

Introduction to automated screening techniques

Oct 7, 9:00 - 10:30

Devi VEYTIA
Postdoctoral researcher
École Normale Supérieure

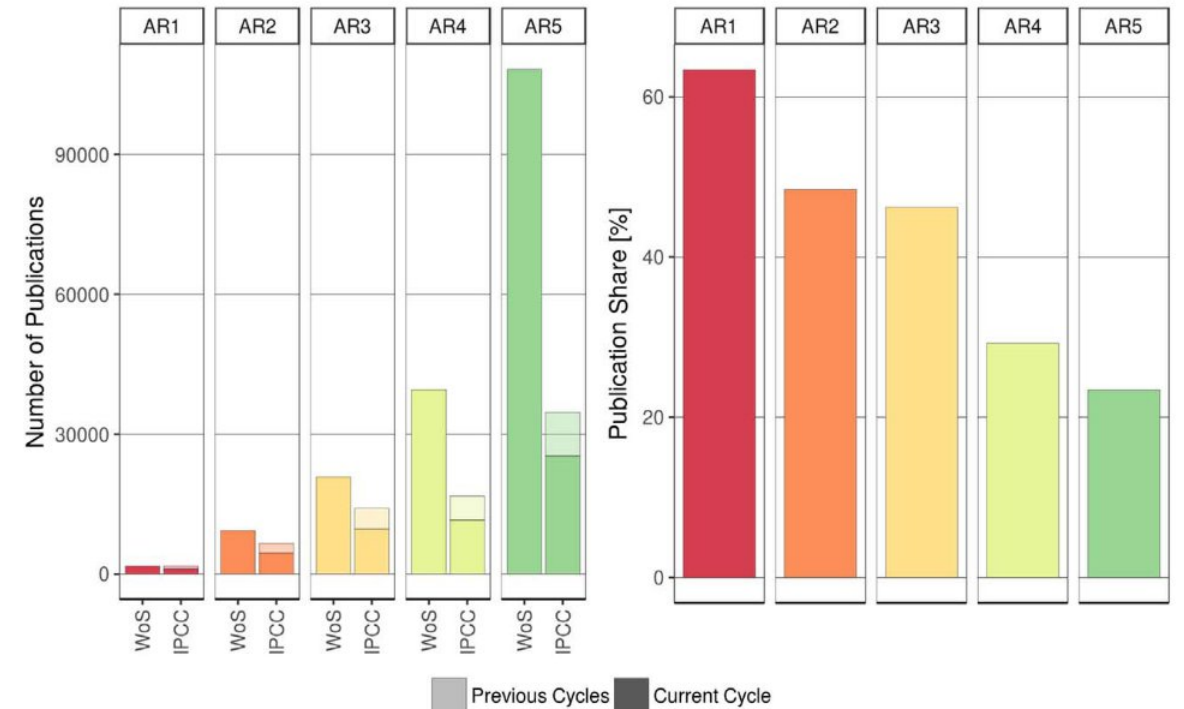
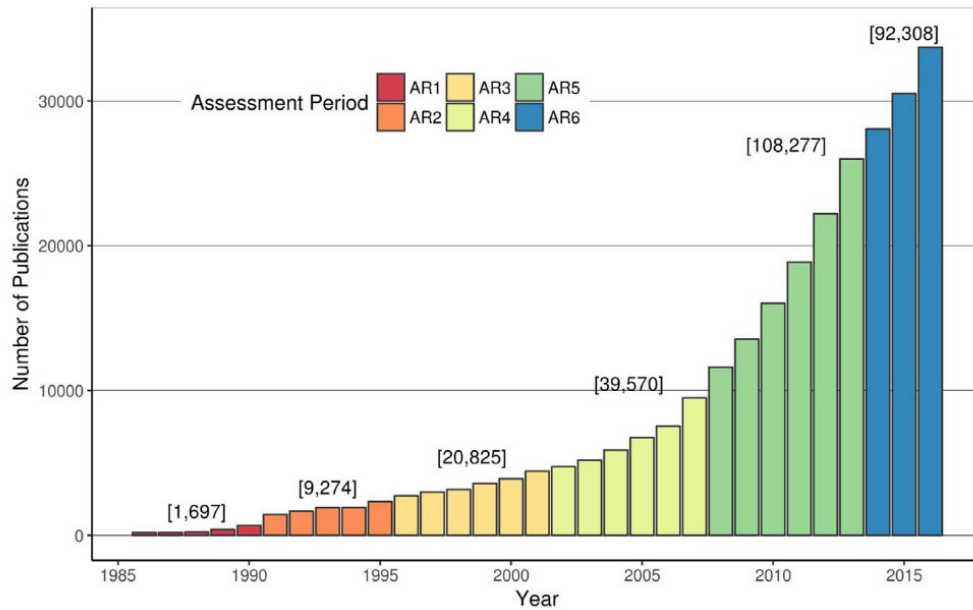
Outline

1. What is the need?
2. The solution: Machine learning?
3. Previous applications of machine learning in evidence syntheses
4. Recent advancements: Transformer model
5. TD: Encoder model for screening:
6. Other applications/examples

What is the need ?

Publications is exponentially increasing

What are the implications for evidence syntheses?



The solution: Machine learning?

taskade.com



AI Research Question Prompt Generator

Unleash your academic potential with the click of a button! Our Research Question Prompt Generator is your secret weapon to effortlessly crafting thought-provoking questions that'll make your studies and papers stand out. Try it now and let inspiration strike!

Elicit.com Analyze research papers at superhuman speed

Automate time-consuming research tasks like summarizing papers, extracting data, and synthesizing your findings.



AI Agent Launched

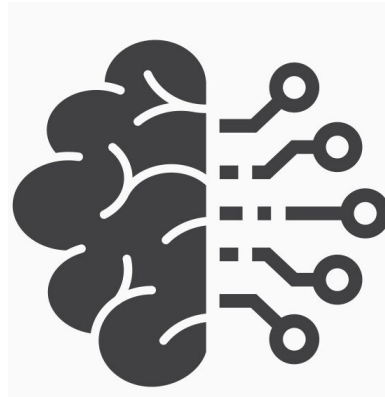
AI Assistant to Automate Everyday Research Tasks

Write a research task or choose one below and SciSpace Agent will use the best AI Models, Tools and Data to complete it for you.

Give me any task to work on...



Deep Search



AI Search Engine for Research

Find & understand the best science, faster.

<https://consensus.app/>

Ask the research...



Does exercise improve cognition? Q

Can cash transfers reduce poverty? Q

Are statins effective in the elderly? Q

Can mindfulness help with sleep? Q

Try an example search



ML in evidence syntheses is new, untested and unreliable

Evidence syntheses have been using unsupervised ML (e.g. topic models and clustering algorithms) for *decades*.

Topic models:

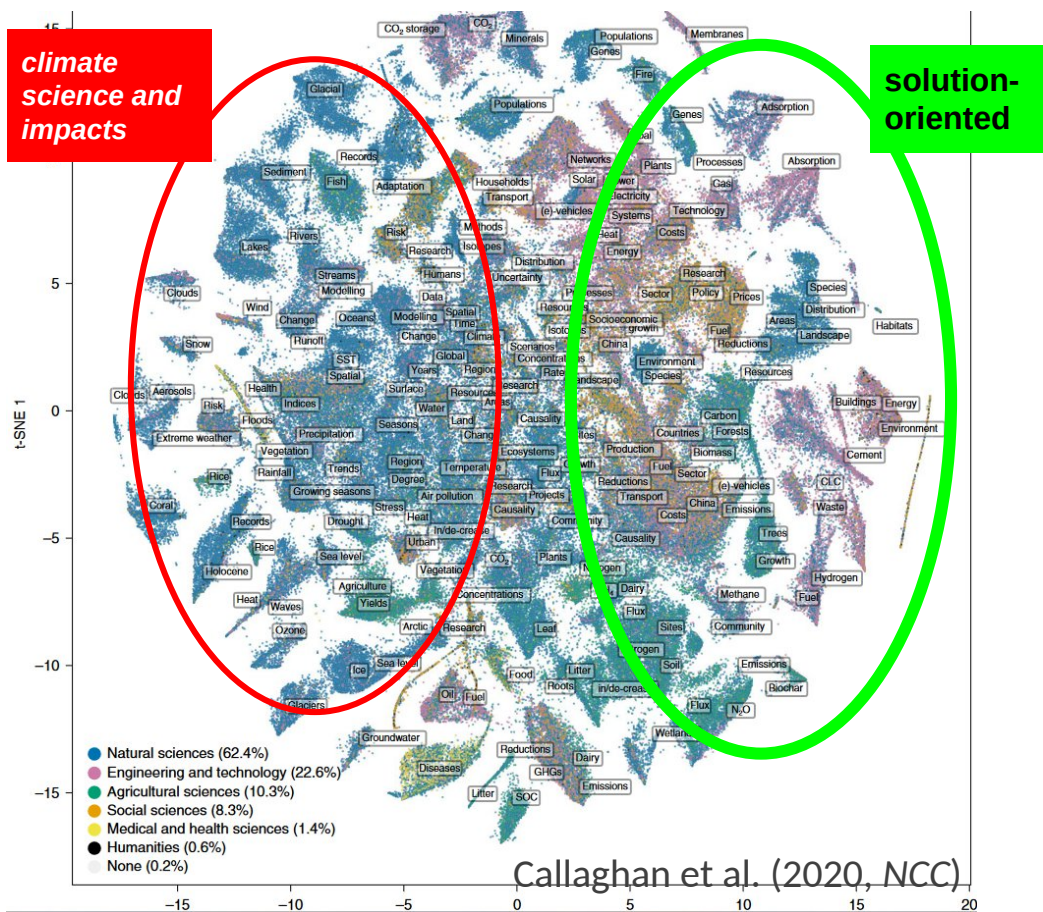
- Describe (and often cluster) documents by quantifying words that occur frequently together
- Pros: No need for hand-labelled data or advanced coding knowledge → fast to implement
- Cons: No control over the topics → generated automatically by identifying patterns that explain variability in the data, but these groupings may not be interesting to your research question/conceptual framework of the review

Topic model use cases

Scoping topics of a corpus to refine your research question (e.g. ~ systematic map)

As a screening method to only screen relevant clusters, or screen clusters that are likely irrelevant and exclude the irrelevant references (e.g.

<https://www.nactem.ac.uk/robotanalyst/>)



So why now? Why is ML growing in popularity?

1. Need → the volume of evidence is becoming an increasingly limiting factor
2. Recent advancements in language models that allow for a more sophisticated interpretation and analysis of text data
 - a. Access to large amounts of text data for training
 - b. Mathematical advancements that reduce computation requirements (backpropagation)
 - c. Advancements in the model architecture (ie [Transformer models](#)), and
 - d. Video gamers

So what is a language model?

- Language models are probabilistic models. They are designed to predict the most likely next word in a sequence of text based on the context provided by the preceding words
- Large Language Models are language models trained on large amounts of data using a transformer architecture

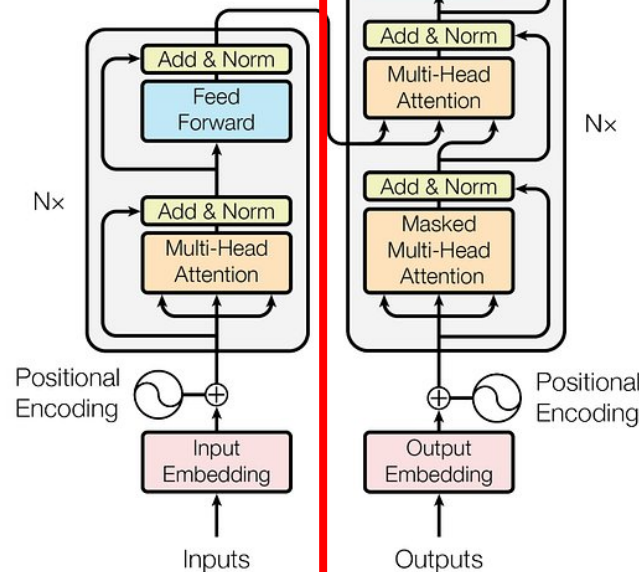


What is a transformer model?

Text classification, where the model must classify a piece of text into one of several predefined categories.

The encoder takes in a sequence of tokens and produces a fixed-size vector representation of the entire sequence, which can then be used for classification.

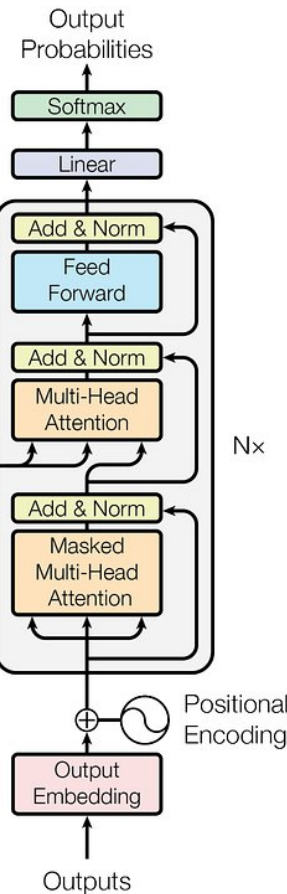
BERT Encoder



Language generation, where the model must generate a sequence of words based on an input prompt or context.

The decoder takes in a fixed-size vector representation of the context and uses it to generate a sequence of words one at a time

GPT Decoder



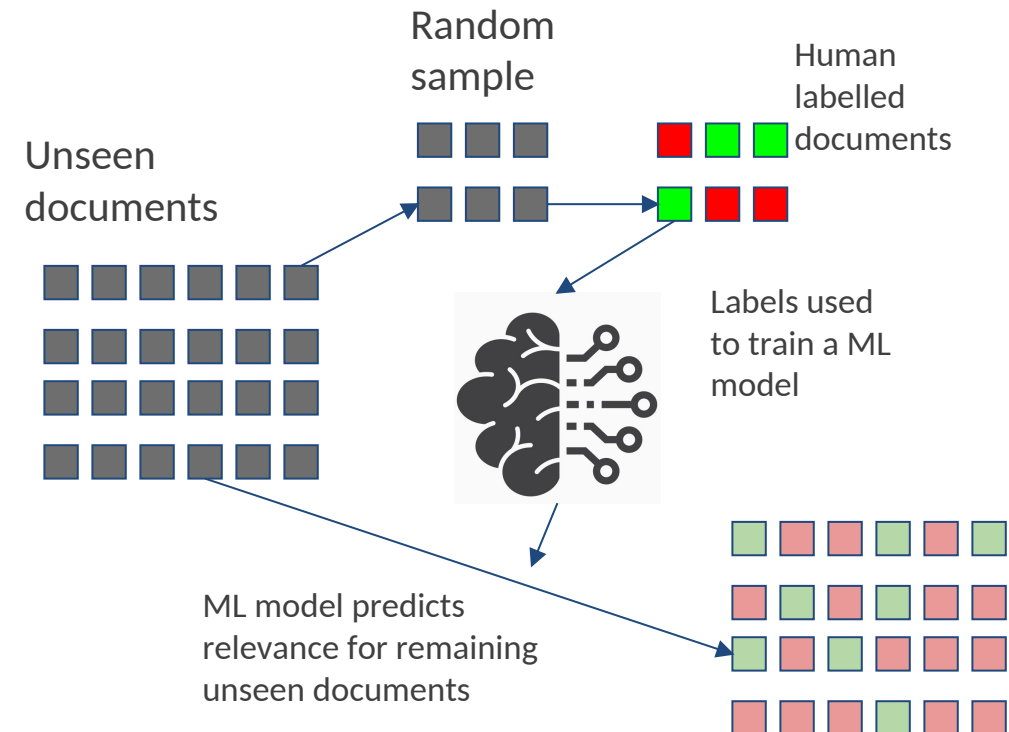
TD: Encoder block for classification

Literature review example:

Screening → training the model to classify unseen text as relevant or not relevant

How do we train and evaluate a classification model? → Cross validation

- Need to provide it with a human-labelled dataset of text and screening decisions to learn from
- We divide this data into a training set and test set.
- Use the test set to train the model
- Use the model to make predictions for the test set, and evaluate how well it does against the 'true' labels



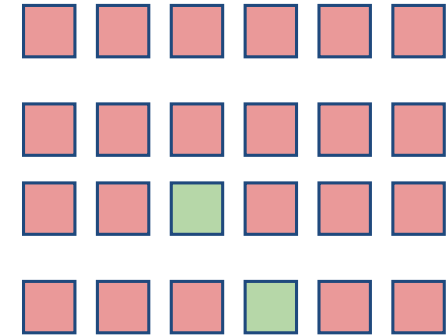
TD: Encoder block for classification

How to we evaluate the model?

Accuracy: Out of all the predictions we made, how many were true?

Example: We have a data set with 100 articles, and 2 are relevant. Our model could predict every document as irrelevant and it would have 98% accuracy, but it would be a terrible model for our purpose (finding the relevant articles)

So we need a way to maximize our *recall* of relevant papers, while avoiding including too many irrelevant papers (*precision*)



TD: Encoder block for classification

How to we evaluate the model?






Recall: how good the model is at finding all the relevant articles (TP/TP+FN)?

Precision: Out of all the articles included, how many are truly relevant (TP/TP+FP)?

Generally aim for $F1 > 0.7$ for simple (binary) classification problems. For more complex multi-label, F1 scores can be as low as 0.45. But, this implies a larger amount of error – deciding whether a model is fit for purpose depends on the purpose. For a map this is probably sufficient to describe general patterns in the data, but probably not for a rigorous meta-analysis.

11

F1 Score: combines recall and precision. Because there is a trade-off between precision and recall, F1 measures how the model makes that trade-off.

		Predicted	
		Animal	Not animal
Actual	Animal	 	
	Not animal		 

True Positives	2
True Negatives	3
False Positives	0
False Negatives	1

Accuracy	83%	$\frac{3+2}{3+2+0+1}$
Precision	75%	$\frac{3}{3+1}$
Recall	100%	$\frac{3}{3+0}$
F1 score	86%	$2 \cdot \frac{0.75 \cdot 1}{0.75 + 1}$

Model 4

TD: Encoder block for classification

Step 2: Represent text numerically

2.1: Tokenisation

Take raw text, convert it into tokens (numbers). [See example of OpenAI's tokenizer](#)

The CESAB literature review course is awesome

Clear
Show example

Tokens	Characters
8	45

The CESAB literature review course is awesome

[976, 105985, 2912, 23216, 3358, 4165, 382, 15339]

Sequence of words



Integers

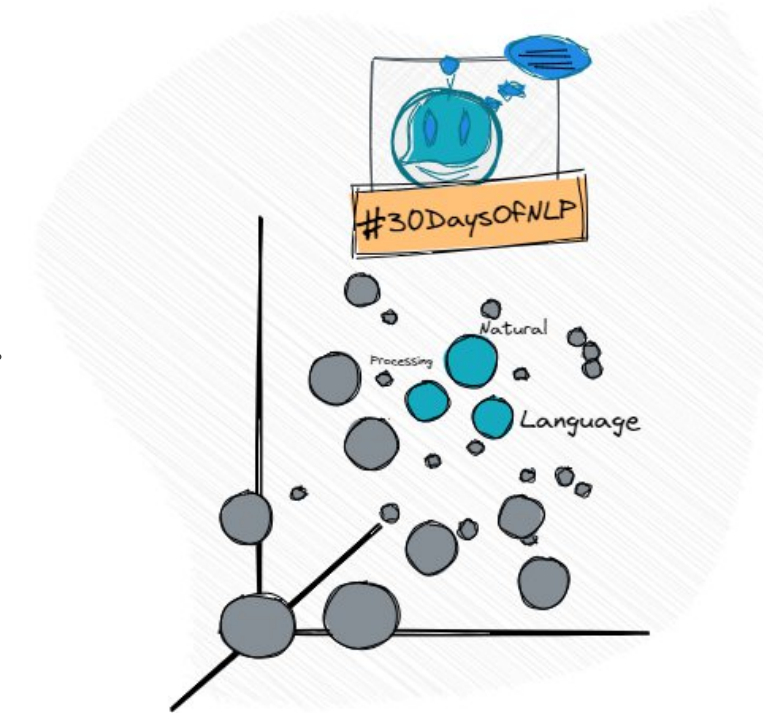
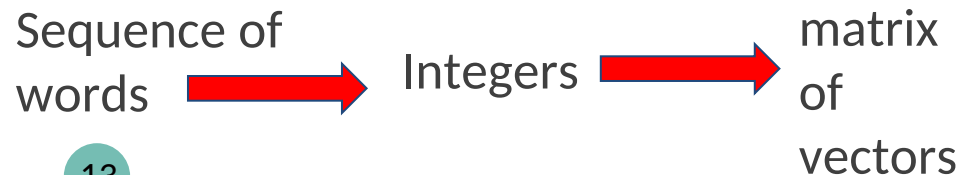
TD: Encoder block for classification

Step 2: Represent text numerically

2.2: Input embedding + positional encoding

For each token, you give the model a vector of numbers to train, as well as information about its position in relation to the other words in the sentence.

- Each token (integer) is accompanied by a vector
- A sequence of vectors → an embedding matrix where the i -th row of the embedding matrix = the embedding of the i -th token (this is a trainable parameter).
- At the end of each vector, adds information on position of the token in the sentence



Why: to train matrix that represents high-dimensional vector space where semantic relationships between tokens become mathematical relationships

TD: Encoder block for classification

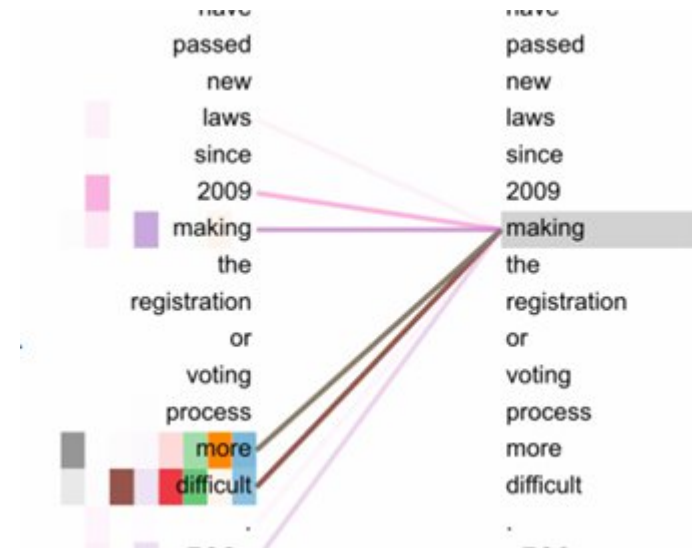
Step 3: Attention

The model uses these inputs to calculate the relative meaning of the word inside the sentence, because depending on the context, the same word can have different meanings

*'more' and 'difficult' contribute to the word 'making'.
So rather than its generic meaning, it's related to
making something more difficult.*

Step 4: Classification

Encoder produces an output matrix numerically describing the text, which is then matched with the human label (e.g. relevant or not) to train a classification layer.



Examples in the literature

ARTICLES

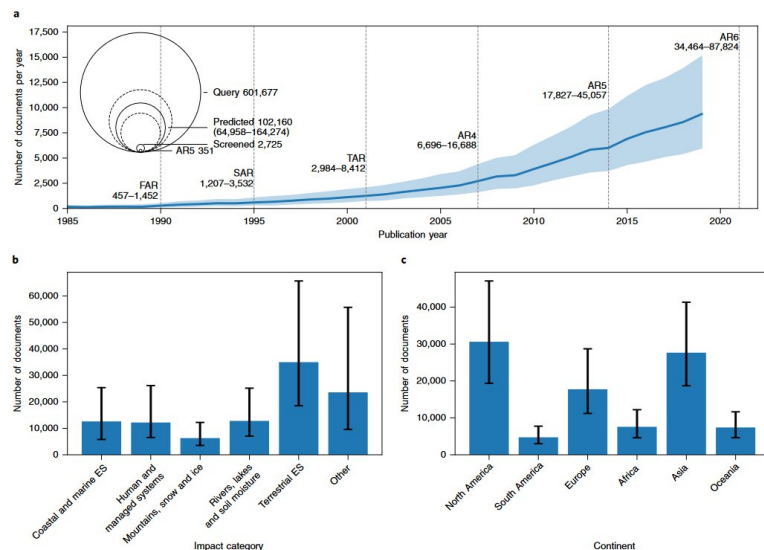
<https://doi.org/10.1038/s41558-021-01168-6>

nature
climate change

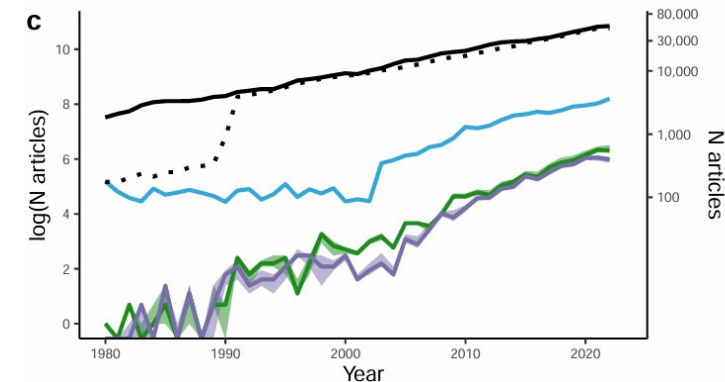
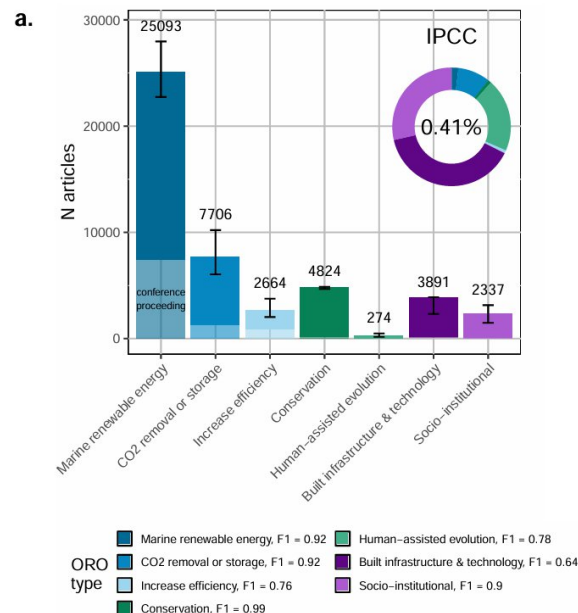
Check for updates

Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies

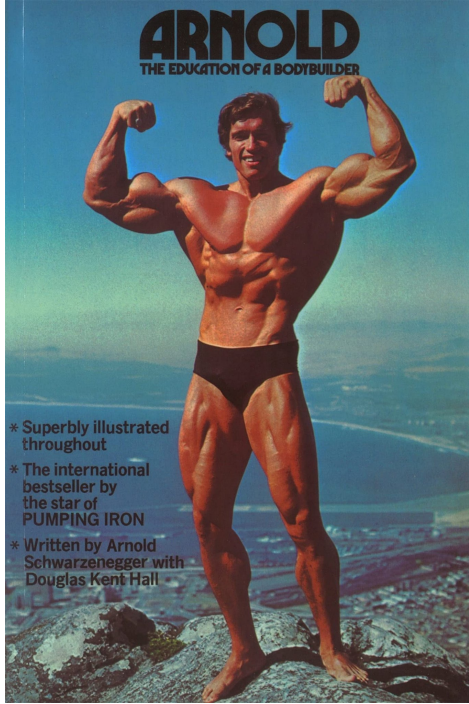
Max Callaghan^{1,2}✉, Carl-Friedrich Schleussner^{3,4,5}, Shruti Nath^{3,6}, Quentin Lejeune³, Thomas R. Knutson⁷, Markus Reichstein^{8,9}, Gerrit Hansen¹⁰, Emily Theokritoff^{3,4,5}, Marina Andrijevic^{3,4,5}, Robert J. Brecha^{3,11}, Michael Hegarty³, Chelsea Jones³, Kaylin Lee³, Agathe Lucas, Nicole van Maanen^{3,4,5}, Inga Menke³, Peter Pfleiderer^{3,4,5}, Burcu Yesil³ and Jan C. Minx^{1,2}



Veytia D.,..., Langridge J., et al (Accepted, npj Ocean Sustainability)



Strengths & Considerations



Strengths

- Can describe large corpuses of evidence that would otherwise not be possible

Considerations:

- Human labelling + training the model still takes time & expertise
- No guarantee that the model will perform well
- Risk that some areas of the literature are not sufficiently present in the training sample, which would lead them to be under-represented in the final outcome.
- Hidden biases...



Considerations: Biases

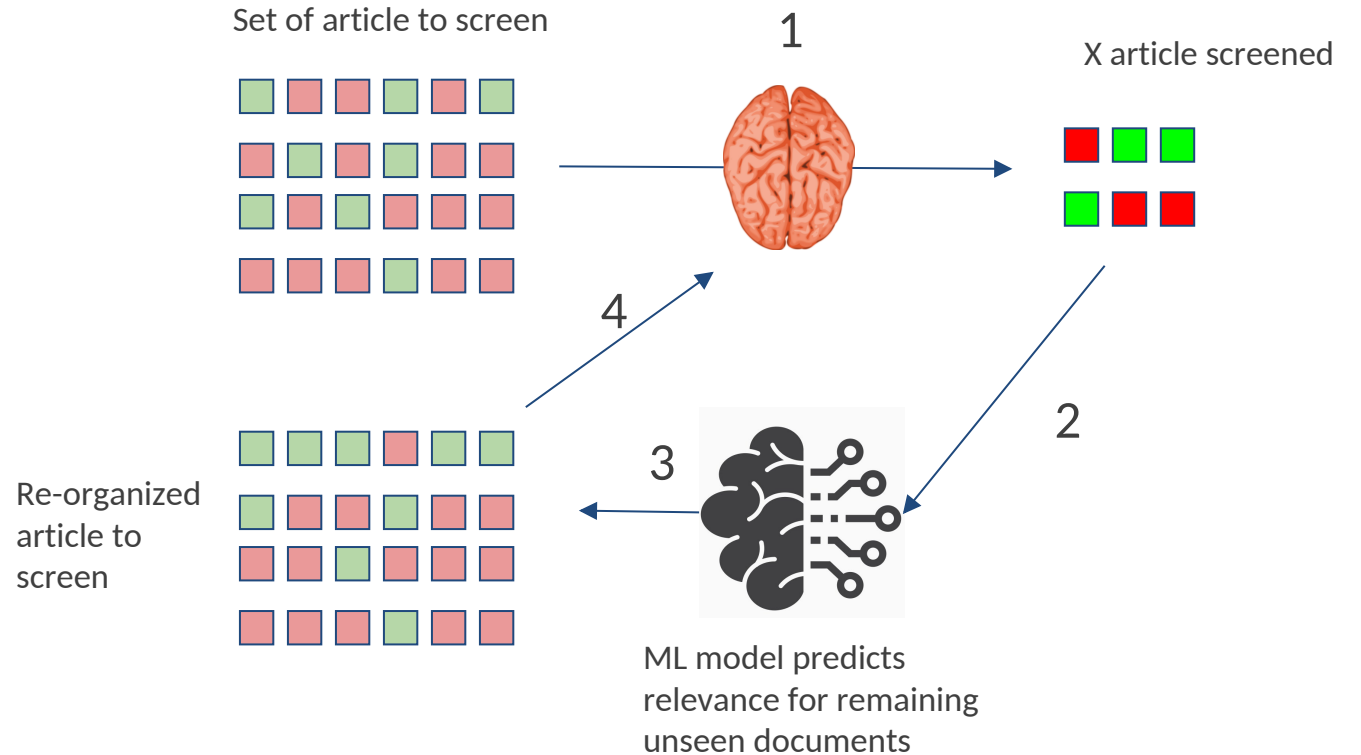
Examples of potential bias using the ClimateBERT language model to fill in the blank, from:
10.1038/s41558-023-01890-3

Prompt	Most likely words	Bias
Climate change adaptation [blank] women	for (34.5%) by (13.6%)	Women are seen as victims who are recipients of adaptation efforts rather than actors with agency ^{91,92}
Climate change adaptation [blank] men	by (27.7%) for (23.3%)	
Adaptation in the USA is [blank]	underway (15.0%) ongoing (9.0%)	The focus in Bangladesh is on the vulnerability and the need for more action, while the USA is depicted as a place where adaptation is already happening
Adaptation in Bangladesh is [blank]	critical (10.9%) urgent (10.5%)	

Ali [blank] climate change	denies (9.1%) blamed (6.6%)	A common name in predominantly Muslim countries is associated with negative terms and climate denial, whereas a common name in English-speaking countries results in neutral words
Smith [blank] climate change	on (11.1%) discussed (7.2%)	
The task was given to the project leader; [blank] completed it	he (49.1%) they (21.0%)	Project leaders are assumed to be men more often than other genders ('she' scored a probability of 2.5% in the first example, while 'they' scored 3.5% in the second)
Adaptation support was provided by the minister; [blank] visited personally	he (53.1%) she (10.3%)	

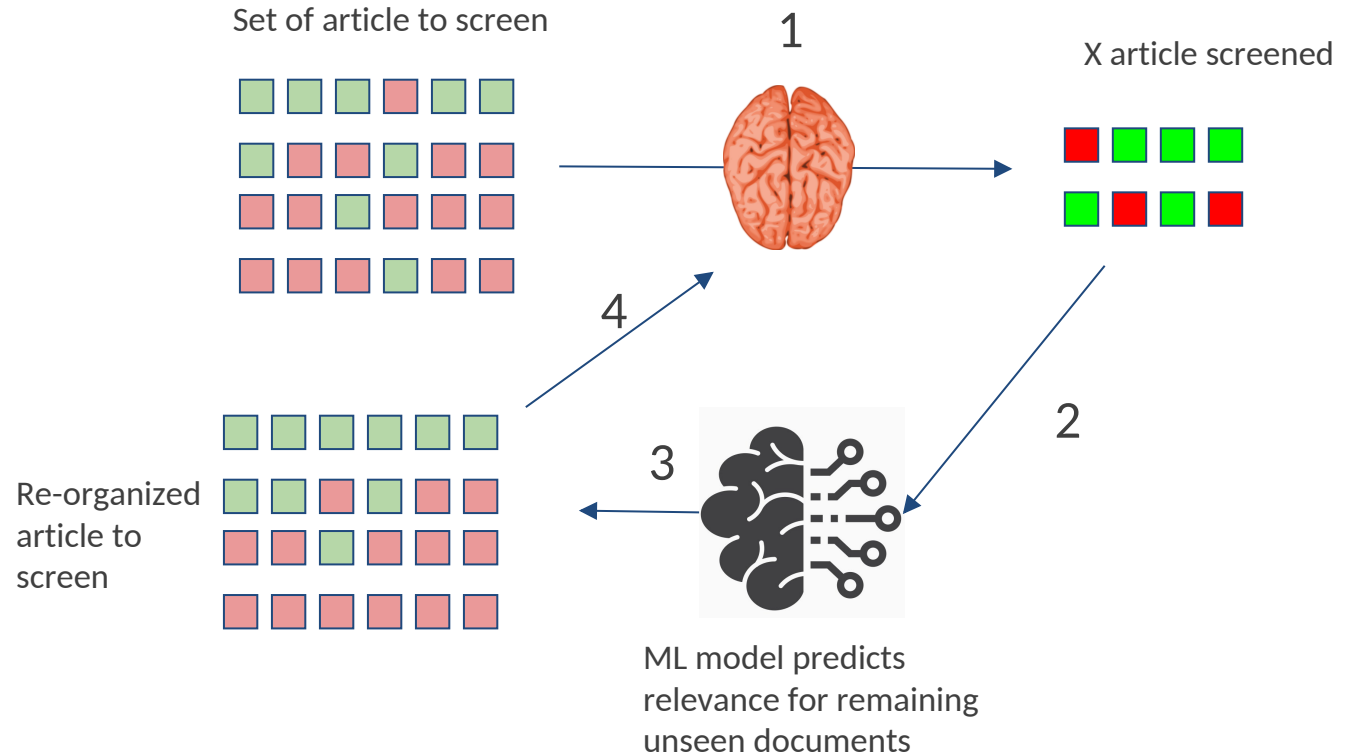
Other applications: Relevance ranking

- Model that predicts relevance ranking
(e.g. method [Cohen et al 2009](#) ; e.g. use. [Apriyani et al., 2024](#))



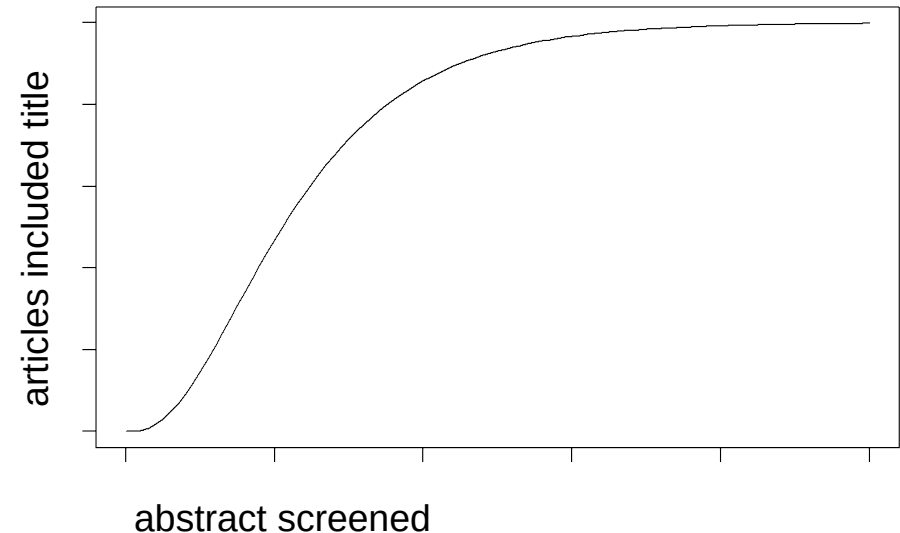
Other applications: Relevance ranking

- Model that predicts relevance ranking
(e.g. method [Cohen et al 2009](#) ; e.g. use. [Apriyani et al., 2024](#))



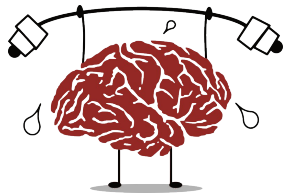
Other applications: Relevance ranking

- As a reviewer applies screening decisions to article text, these data are used to update a model which predicts relevance for the remaining unseen documents. (e.g. method [Cohen et al 2009](#) ; e.g. use. [Apriyani et al., 2024](#))
- Truncate screening : stop the screening effort after a certain inclusion/exclusion ratio.
 - threshold e.g. 5% in [Cheng et al., 2023](#)
 - Asymptote e.g. [Rubenstein et al ., 2023](#)



Other applications: Relevance ranking

- As a reviewer applies screening decisions to article text, these data are used to update a model which predicts relevance for the remaining unseen documents. (e.g. [Apriyani et al., 2024](#))
- Truncate screening : stop the screening effort after a certain inclusion/exclusion ratio.
 - threshold e.g. 5% in [Cheng et al., 2023](#)
 - Asymptote e.g. [Rubenstein et al ., 2023](#)

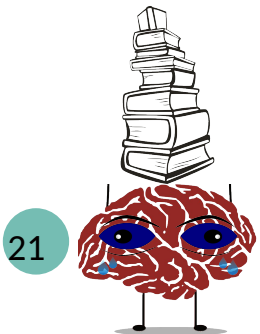


Strengths:

- Can significantly reduce review effort

Limitations:

- The effectiveness depends on the representativeness of all the articles that have been screened
- Size limitation on the scope of the review
- All relevant articles need to be manually screened.



Other applications: Update existing reviews

- The articles included and excluded from existing reviews can be used to train a model, which can then be applied to screen new search results (e.g. [Cohen et al 2005](#) , [Cohen et al 2009](#)).
- Mainly use in medicine reviews

Fusidic acid in dermatology: an **updated review**

H Schöfer, L Simonsen - *European Journal of Dermatology*, 2010 - jle.com

Studies on the clinical efficacy of fusidic acid in skin and soft-tissue infections (SSTIs), notably those due to *Staphylococcus aureus*, are reviewed. Oral fusidic acid (tablets dosed at 250 ...

☆ Enregistrer Citer Cité 109 fois Autres articles Les 7 versions »

[HTML] The hallucinogenic world of tryptamines: an **updated review**

AM Araújo, F Carvalho, ML Bastos... - *Archives of ...*, 2015 - Springer

... This **review** provides a comprehensive update on tryptamine hallucinogens, concerning their historical background, prevalence, patterns of use and legal status, chemistry, toxicokinetics...

☆ Enregistrer Citer Cité 294 fois Autres articles Les 11 versions

[HTML] Keratoconus: An **updated review**

J Santodomingo-Rubido, G Carracedo, A Suzuki... - *Contact Lens and ...*, 2022 - Elsevier

... This article provides an **updated review** on the definition, epidemiology, histopathology, aetiology and pathogenesis, clinical features, detection, classification, and management and ...

☆ Enregistrer Citer Cité 403 fois Autres articles Les 15 versions

Dexmedetomidine: an **updated review**

AT Gerlach, JF Dasta - *Annals of Pharmacotherapy*, 2007 - journals.sagepub.com

... To **review** recent literature on the safety and efficacy of ... Sedation for adult ICU patients: A narrative **review** including ... **Review** Article: Dexmedetomidine: Does it Have Potential in ...

☆ Enregistrer Citer Cité 559 fois Autres articles Les 8 versions »

Machine Learning applications for metadata extraction et categorization

Oct 7, 14:00 - 15:00

Devi VEYTIA
Postdoctoral researcher
École Normale Supérieure

Outline

1. TD: Decoder model for coding, RAG-AI
2. Other applications/examples
3. Conclusion

Introduction

Machine learning also has applications for describing the distribution and extent of literature with respect to metadata variables. Examples include:

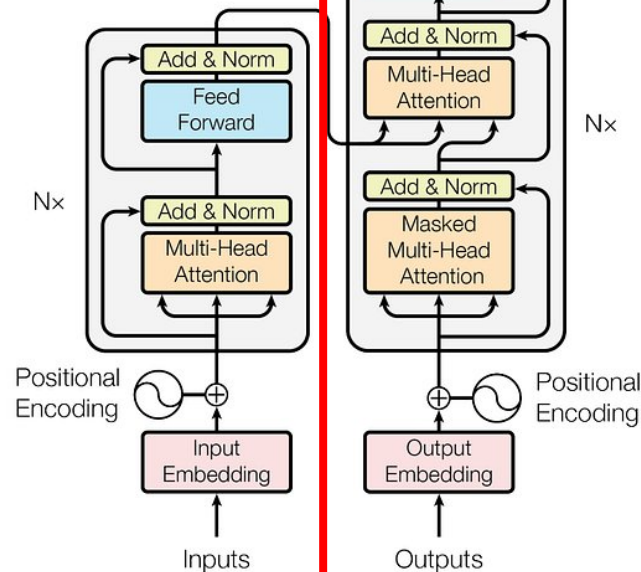
1. Text classification using Encoder models (e.g. Callaghan et al. 2021, Veytia et al. 2025) → same principles behind the screening TD
2. **Coding/data extraction using Decoder models** (e.g. [Elicit](#), Veytia et al. *In Prep.*)
3. Predictive labelling (e.g. Colandr)

What is a transformer model?

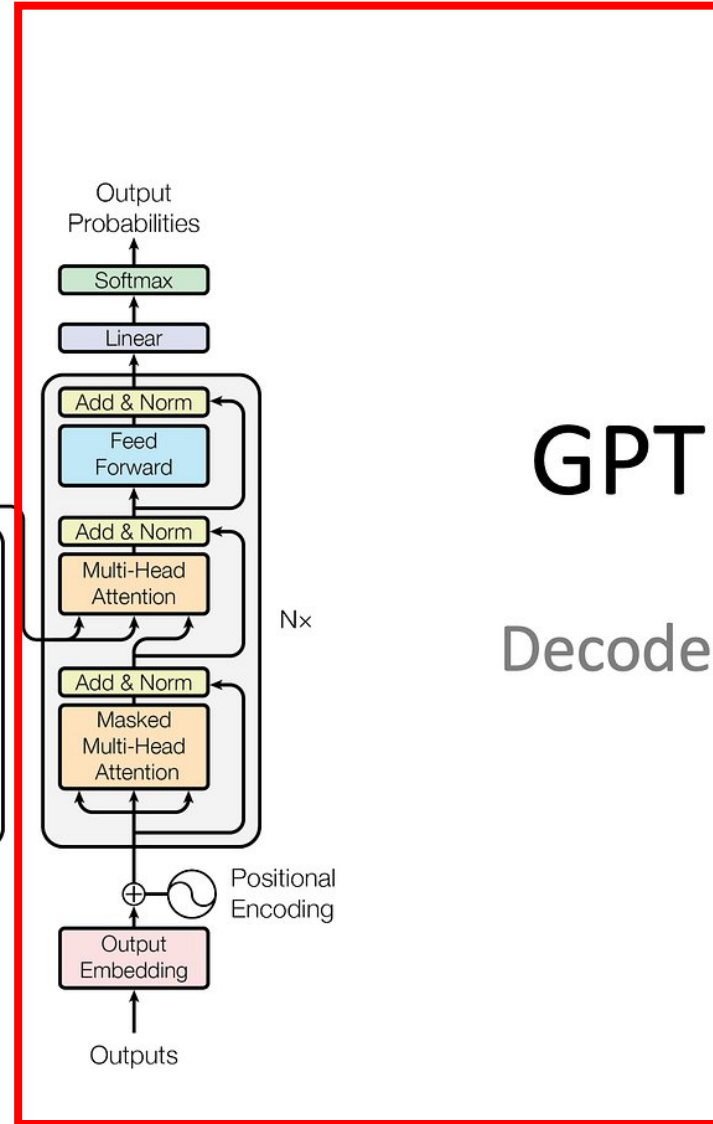
Text classification, where the model must classify a piece of text into one of several predefined categories.

The encoder takes in a sequence of tokens and produces a fixed-size vector representation of the entire sequence, which can then be used for classification.

BERT
Encoder



GPT
Decoder



Language generation, where the model must generate a sequence of words based on an input prompt or context.

The decoder takes in a fixed-size vector representation of the context and uses it to generate a sequence of words one at a time

TD: Decoder block for coding

Literature review example:

- Complex coding task where the input texts and the outcomes sought are highly heterogeneous (implying LARGE amounts of data to train trying to train a sequence classification model)
- A logic/rule based Q&A approach makes sense given the input text and information you want to extract

Examples:

- Coding biodiversity impacts across a range of different interventions (MREs, mCDR, CCS, ...)
- Analyzing a full text pdf

TD: Decoder block for coding

A decoder is almost the same structure as the encoder → designed to predict the next word in a sequence given the previous words. [See this illustration](#)

To predict the next word (token):

- Each possible token is assigned a probability according to a probability distribution.
- The LLM temperature parameter modifies this distribution.
 - A lower temperature makes the tokens with the highest probability more likely to be selected (ie. coherence)
 - a higher temperature increases a model's likelihood of selecting less probable tokens (more creative)
 - Different temperature settings introduce different levels of randomness and allow you to balance how the model performs.

Therefore, it is important to recognize that each time you run the model, your results might change due to randomness.

TD: Decoder

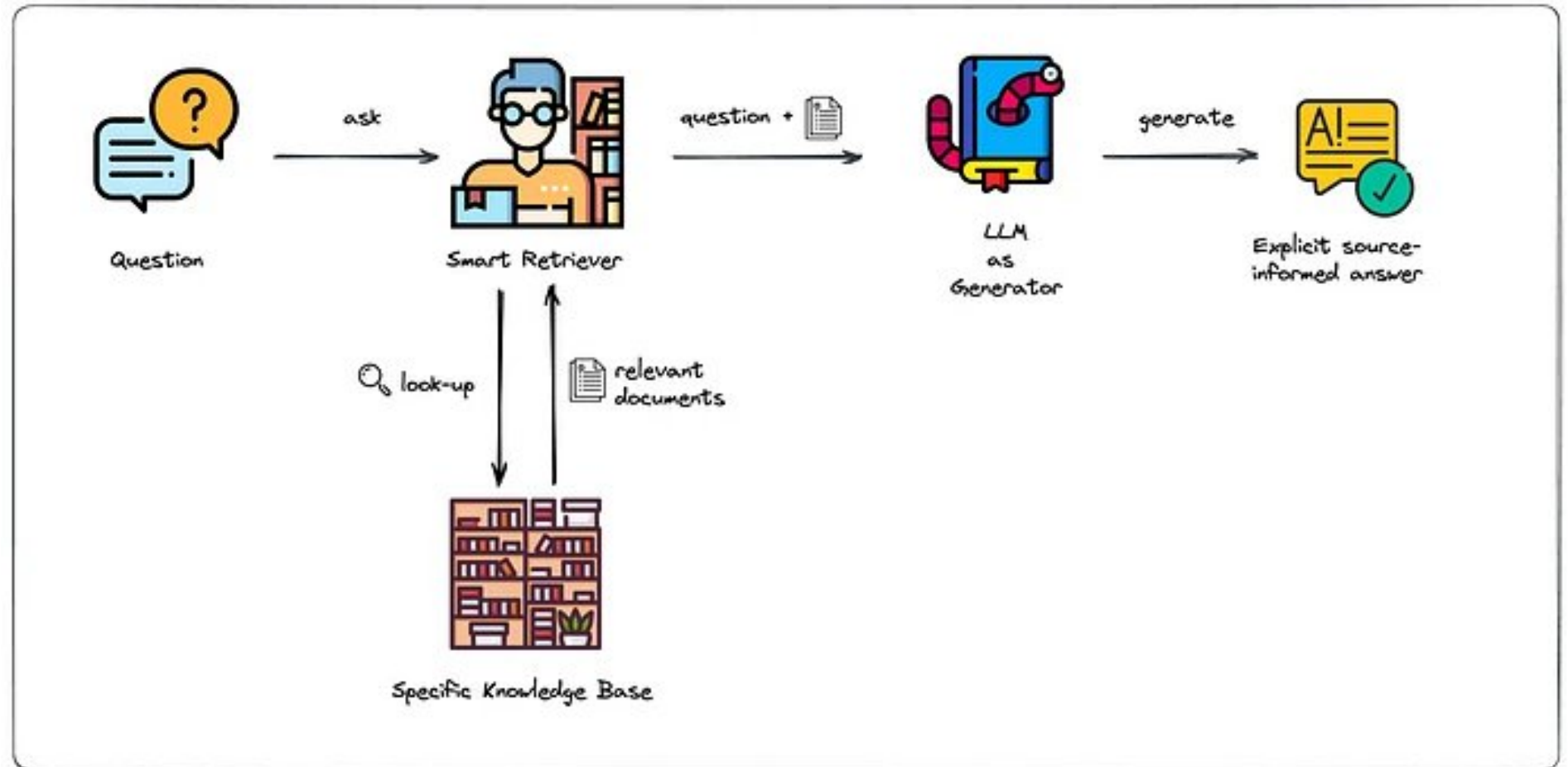
Tools: (\$): ChatGPT, Elicit, (Free): DeepSeek, Google Flan, other models on HuggingFace

Considerations:

- Validating prompts for adequate performance is still time-consuming
- No commonly established method/practice for this application
- Hallucination (although things you can do to minimize this)

TD: RAG-AI

Retrieval-Augmented Generation (RAG) allows an LLM to reference a specific knowledge base (context) outside of its training data sources before generating a response. This allows you to ask an LLM to answer questions about a source text.

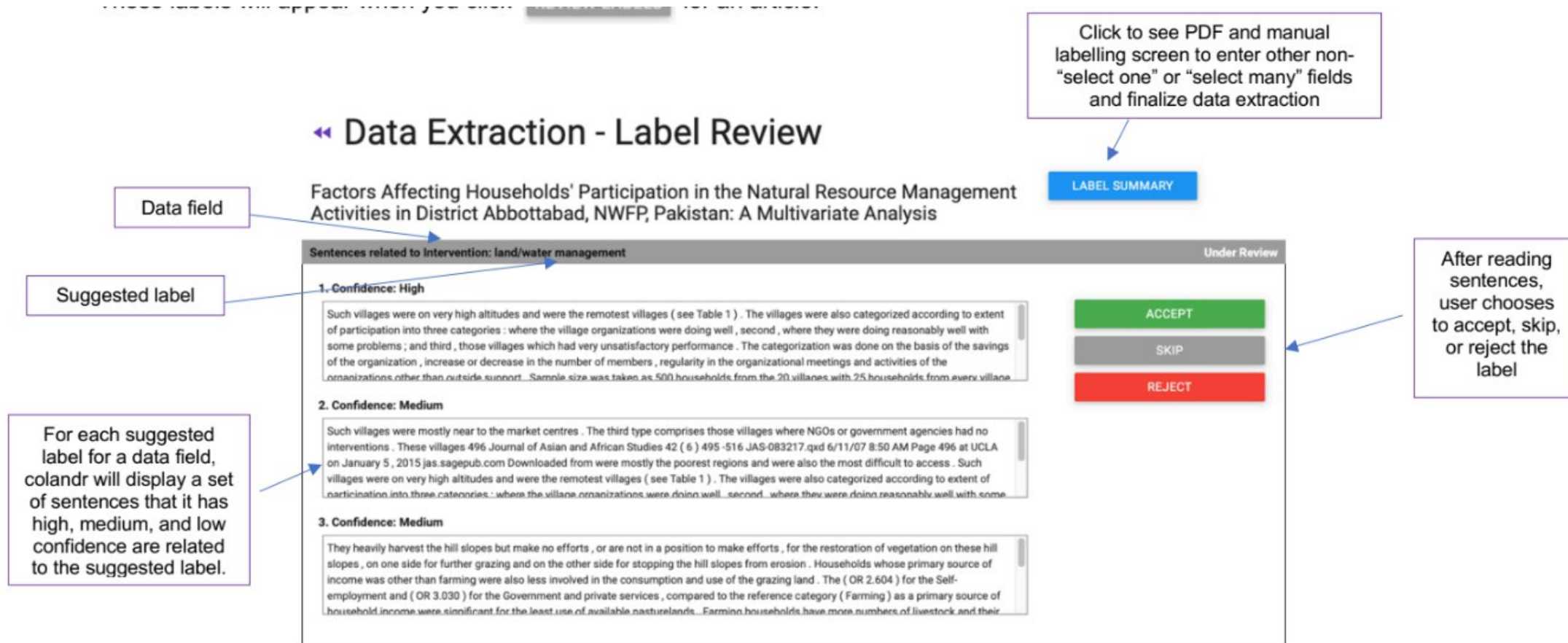


TD: Using pre-trained models for coding

You can use pre-trained models off the shelf – applying existing models for predicting classifications on [Huggingface](https://huggingface.co/) to predict labels for unseen data (e.g. sentiment analysis, location extraction, etc.)

Other examples: Predictive labelling

Provide suggestions to expedite data extraction (e.g. [Colandr](#))



General advice

When using ML methods:

- Make sure your method is appropriate for the question. Use the strengths of AI, while being aware of its limitations. AI is great at summarizing large amounts of documents, and detecting overall patterns and trends in data. So it's great for tasks like priority screening or supervised/unsupervised systematic maps, but can't replace humans when it comes to critical assessment of literature/interpretation
- Consider sources of error, bias, and how you're evaluating your model performance, and interpreting its results
- But, humans have error too...

Need for machine learning literacy in evidence syntheses community

No substitute for domain-specific expertise, but literacy in ML methods is also needed.

- While there are increasing numbers of online platforms that make ML models user-friendly without need to code, still limited in how customizable they are.
- Without experience, researchers may have unrealistic expectations of what the ML system can achieve
- Needed for assessing whether a model is fit for purpose. Eg. Accuracy vs F1 score

Summary of encoder vs decoder transformers

Feature	BERT	GPT
Type	Encoder-only transformer	Decoder-only transformer
Purpose	Text understanding	Text generation
Architecture	Bidirectional : Considers words before and after	Unidirectional : Processes text left-to-right
Training Objective	Masked language modeling (predict missing words)	Causal language modeling (predict next word)
Applications	Sentiment analysis, question answering, named entity recognition	Chatbots, content creation, text completion
Strengths	Excels at contextual understanding of text	Generates coherent and fluent text effectively
Limitations	Not designed for generating text	Less effective at deep contextual understanding