January 2024

White Paper

Bridge
Evidence that matters

Artificial Intelligence in Systematic Literature Reviews

Part 1 | AI-aided Title/Abstract Screening

Bridge has a well-established interest in exploring the role of artificial intelligence (AI) in delivering systematic literature reviews (SLRs). With the emergence of novel and powerful large language models (LLMs), we restarted our research comparing the performance of these models against our extensive in-house datasets, which have been robustly human-QC'ed for over 100 SLRs of different types. In this first paper of the series, we share our methodology and results on AI performance in title/abstract screening. We will release further papers as our research continues.

Introduction

Literature reviews are time- and resource-intensive. A typical large systematic literature review (SLR) requires many months to complete; such SLRs were estimated to cost USD 141,194 each in 2019.[1] This has led to an increasing focus on automating literature review tasks using artificial intelligence (AI) – specifically, machine learning (ML), which uses algorithms to identify patterns in data (ML may be supervised or unsupervised), and natural language processing (NLP), which enables computers to understand, interpret, generate, and respond to human language in a meaningful manner.[2,3]

Historically, tests on using AI in various SLR tasks have yielded mixed results. Furthermore, there is no consensus on a standard approach to validate the performance of AI models on specific tasks. Over the last few years, bidirectional encoder representations from transformers (BERT) models pre-trained on relevant data have been assessed for their ability to classify data appropriately while conducting SLRs.[3,4,5]

Some positive findings have been demonstrated in the area of title/abstract (ti/ab) screening.[6,7,8] However, the following issues have been reported:

1. **Variable accuracy:** Ti/ab screening accuracy levels have not been replicated across the literature [4,7,8,9,10]
2. **Limited focus:** The available data has mostly focused on interventional studies with safety/efficacy/effectiveness outcomes. Limited tests have been reported for observational studies (which have generally yielded lower accuracy) [8,11,12]
3. **Methodological challenges:** Most automated approaches have implemented a probabilistic approach, with citations ranked by probability of being included. This differs from traditional human screening, therefore, it is not always clear when it is safe to dispense with manual screening altogether [11]
4. **Non-intuitive approach:** Most automated approaches appear to provide a single final assessment of inclusion/exclusion of a citation, while traditional human screening also provides a reason for exclusion

Given this, it is not entirely surprising that a recent review on automated literature reviews concluded that "no single platform appeared to be sufficiently accurate and reliable to date" [13]

In the past 12 months, the situation has changed significantly. In addition to traditional BERT models, we also have access to an ever-increasing suite of large language models (LLMs) such Generative Pre-trained Transformer (GPT), Large Language Model Meta AI (LLaMa) and GEMINI, which are trained to predict language and writing based on large datasets of written language. [14]

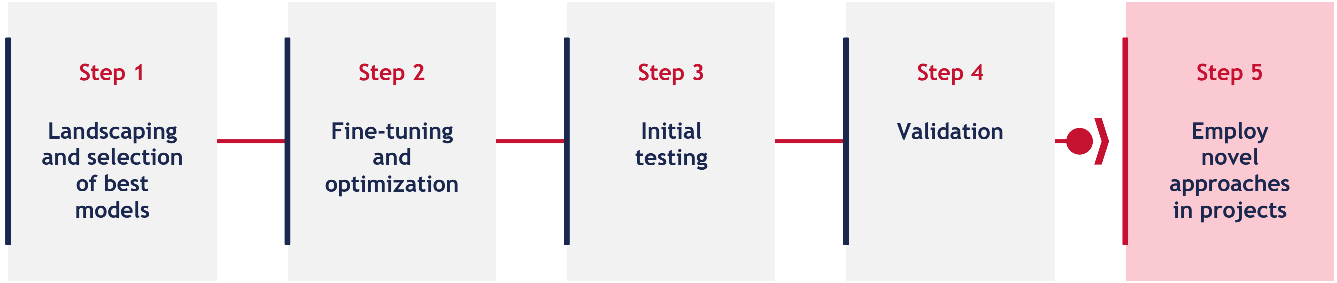
Evaluating newer AI models

At Bridge, we have examined the accuracy of available AI models over the years against our ‘gold-standard’ reference datasets*. Our position in 2021, after an extensive research program using the best commercially available AI tools of that time, was that *cautious optimism* was warranted, but that full-scale adoption of AI tools into SLRs was not yet feasible. [15]

After novel LLMs became available, we instituted a comprehensive program to evaluate their performance across all key stages in a literature study; namely, ti/ab screening, full-text screening, simple data extraction and simple text summaries. Evaluation at each stage consists of a structured five-step process (**Figure 1**).

In this first paper in our ***AI in SLRs*** series, we present our findings on AI-aided ti/ab screening, which is now at step 5.

Figure 1: Structured approach to AI testing



Step 1

Landscaping and selection of best models

Step 2

Fine-tuning and optimization

Step 3

Initial testing

Step 4

Validation

Step 5

Employ novel approaches in projects

Step 1: Landscaping of potential models

After an initial scoping exercise, we identified the five most promising AI models based on pre-specified criteria (**Table 1**).

Table 1: Key models identified after landscaping process

Name	Type	Brief description
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	Pre-trained BERT model	Pretrained language model specifically designed for biomedical natural language processing tasks. It was trained from scratch using abstracts and full-text articles from PubMed and PubMed Central
Medicalai/ ClinicalBERT	Pre-trained BERT model	Pretrained language model, trained on a large multicenter dataset with a large corpus of 1.2 billion words on diverse diseases
BiomedVLP-CXR-BERT-general	Pre-trained BERT model	Trained from a randomly initialized BERT model via MLM on PubMed abstracts and clinical notes from the publicly-available MIMIC-III and MIMIC-CXR. This general model is expected be applicable for research in clinical domains other than chest radiology through domain-specific fine-tuning
BERT-base-uncased	Pre-trained BERT model	Model pretrained on the English language using an MLM objective
GPT-4	Multimodal generative AI model	Multimodal LLM created by OpenAI, and the fourth in its series of GPT foundation models; pretrained using both public data and "data licensed from third-party providers"

BERT=Bidirectional Encoder Representations from Transformers; GPT-4=Generative Pre-trained transformer; LLM=large language model; MLM=masked language modelling.

The BERT models were accessed via the Hugging Face platform, while GPT-4 was accessed using the OpenAI application programming interface (API).

* Bridge ‘gold standard’ reference datasets are fully QC’ed human-screened ti/ab screening datasets available in-house across >100 reviews

Step 2: Fine-tuning and optimization

Fine-tuning the BERT models

Title/abstract screening itself has several steps, corresponding to the population, indication, comparator, outcomes, and study design (PICOS) parameters. Different models appear to have different advantages; therefore, the BERT models were further trained using Bridge reference datasets for each screening step. In this way, we developed fine-tuned versions of *each* (pre-trained) BERT model for *each* screening parameter.

Optimizing GPT4

GPT4 did not require fine-tuning (note that fine-tuning functionality in GPT4 was not available when we conducted our ti/ab screening assessment in Q2-Q3 2023),but required the use of the most appropriate prompts. We automated this task with Python libraries including Langchain (for prompts structuring and engineering) and OpenAI for API access. GPT prompts were created separately for each screening parameter.

Step 3: Initial testing

Initial testing of individual ti/ab screening parameters

The optimal fine-tuned BERT models and GPT4 were then evaluated against human-screened ti/ab screening datasets for five SLR projects. These SLRs were selected to cover a range of research question complexities, therapy areas, study designs of interest, and outcomes (**Table 2**).

Each screening parameter (as per the PICOS) was evaluated separately, i.e., the classification of ti/abs by the AI models was compared with the human classification. As an example of the output from initial testing for individual screening parameters, **Table 3** shows the accuracy of a range of fine-tuned models for the 'review' parameter for a single project.

Table 2: Projects for initial testing

Indication	Outcomes	Study designs
Anemia associated with chronic kidney disease	Safety Efficacy Effectiveness Tolerability	Interventional studies Prospective cohort studies
Metastatic castration-resistant prostate cancer	Safety Efficacy Effectiveness Tolerability	Interventional studies
Atopic dermatitis	Burden of illness (clinical, humanistic, economic burden)	Observational studies Systematic literature reviews Narrative reviews
Soft tissue sarcoma	Safety Efficacy Effectiveness Quality of life Tolerability	Interventional studies Prospective cohort studies Retrospective cohort studies
Borderline personality disorder	Burden of illness (clinical, humanistic, economic)	Observational studies Systematic literature reviews Narrative reviews

Table 3: Accuracy of different AI models in classifying articles as reviews (vs primary studies) from ti/abs in the anemia-CKD project

	Accuracy	Sensitivity	Specificity	NPV	PPV
Model A	95%	97%	92%	94%	96%
Model B	95%	96%	94%	96%	95%
Model C	93%	97%	86%	91%	96%
Model D	95%	97%	92%	94%	96%
Model E	91%	90%	94%	95%	87%
Model F	95%	98%	91%	97%	94%

CKD=chronic kidney disease; NPV=Negative predictive value, PPV=Positive predictive value; ti/ab=title/abstract.

Initial testing of combined ti/ab screening parameters

The best-performing models (including BERT and GPT4) with the highest accuracy were selected, with the emphasis on sensitivity to ensure that no key articles (or very few) were excluded. These models were then used in sequence for each step in the screening flowchart (akin to human screening) to determine the final eligibility decision for ti/abs, and the screening accuracy was calculated (**Table 4**).

In addition to ti/ab sensitivity, we also calculated the 'practical' sensitivity, which – in our view - is the more relevant parameter, i.e., of the publications that were eventually included for final reporting in the project*, how many were correctly identified for inclusion by AI during ti/ab screening.* Thus, taking the example of the atopic dermatitis SLR, of the total number of citations that were eventually used for data reporting, 98% were correctly included by automated ti/ab screening, and only 2% were incorrectly excluded (false negatives). Overall, 'practical' sensitivity ranged from 89% to 98% across projects, and specificity from 68% to 89% (**Table 4**).

Table 4: Results from initial testing for the five projects

Project topic	Total hits	Accuracy	Ti/ab Sensitivity	'Practical' Sensitivity	Specificity	PPV	NPV	TP	FP	TN	FN
Anemia-CKD	3740	89%	89%	96%	89%	54%	98%	428	370	2891	51
mCRPC	3885	85%	86%	90%	85%	62%	95%	748	456	2555	126
Atopic dermatitis	7356	69%	93%	98%	68%	9%	100%	222	2247	4869	18
Soft tissue Sarcoma	2149	82%	82%	95%	82%	38%	97%	212	343	1549	45
Borderline personality disorder	3464	74%	82%	89%	73%	11%	99%	113	884	2443	24

CKD=chronic kidney disease; FN=false negative; FP=false positive; mCRPC=metastatic castration-resistant prostate cancer; NPV=negative predictive value; PPV=positive predictive value; Ti/ab=title abstract; TN=true negative; TP=true positive.

Step 4: Validation

Having tested the AI models across 5 projects, we then validated them across additional datasets: two internal datasets where the Bridge AI team had no prior knowledge of any results, and an external Cochrane SLR where the AI team were 'blinded' to results. We also assessed a workload/time reduction metric to evaluate the efficiency of the AI-aided screening compared to human screening. The 'practical sensitivity' and specificity remained high across the three projects, providing evidence that the accuracy of the fine-tuned BERT models and GPT prompts was transferable across projects. The workload reduction was >90% in two of three projects (**Table 5**).

Table 5: Results from validation testing

	Total hits	Accuracy	Ti/ab Sensitivity	'Practical' Sensitivity	Specificity	Workload reduction	PPV	NPV	TP	FP	TN	FN
Advanced met. NSCLC (safety/ efficacy)	2871	97%	84%	98%	98%	91%	56%	99%	87	68	2699	17
Depression in cancer [Cochrane] (safety/ efficacy)	5550	98%	63%	83%	98%	98%	28%	100%	33	84	5414	19
Post-traumatic stress disorder (burden of illness)	7065	65%	65%	96%	65%	64%	4%	99%	102	2433	4475	55

FN=false negative; FP=false positive; met.=metastatic; NPV=negative predictive value; NSCLC=non-small-cell lung cancer; PPV=positive predictive value; Ti/ab=title abstract; TN=true negative; TP=true positive.

Reflection on findings and next steps

Overall, the accuracy of the models in ti/ab screening was very high. The accuracy for conceptually more straightforward, though still quite complex, efficacy/effectiveness/safety-focused SLRs was at or above high-performing human standards. [16] For the more conceptually challenging burden-of-illness SLRs, the ti/ab screening performance was impressive, and we believe will only continue to improve. These models were tuned for sensitivity and achieved a high 'practical' sensitivity of 83%-98% across the SLRs. At the same time, specificity is also critical for practical workload reduction, and this is an area where we saw the biggest advance over our previous research. [15] The workload reduction was 64%-98% across the SLRs, corresponding to many days/weeks of time saved in ti/ab screening per SLR.

Bridge is already in the process of sharing the detailed findings of its research to date across its client base, and as of January 2024, two AI-enabled literature reviews have been completed for clients, and a further seven are underway in which AI-enabled screening is central to the methodology.

Alongside the adoption of AI tools into client work, our AI workstream research program continues. We plan to share our findings [positive and negative] on the performance of these models on full text screening, data extraction and simple summaries as the data emerges in the coming months. Our very next AI focused white paper will be on the fast-moving field of 'prompt engineering.'

In addition, we have expanded our AI workstream beyond SLRs, to examine use cases across health economics and outcomes research (HEOR) and evidence generation. Once again, we will share our methodology and findings as soon as is practical.

As the field continues to evolve, the potential of AI to transform literature studies, and HEOR more generally, remains significant. With the advent of increasingly high-quality data comparing the efficacy of AI systems to conventional human-led methods, key institutions that effectively set methodological benchmarks for SLRs — such as peer-reviewed journals, Cochrane, and Health Technology Assessment bodies — will need to establish specific validation criteria for AI systems. This will pave the way for AI-enabled SLRs to become broadly recognized and adopted.

Glossary

AI: Artificial Intelligence is a branch of computer science that aims to create systems capable of performing tasks that normally require human intelligence. These tasks include visual perception, speech recognition, decision-making, and translation between languages.

BERT (Bidirectional Encoder Representations from Transformers): BERT is a model based on the transformers architecture for natural language processing pre-training developed by Google. It is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context. The BERT model is pre-trained on a large corpus of text and then fine-tuned for specific tasks like question answering or sentiment analysis. Unlike previous models, BERT takes into account the full context of a word by looking at the words that come before and after it—hence it is bidirectional.

Gemini: Gemini is a family of multimodal large language models developed by Google DeepMind.

GPT: Generative Pre-trained Transformer refers to a series of language processing AI models developed by OpenAI. These models utilize a transformer architecture for deep learning and are pre-trained on a vast corpus of text data. The "generative" aspect refers to the model's ability to generate coherent and contextually relevant text based on input prompts.

In supervised learning, the algorithm is trained on a pre-defined set of training examples, which then facilitate its ability to reach an accurate conclusion when given new data.

In unsupervised Learning, the algorithm is and AI. It is best known for its open-source transformers library and platform (https://huggingface.co/), which provides a collection of pre-trained models and tools for a variety of NLP tasks.

LLaMa: Large Language Model - Meta AI, is a series of large language models developed by Meta AI (previously Facebook AI). These models are designed to process and understand human language, offering capabilities similar to other large language models in tasks like translation, text generation, and information extraction.

LLM: Large Language Models are a type of artificial intelligence models that process, understand, generate, and sometimes translate human language. These models are "large" both in terms of the size of their neural network architecture (having a large number of parameters) and the vast amount of data they are trained on. LLMs are often based on transformer architectures and are trained on diverse datasets from the internet or other large text corpora.

ML: Machine Learning, a subset of AI, involves the development of algorithms that can learn and make predictions or decisions based on data. This learning process is automated and improved upon over time based on experience.

NLP: Natural Language Processing is a field at the intersection of computer science, artificial intelligence, and linguistics. It involves the development of algorithms and systems that enable computers to understand, interpret, and generate human language in a valuable way. Key tasks in NLP include text

NPV = True Negatives / (True Negatives + False Negatives)

PPV: Positive Predictive Value is the proportion of positive test results that are true positives.

PPV = True Positives / (True Positives + False Positives)

'Practical' sensitivity: In the context of AI-aided ti/ab screening, we have defined practical sensitivity to refer to those true positives that were eventually included for final reporting in the review, i.e., the proportion of final actual positives that were correctly identified as such during AI ti/ab screening.

Sensitivity: Sensitivity measures the proportion of actual positives that are correctly identified as such.

Sensitivity = True Positives/ (True Positives + False Negatives)

SLR: A systematic literature review is a methodical and comprehensive approach to identifying, evaluating, and synthesizing all relevant research on a specific topic or research question.

Specificity: Specificity measures the proportion of actual negatives that are correctly identified as such.

Specificity = True Negatives / (True Negatives + False Positives)

TLR: A targeted literature review is a more focused approach than an SLR and is typically used to address specific, often narrower, research questions. A TLR usually does not involve a structured search of the literature, but instead relies on targeted, keyword-based non-exhaustive searches of databases.

given data without predefined labels and is allowed to find structure in its input on its own.

Hugging Face: Hugging Face is a technology company specializing in NLP translation, sentiment analysis, speech recognition, and language generation.

NPV: Negative Predictive Value is the proportion of negative test results that are true negatives.

Authors: Saifuddin Kharawala, Sam Isaacs, Divyanshu Jindal, Paul Gandhi

Acknowledgements: Anadi Mahajan, for support with organizing the training datasets; Digant Gupta, for statistical inputs

References

1. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019 Aug 25;16:100443. doi: 10.1016/j.conctc.2019.100443. Erratum in: Contemp Clin Trials Commun. 2019 Sep 12;16:100450. PMID: 31497675; PMCID: PMC6722281
2. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018 Oct;2(10):719-731. doi: 10.1038/s41551-018-0305-z. Epub 2018 Oct 10. PMID: 31015651.
3. Perlman-Arrow S, Loo N, Bobrovitz N, Yan T, Arora RK. A real-world evaluation of the implementation of NLP technology in abstract screening of a systematic review. Res Synth Methods. 2023 Jul;14(4):608-621. doi: 10.1002/jrsm.1636. Epub 2023 May 25. PMID: 37230483.
4. Ng SH, Teow KL, Ang GY, Tan WS, Hum A. Semi-automating abstract screening with a natural language model pretrained on biomedical literature. Syst Rev. 2023 Sep 23;12(1):172. doi: 10.1186/s13643-023-02353-8. PMID: 37740227; PMCID: PMC10517490.
5. Qin X, Liu J, Wang Y, Liu Y, Deng K, Ma Y, Zou K, Li L, Sun X. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. J Clin Epidemiol. 2021 May;133:121-129. doi: 10.1016/j.jclinepi.2021.01.010. Epub 2021 Jan 21. PMID: 33485929.
6. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. 2022 Apr;144:22-42. doi: 10.1016/j.jclinepi.2021.12.005. Epub 2021 Dec 8. PMID: 34896236.
7. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening - impact on reviewer-relevant outcomes. BMC Med Res Methodol. 2020 Oct 15;20(1):256. doi: 10.1186/s12874-020-01129-1. PMID: 33059590; PMCID: PMC7559198.
8. Gates A, Gates M, DaRosa D, Elliott SA, Pillay J, Rahman S, Vandermeer B, Hartling L. Decoding semi-automated title-abstract screening: findings from a convenience sample of reviews. Syst Rev. 2020 Nov 27;9(1):272. doi: 10.1186/s13643-020-01528-x. PMID: 33243276; PMCID: PMC7694314.
9. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev. 2020 Apr 2;9(1):73. doi: 10.1186/s13643-020-01324-7. PMID: 32241297; PMCID: PMC7118839.
10. Natukunda A, Muchene LK. Unsupervised title and abstract screening for systematic review: a retrospective case-study using topic modelling methodology. Syst Rev. 2023 Jan 3;12(1):1. doi: 10.1186/s13643-022-02163-4. PMID: 36597132; PMCID: PMC9811792.
11. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul 11;8(1):163. doi: 10.1186/s13643-019-1074-9. PMID: 31296265; PMCID: PMC6621996.
12. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ. 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. 2021 May;133:140-151. doi: 10.1016/j.jclinepi.2020.11.003. Epub 2020 Nov 7. PMID: 33171275; PMCID: PMC8168828.
13. Blaizot A, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins M, Lai NM, Chaiyakunapruk N. Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods. 2022 May;13(3):353-362. doi: 10.1002/jrsm.1553. Epub 2022 Feb 28. PMID: 35174972.
14. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev. 2023 Apr 29;12(1):72. doi: 10.1186/s13643-023-02243-z. PMID: 37120563; PMCID: PMC10148473
15. Kharawala S, Mahajan A, Gandhi P. Artificial intelligence in systematic literature reviews: a case for cautious optimism. J Clin Epidemiol. 2021 Oct;138:243-244. doi: 10.1016/j.jclinepi.2021.03.012. Epub 2021 Mar 19. PMID: 33753227.
16. Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. Res Synth Methods. 2019 Dec;10(4):539-545. doi: 10.1002/jrsm.1369. Epub 2019 Jul 18. PMID: 31272125; PMCID: PMC6989049.