *why*. Any interpretation beyond this point risks reflecting human pattern-seeking rather than grounded evidence. This cognitive bias must be acknowledged when interacting with model explanations.

5.3 Findings in Relation to the Study Objectives

Our findings can be summarized along four main objectives. First, we examined whether latent concepts could be extracted directly from ASReview's transformer's hidden embedding layer without retraining. The results show that this is feasible: concepts can be surfaced and presented as sentence-level highlights accessible to users. These reflect decision-relevant internal signals rather than surface-level lexical features. While the extracted concepts encode information used by the model for its final predictions, further work is needed to determine whether they form coherent, user-understandable units. Second, we explored whether the discovered concepts align with inclusion criteria. Cosine similarity

analyses revealed associations between concepts and criteria, suggesting that the model's latent space captures aspects of these criteria. However, this evidence is correlational and does not imply that the criteria are explicitly encoded as separable concepts. Risks of cognitive bias and interpretability limitations (see Section 5.2) further constrain the strength of this claim. Third, we considered the practical usefulness of the explanations for reviewers. While we demonstrate sentence-level alignment with concept activation, their effectiveness in practice remains inconclusive. A controlled user study would be needed to assess their impact on reviewer performance. Importantly, high concept importance should not be taken as evidence of causal contribution to inclusion decisions, and any implementation must make this distinction explicit to avoid misinterpretation. Finally, we assessed the pipeline's practicality for on-the-fly use in ASReview. Results indicate that the method produces concept-level explanations with low latency on standard laptop hardware. By reusing the fixed NMF basis, Sobol importances, and ASReview's feature extractor, and only recomputing sentence-level occlusions as needed, the approach enables ad-hoc execution without retraining. The design is model-agnostic and suitable for integration into active learning workflows. Looking ahead, concept labeling remains an open challenge. Assigning coherent labels without introducing bias is difficult, but one promising direction is to leverage similarity between concepts and sentences. Because active learning keeps the reviewer in the loop, the system could accept reviewer-proposed labels or criteria and return similarity scores to each concept, optionally with representative sentences. Our results suggest this interaction is viable. The approach is simple to implement, adds minimal computational cost, and enables less biased, user-steerable exploration of the concept space to improve understanding.

6 Conclusion

This study introduced CLARIFY, a post-hoc, concept-based explanation pipeline for active learning screening in systematic reviews. Using embeddings from the final hidden layer, the method factorizes the representation space with NMF to discover latent concepts. It integrates without retraining and reuses model elements to minimize computational cost. The pipeline was evaluated on three SYNERGY datasets, producing ranked concept importance, concept–criterion similarity measures, and sentence-level highlights. All code and results are openly available [22].

CLARIFY demonstrates that post-hoc, concept-based explanations can be integrated into active learning screening without retraining or heavy computation, while preserving model-agnosticism, with about 20 seconds per highlighted abstract in our setup. By surfacing decision-relevant signals, the method moves transformer-based screening models toward greater transparency and interactivity. Practical reviewer benefit requires validation in controlled studies. We see this work as a step toward explainable systematic review tools while employing black-box machine learning models, by (re)enabling accountability in AI-assisted screening.

7 Appendix

.1 Usage of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used *Open Source* Generative AI in order to increase language readability. After use of this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

.2 PseudoCode

```
INPUT
  Dataset of abstracts with titles and labels
  Feature extractor FE
  Classifier CLF, Balancer BAL, Querier QRY
  K = number of NMF concepts
  S = number of Sobol designs

OUTPUT
  Global concept importances
  Per-abstract highlighted sentences with (concept, intensity)

1. EMBED AND NORMALIZE
    FOR each record i IN corpus DO
        SET text_i = CONCAT(title_i, abstract_i)
        OBTAIN x_i = FE.EMBED(text_i)
  ENDFOR
  SET A = STACK(x_i)
  COMPUTE A = NORMALIZE(A, θ_min, θ_max)  // store θ_min, θ_max for reuse

2. ACTIVE LEARNING
  INIT cycle with CLF, BAL, QRY on features A
  WHILE stopping criterion NOT met DO
    CALL QRY to obtain next records
    OBTAIN labels for queried records
    FIT CLF on labeled set with BAL
  ENDWHILE
  SET A_pos = SUBSET of A WHERE label_i = 1

3. CONCEPT FACTORIZATION
  CALL NMF.FIT on A_pos with K components
  SET W = concept basis, U = activations for A_pos

4. GLOBAL CONCEPT IMPORTANCE (SOBOL TOTAL-ORDER)
  FOR each positive embedding a_i IN A_pos DO
    OBTAIN Sobol perturbations guided by concept base W
    ESTIMATE classifier variance using JANSEN_TOTAL_ORDER
    ACCUMULATE importance scores
```

```
  ENDFOR
  COMPUTE S_global = average accumulated scores
  COMPUTE τ = MEAN(S_global) + STD(S_global)
  SET C_top = { k | S_global[k] > τ }

5. SENTENCE-LEVEL OCCLUSION WITH FIXED W
  FOR any positive record i DO
    SET T_full = CONCAT(title_i, abstract_i)
    SET u_full = U[i]

    SPLIT T_full into components = sentences
    FOR each component_j IN components DO
      FORM T_minus_j by removing component_j from T_full
      OBTAIN a_minus_j = FE.EMBED(T_minus_j)
      COMPUTE a_minus_j = NORMALIZE(a_minus_j, θ_min, θ_max)
      OBTAIN u_minus_j = NMF.TRANSFORM_W(a_minus_j, W)
      COMPUTE Δu_j = u_full - u_minus_j
      RESTRICT Δu_j to indices in C_top
      SCALE Δu_j to [0, 1]
      SET concept_id = ARGMAX(Δ u_j[k])
      SET intensity  = MAX(Δ u_j[k])
      ASSIGN component_j WITH (concept_id, intensity)
    ENDFOR
  ENDFOR

6. OUTPUT
  DISPLAY S_global and C_top
  FOR each abstract_i DO
    DISPLAY components with assigned concept labels and intensities
  ENDFOR
```

Bibliography

[1] Boutsidis, C., Gallopoulos, E.: Svd based initialization: A head start for nonnegative matrix factorization. Pattern recognition **41**(4), 1350–1362 (2008)

[2] de Bruin, J., Lombaers, P., Kaandorp, C., Teijema, J.J., van der Kuil, T., Yazan, B., Dong, A., van de Schoot, R.: Asreview lab v2: Open-source text screening with multiple agents and oracles. Available at SSRN 5136987 (????)

[3] De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., Van de Schoot, R.: SYNERGY - Open machine learning dataset on study selection in systematic reviews (2023), `https://doi.org/10.34894/HE6NAQ`, URL `https://doi.org/10.34894/HE6NAQ`

[4] Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D.L., Tummers, L., Teijema, J.J., van de Schoot, R.: Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the average time to discover relevant records. Systematic Reviews **12**(1), 100 (2023)

[5] Gates, A., Johnson, C., Hartling, L.: Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the abstrackr machine learning tool. Systematic reviews **7**, 1–9 (2018)

[6] Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Mueller, K.: Explainable active learning (XAL): an empirical study of how local explanations impact annotator experience. CoRR **abs/2001.09219** (2020), URL `https://arxiv.org/abs/2001.09219`

[7] Hall, T., Beecham, S., Bowes, D., Gray, D., Counsell, S.: A systematic literature review on fault prediction performance in software engineering. IEEE Transactions on Software Engineering **38**(6), 1276–1304 (2012), `https://doi.org/10.1109/TSE.2011.103`

[8] Jain, S., Wallace, B.C.: Attention is not explanation. arXiv preprint arXiv:1902.10186 (2019)

[9] Jeyaraman, M., Muthu, S., Ganie, P.A.: Does the source of mesenchymal stem cell have an effect in the management of osteoarthritis of the knee? meta-analysis of randomized controlled trials. Cartilage **13**(1_suppl), 1532S–1547S (2021)

[10] Jourdan, F., Picard, A., Fel, T., Risser, L., Loubes, J.M., Asher, N.: Cockatiel: Continuous concept ranked attribution with interpretable elements for explaining neural net classifiers on nlp tasks. arXiv preprint arXiv:2305.06754 (2023)

[11] Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. Advances in neural information processing systems **13** (2000)

[12] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. nature **401**(6755), 788–791 (1999)

[13] Lee, S., Shakir, A., Koenig, D., Lipp, J.: Open source strikes bread - new fluffy embeddings model (2024), URL https://www.mixedbread.ai/blog/mxbai-embed-large-v1

[14] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Curran Associates, Inc. (2017), URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[15] Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics **113**, 103655 (2021)

[16] Menon, J., Struijs, F., Whaley, P.: The methodological rigour of systematic reviews in environmental health. Critical Reviews in Toxicology **52**(3), 167–187 (2022)

[17] Rathbone, J., Hoffmann, T., Glasziou, P.: Faster title and abstract screening? evaluating abstrackr, a semi-automated online screening program for systematic reviewers. Systematic reviews **4**, 1–7 (2015)

[18] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. CoRR **abs/1602.04938** (2016), URL http://arxiv.org/abs/1602.04938

[19] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144 (2016)

[20] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

[21] Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)

[22] Teijema, J.: jteijema/asr-xai: v0.1 (Aug 2025), https://doi.org/10.5281/zenodo.16797395, URL https://doi.org/10.5281/zenodo.16797395

[23] Teijema, J.J., de Bruin, J., Bagheri, A., van de Schoot, R.: Large-scale simulation study of active learning models for systematic reviews. International Journal of Data Science and Analytics pp. 1–22 (2025)

[24] Teijema, J.J., Ribeiro, G., Seuren, S., Anadria, D., Bagheri, A., van de Schoot, R.: Simulation-based active learning for systematic reviews: A scoping review of literature (2023)

[25] Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al.: An open source machine learning framework for efficient and transparent systematic reviews. Nature machine intelligence **3**(2), 125–133 (2021)

[26] Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion **76**, 89–106 (2021)