AI as a tool in evidence synthesis

James Thomas

EPPI Centre
Evidence for Policy & Practice

About me

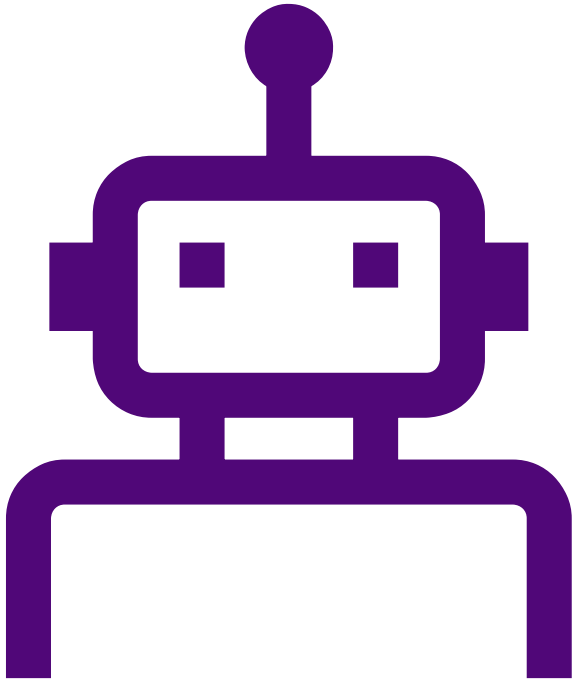
- Co-director of EPPI Centre, UCL
- Do a wide variety of evidence synthesis – mostly for Department of Health & Social Care
  - Addressing questions beyond effectiveness
  - Methodological development
- Evidence synthesis methods
- Long-standing area of work in making the review process more efficient using new technologies

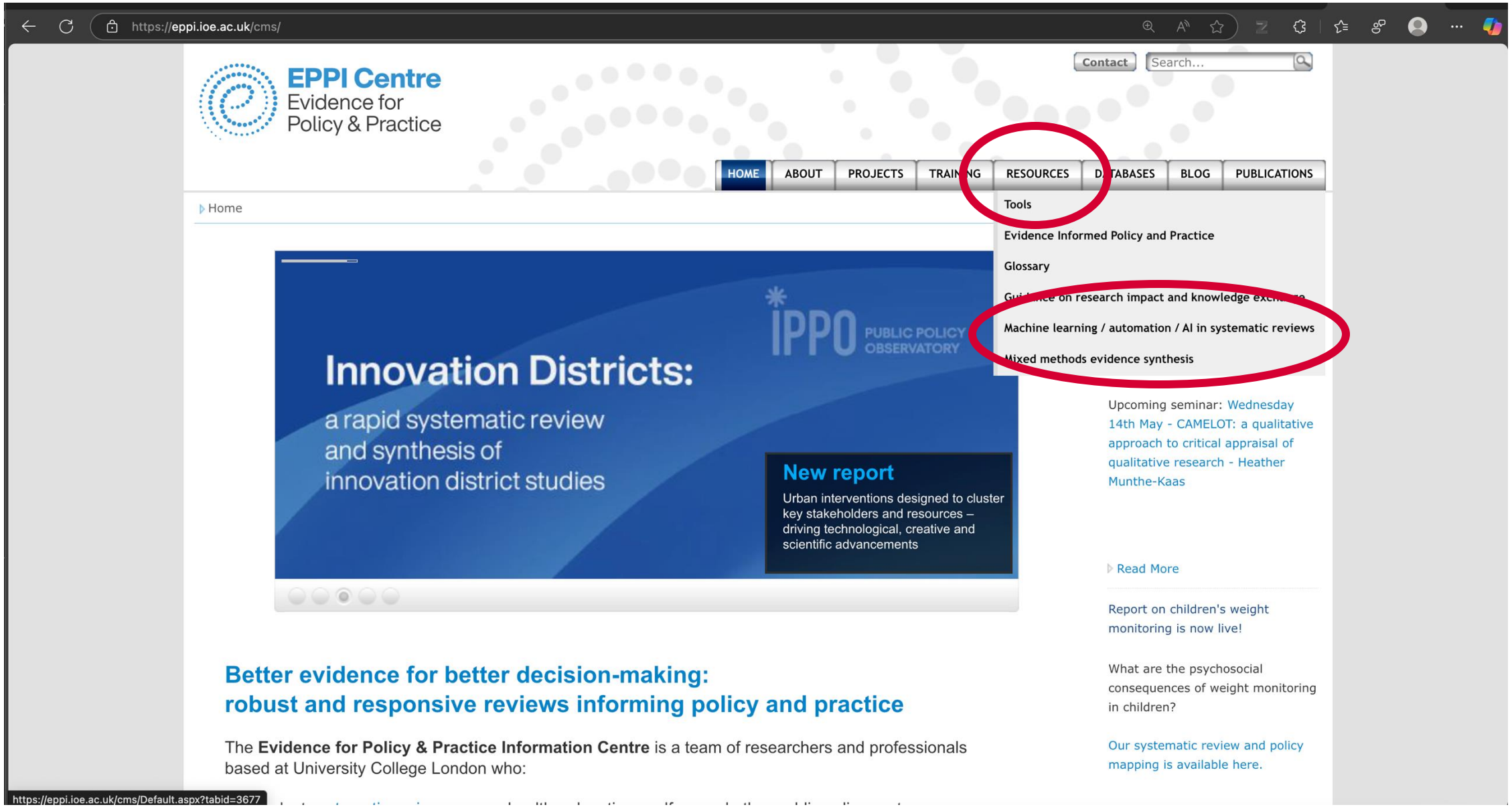
Acknowledgements and declaration of interests

- I am employed by University College London; receive funding from the funders below for this and related work; lead EPPI-Reviewer software development
- Cochrane roles: Review author; Co-convenor Joint Artificial Intelligence Methods Group[NEW!]; Co-Senior Scientific Editor Cochrane Handbook; support Cochrane with information technologies (EPPI-Reviewer and machine learning)
- Guidance for responsible use of AI in systematic reviews (RAISE)
- Parts of this work funded by: Wellcome Trust, National Institute for Health Research (NIHR), Education Endowment Foundation, Youth Endowment Foundation

In this session

- Introduction to AI / machine learning / automation tools for evidence synthesis (and how they work)
- Hands-on experimentation with LLM tools for evidence synthesis
- Please feel free to ask questions as we go





https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677

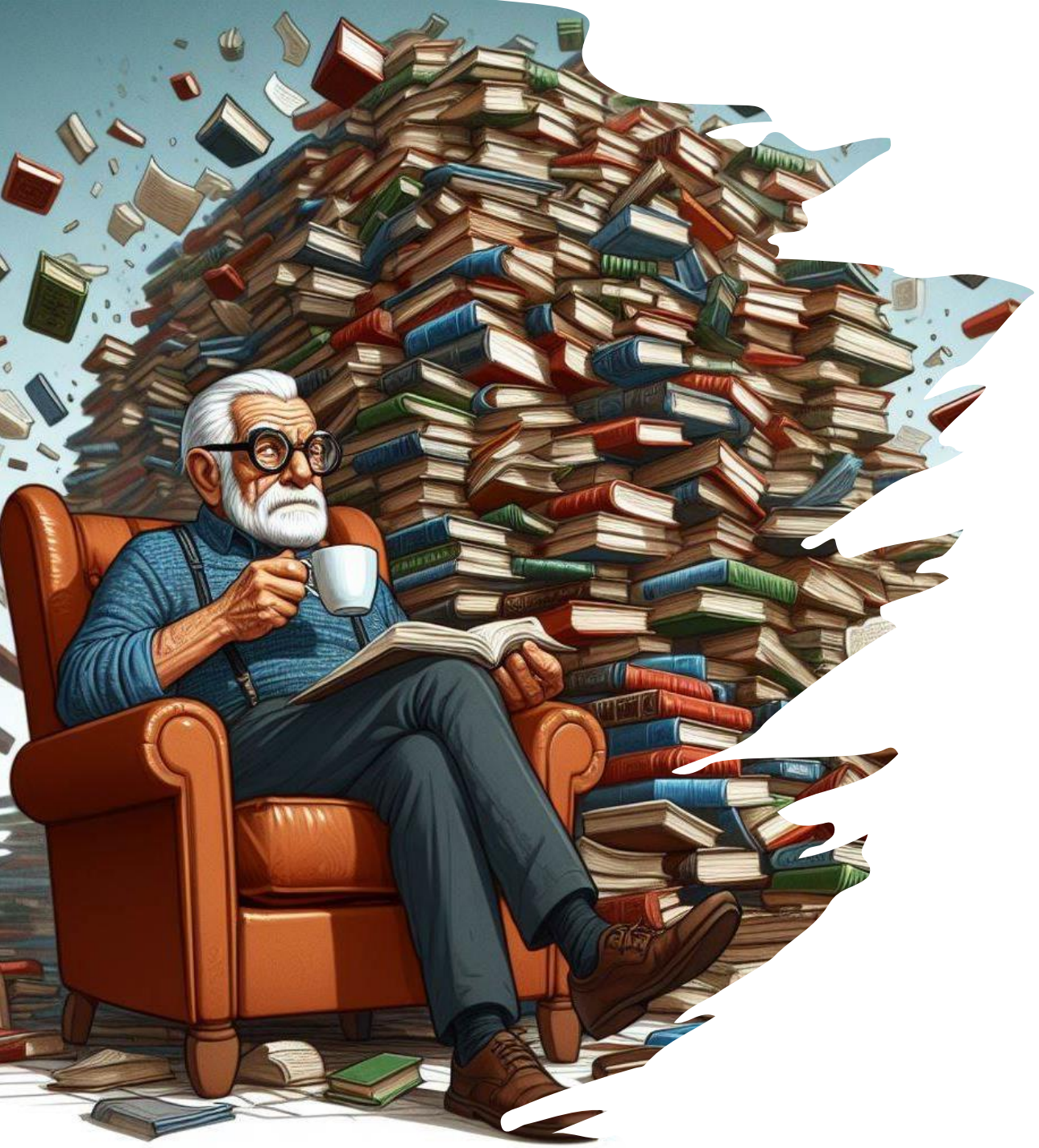
Evidence synthesis priorities

Evidence syntheses are often used to inform decisions that affect people's lives

Evidence synthesists favour accuracy over efficiency

Highly sensitive searches are required to avoid selection bias

Highly accurate quality assurance processes are required to avoid human error

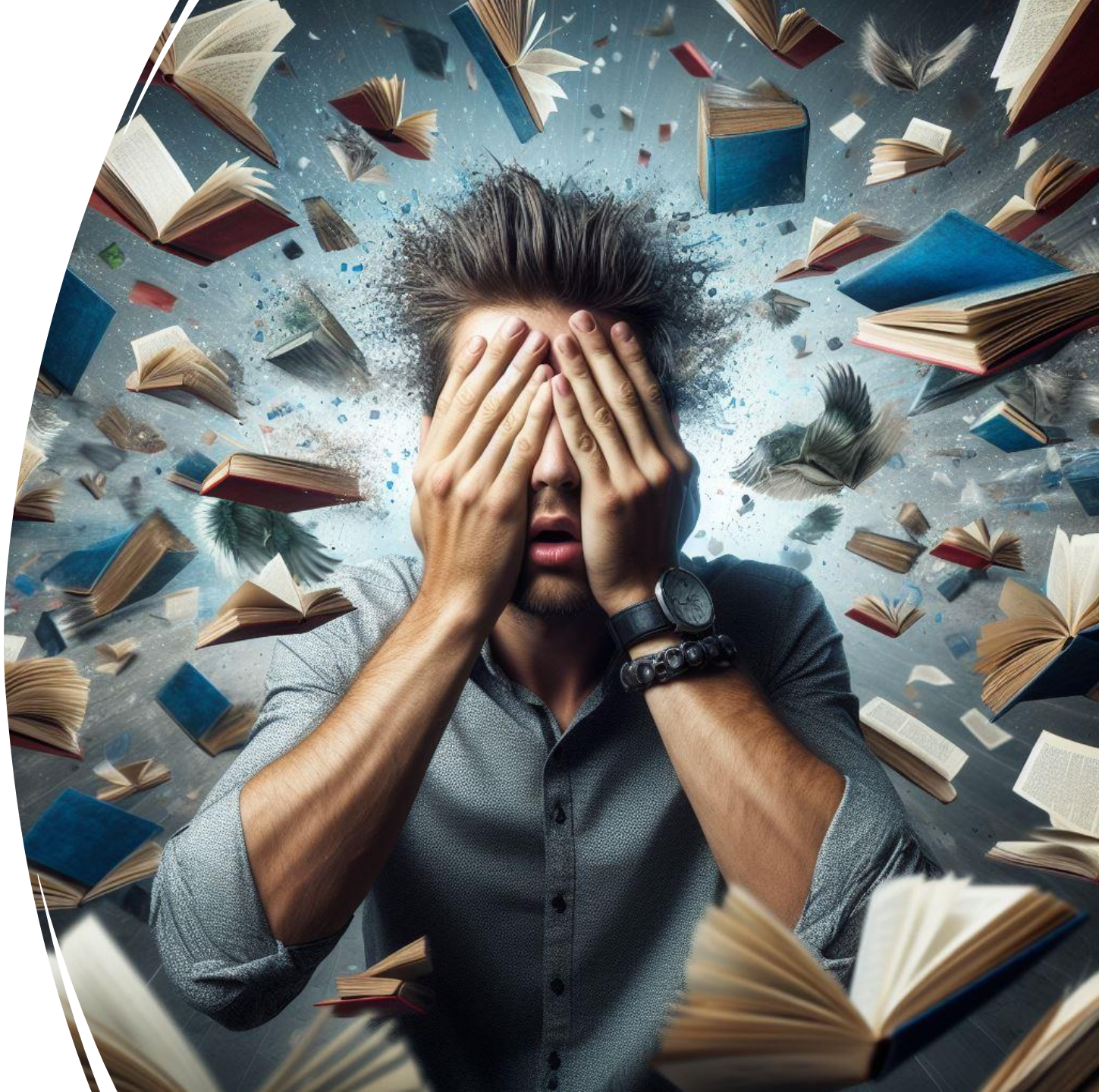


Impact of these priorities

- An inefficient, resource-intensive process has evolved that produces reliable, but expensive and time consuming, reviews
- We cannot keep pace with the deluge of new research being published
- E.g. in the Cochrane Reviews published March 2014, > 163k citations were screened; 6,599 full text reports were read; and 703 studies were included
- That’s about 2 million records per year

This means

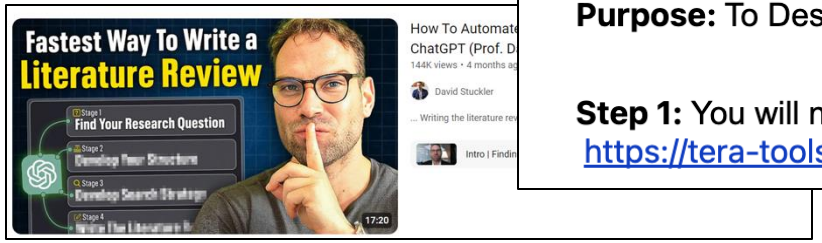
- Only a fraction of available studies are included in evidence syntheses
- Evidence synthesis does not cover all questions/ domains comprehensively
- We don’t even know when reviews *need* to be updated



Are there AI tools we can use?

Are there AI tools we can use?

Are there AI tools we can use?



INSTITUTE FOR
Evidence-Based Healthcare

TERA

Review Wizard

Scope of action: Write the method section for your review.
Purpose: To Design and write your review methods section.

Step 1: You will need to be logged in and click on Review Wizard URL.
https://tera-tools.com/methods-wizard

Are there AI tools we can use?

Consensus

Find the best science, faster.

Consensus Product Update 3/24

We are excited to announce the launch of one of our most requested features ever: **upload and chat with a PDF within Consensus** ✨

Less PDF scrolling, more time-saving analysis. This new feature allows you to apply Consensus's models to your research paper library. Upload and chat with the full-text of your papers to ask about key figures, methodological details, novel insights and more!

This launch marks the start of a string of major changes to Consensus in the coming weeks. Upcoming changes will unlock a whole new level of AI analysis including full-text access, multi-paper upload & analysis, and more!

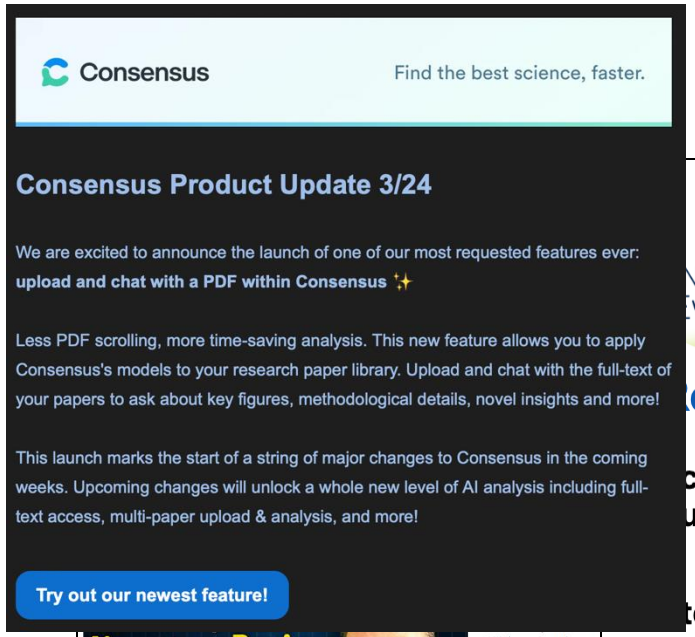
Try out our newest feature!

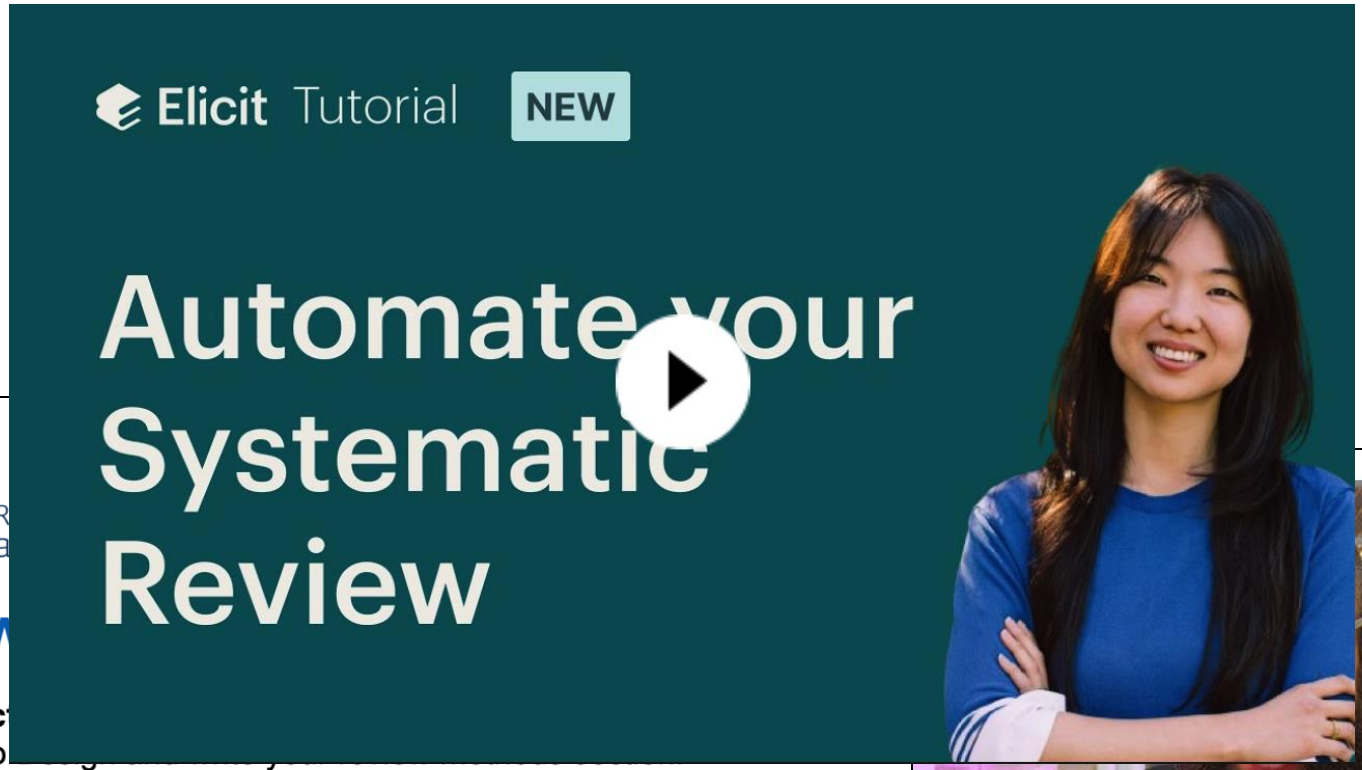
TERA

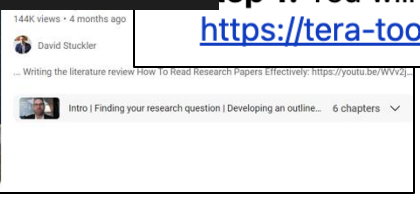
tion for your review.
ew methods section.

d click on Review Wizard URL.
d

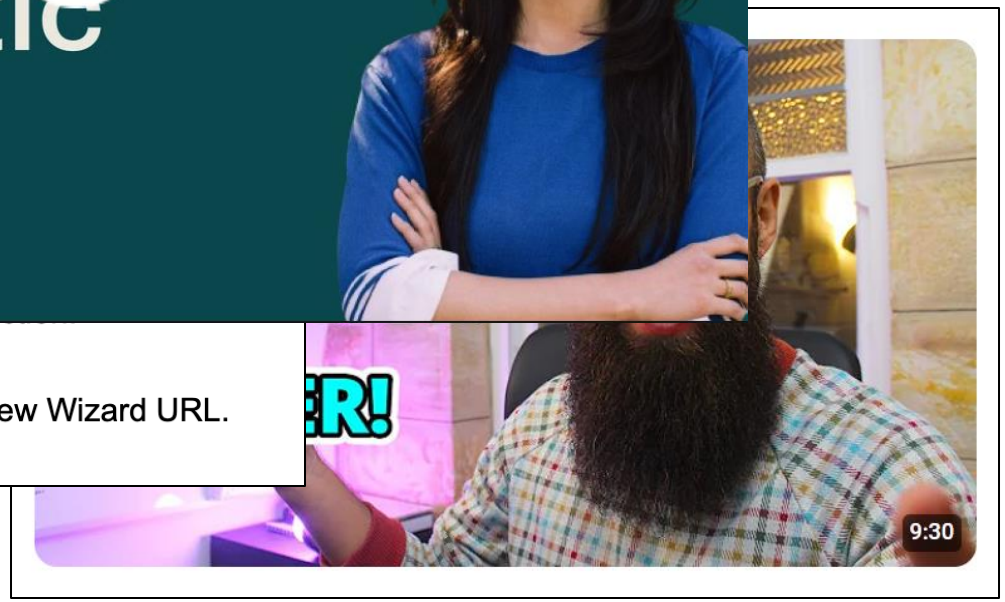
Are there AI tools we can use?



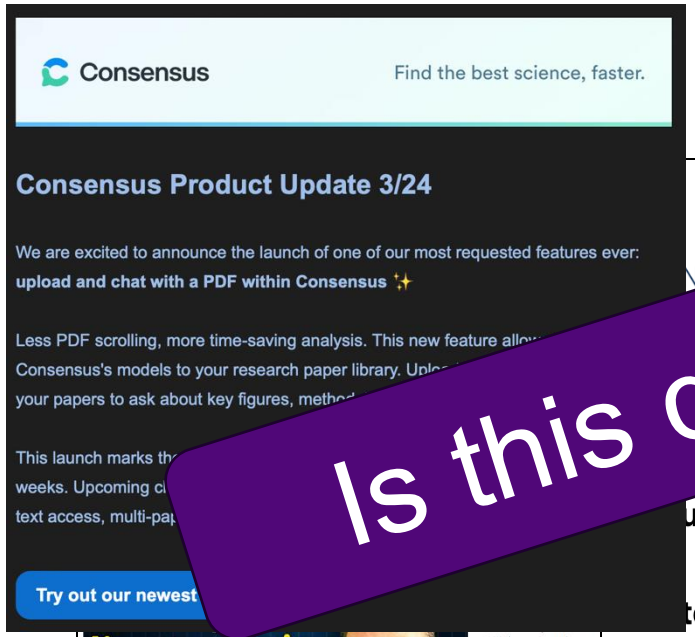




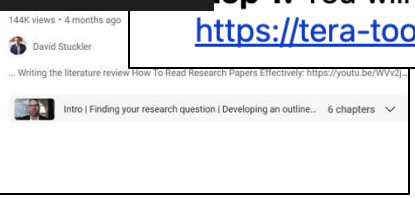
tep 1: You will need to be logged in and click on Review Wizard URL.
https://tera-tools.com/methods-wizard



Are there AI tools we can use?



Is this question out of date now?!



ep 1: You will need to be logged in and click on Review Wizard URL.
https://tera-tools.com/methods-wizard

~~Are there AI tools we can use?~~
There are a lot of AI-based evidence synthesis tools!

- Can we use them?
- Should we use them?
- And are we already being out-evolved if we're not using AI?
- Important to understand a bit about how automation tools work to make good decisions about using them

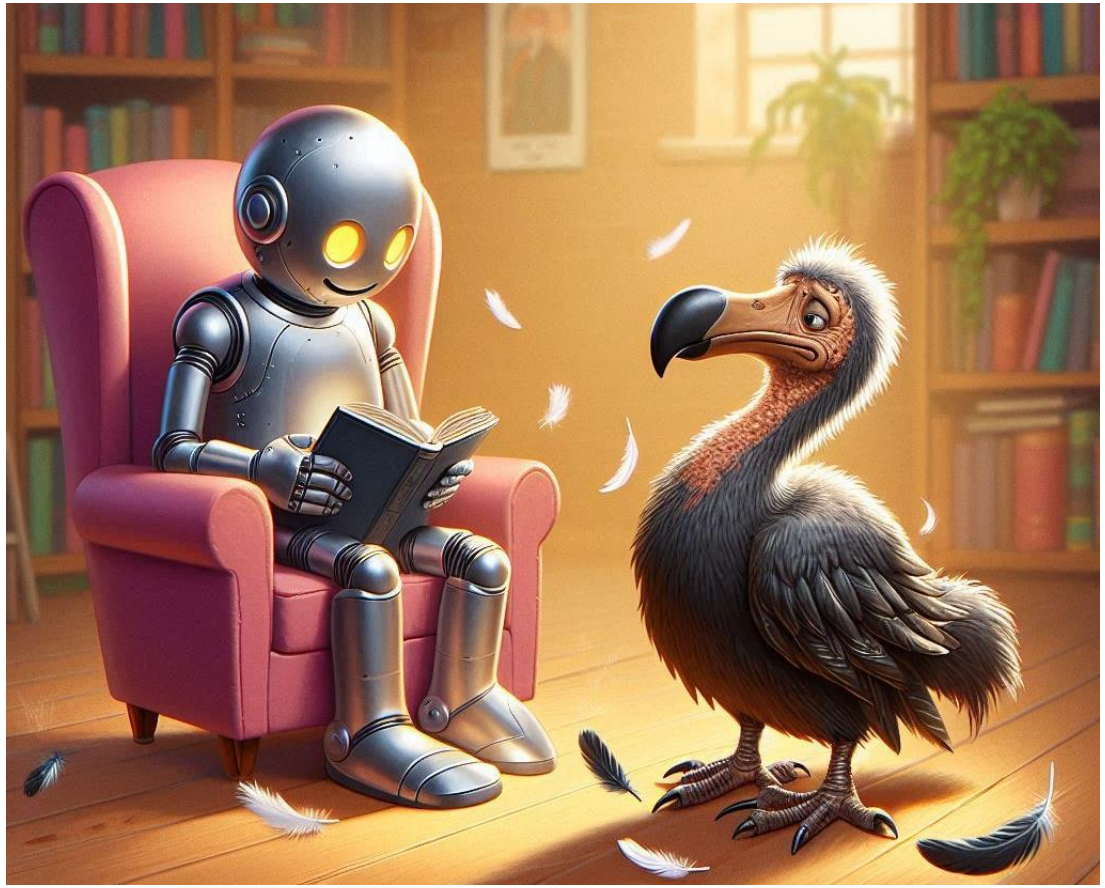


Image generated with the help of Microsoft Copilot

Four machine learning / automation paradigms

- Rules-based approaches
  - (strictly speaking, not *machine learning*)
- Unsupervised approaches
- Supervised approaches
- Generative approaches (‘Gen AI’)

- Covering in terms of technology not purpose, so we can consider their strengths and weaknesses more easily

Rules-based approaches

As you might guess… a set of rules is constructed by humans and given to the machine

For example

Look up a simple set of words	Use of synonyms	If a given phrase is present, apply a given code	Many citation duplicate-checking algorithms

Rules can be accurate… but fragile

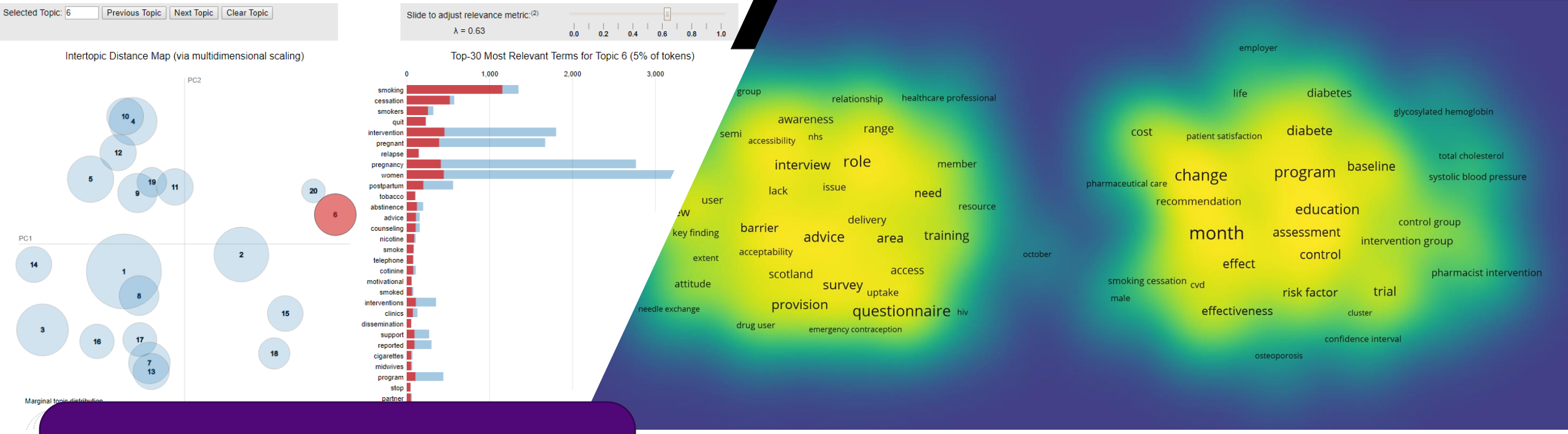
If you stick within the rules, you get the anticipated results

If you stray outside – even a little bit – the rule can fail altogether

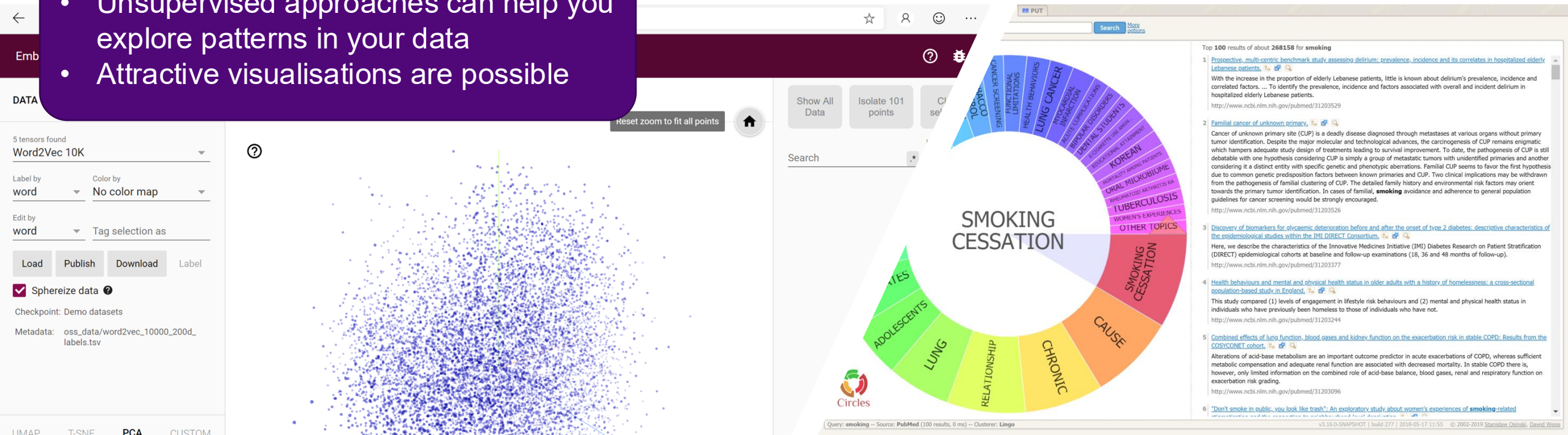
No grey area – it works, or completely fails

Unsupervised approaches

- The machine is given no rules…
- And simply identifies patterns in the data
- E.g.
  - Relationships between words
  - Clustering documents



- Unsupervised approaches can help you explore patterns in your data
- Attractive visualisations are possible



Unsupervised approaches lack control

Very powerful – can reveal relationships in the data which are not necessarily obvious

Very efficient – data often need no preparation

But… you don’t get to tell the machine which classifications to make

Supervised approaches

Humans prepare ‘training’ data – containing data + labels which describe the desired classification

For example

Image recognition

Text classification

Good supervision is required…

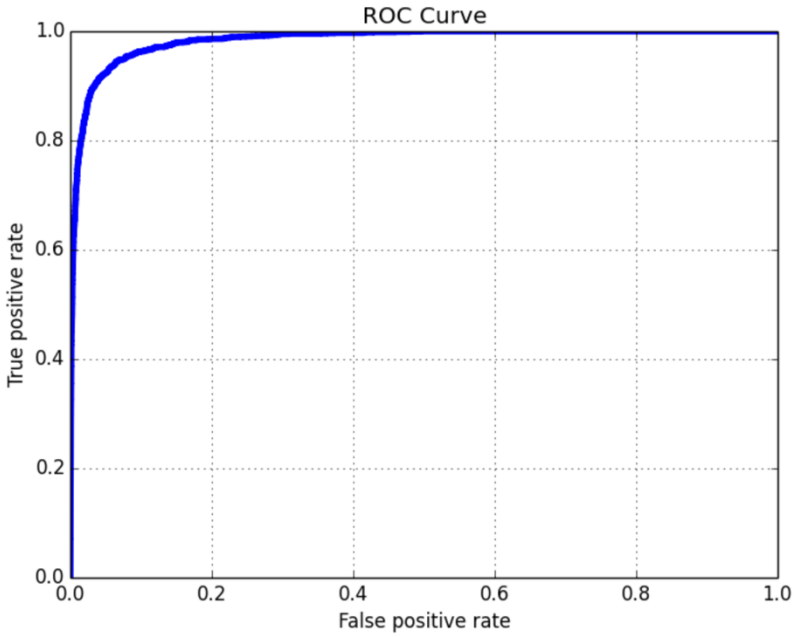
Very dependent on quality and coverage of training data

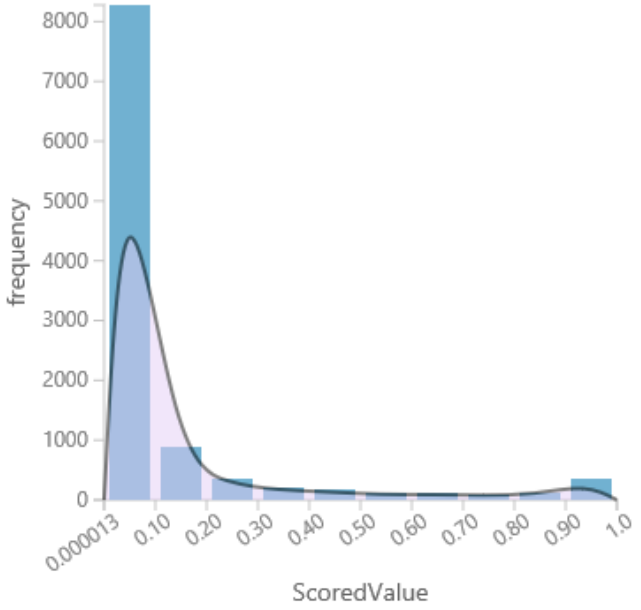
Performance very dependent on context

For example…

Example of study classification: RCT Classifier

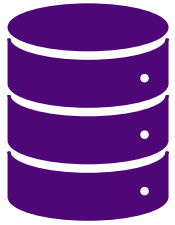
- A classifier was built using more than 280,000 records from Cochrane Crowd
- It is ‘simply’ applying single classification (RCT / not RCT)
- It has been calibrated to achieve a recall = 99% on the McMaster ‘Hedges’ dataset
  - Calibration = ranking the ‘test’ dataset by score
  - BUT precision is low
- It is very accurate!
  - But not all supervised learning can be so accurate, as lots of high-quality training data are needed



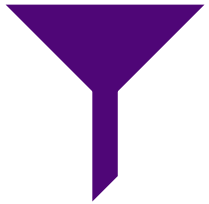


Generative approaches

ChatGPT (or other LLM chatbot)



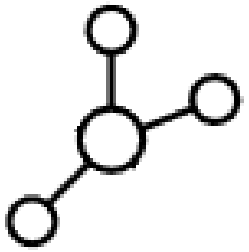
LLM-based database querying and summarisation



LLM-based information extraction

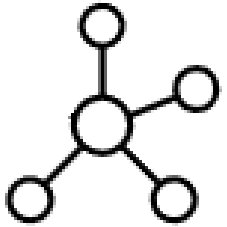
Building a GenAI chatbot

Pretraining
(unsupervised ML))



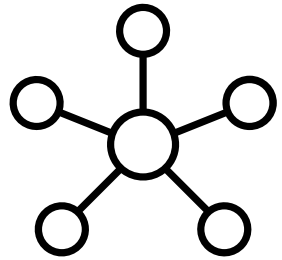
‘Naïve’ model
Cannot ‘chat’; next-word prediction only

Fine-tuning
(supervised ML)



Model can now ‘chat’ and answer questions

RLHF
(supervised ML)

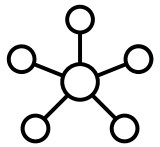


Model produces ‘better’ and less toxic answers

Generative LLM operation

Input

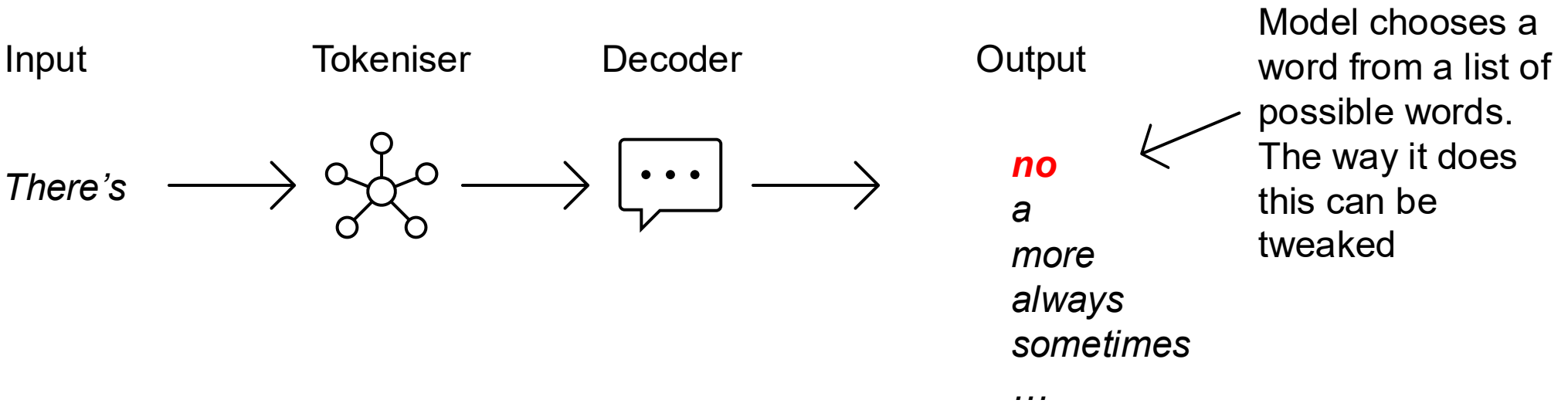
Tokeniser



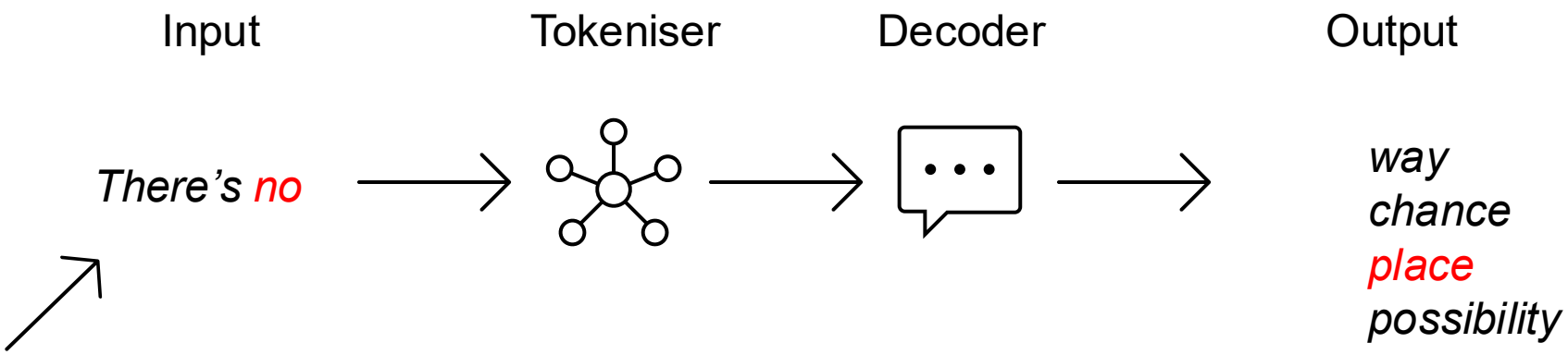
Decoder

Output

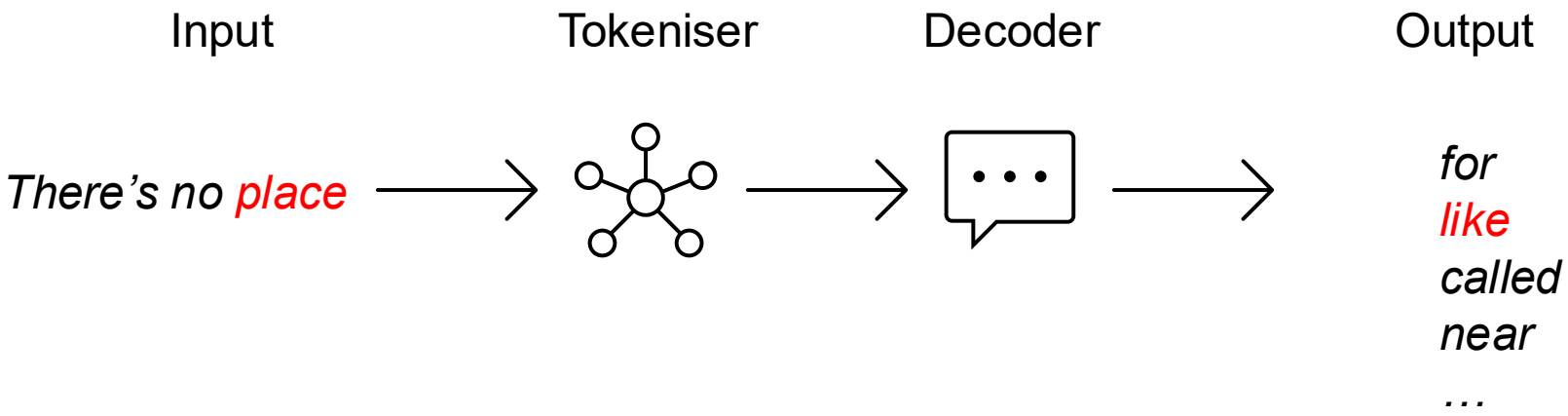
Generative LLM operation



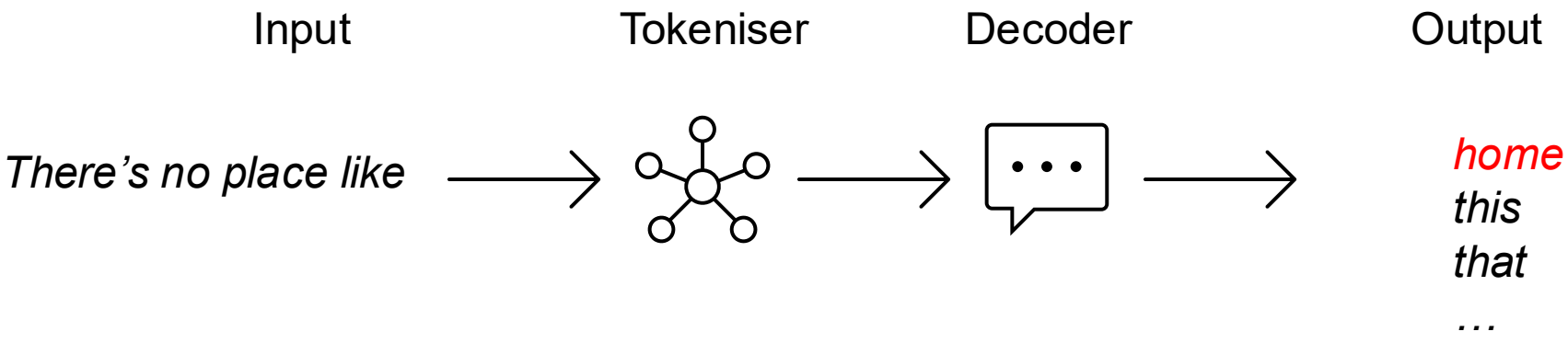
Generative LLM operation



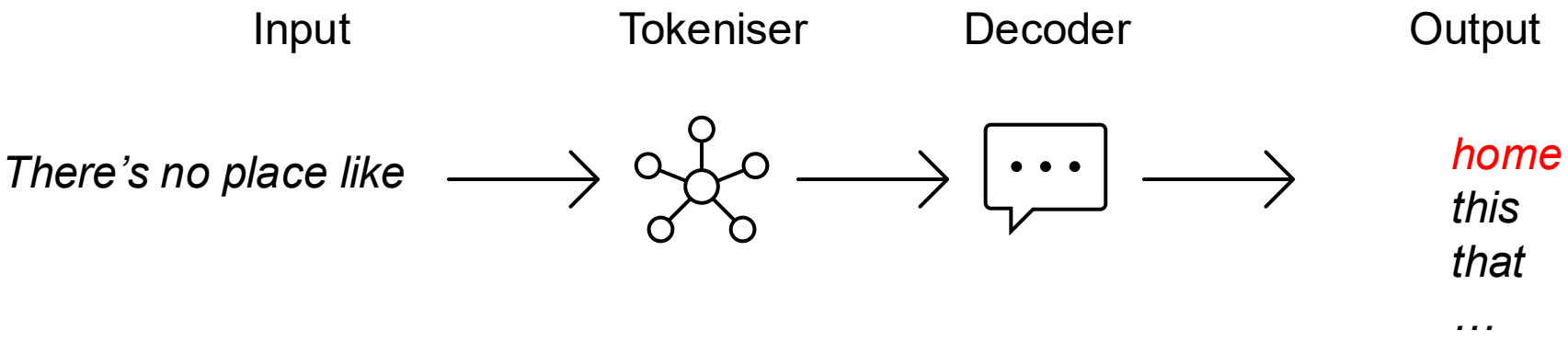
Generative LLM operation



Generative LLM operation

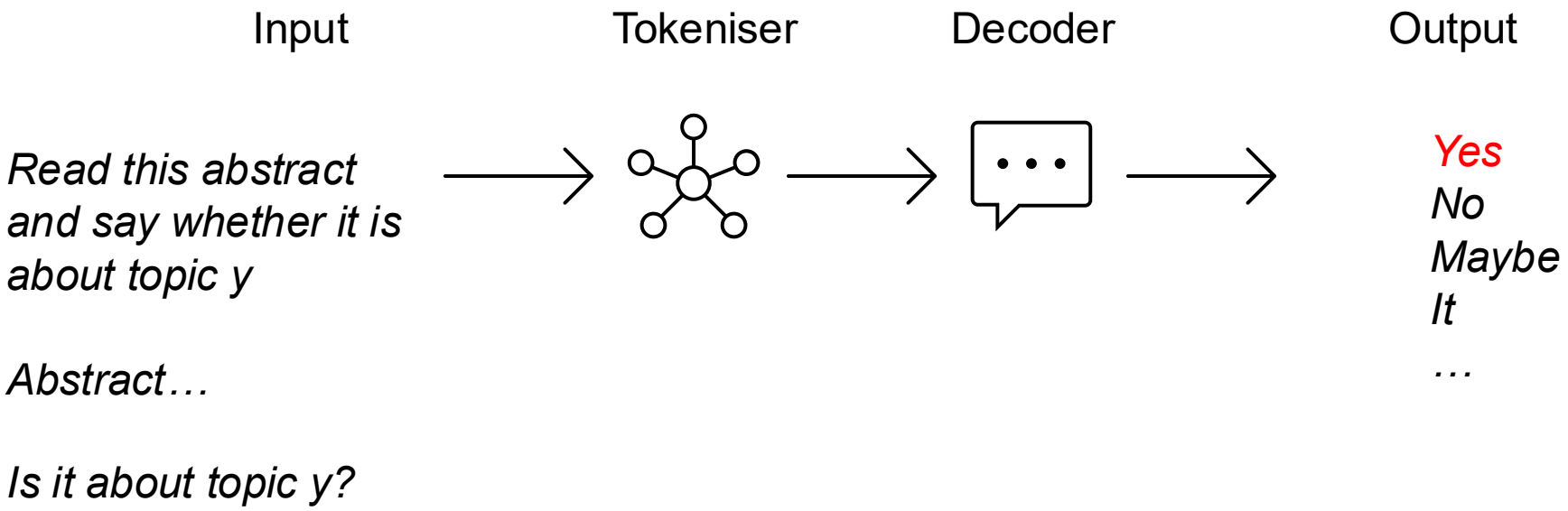


Generative LLM operation



Important to bear in mind that the system does not plan ahead
…and at no point does it check the accuracy of what is 'said'

Generative LLM operation



Instead of the prompt containing '"There's no place like…" it could contain a question about a passage of text that is also in the prompt

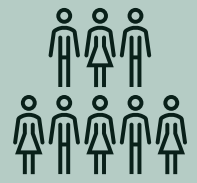
A major contrast with supervised machine learning:

‘zero shot’ or ‘in context’ learning

Image generated with the help of Microsoft Copilot

Why zero-shot learning is a gamechanger

Development and evaluation of the Cochrane RCT Classifier
(Using conventional supervised machine learning)



Conventional machine learning model trained on 280,000 records from Cochrane Crowd

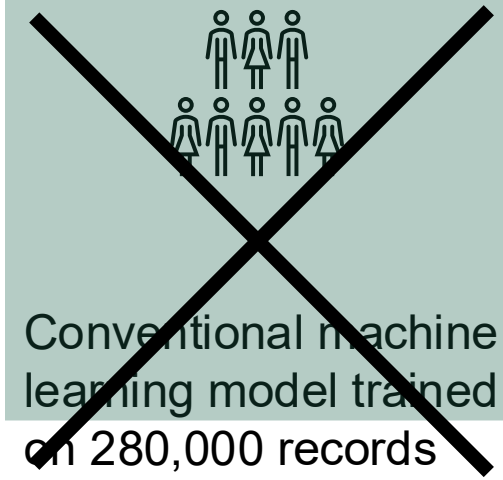
Model was calibrated to achieve 99% recall on a second ('Hedges') dataset (~50,000 records)

Model was validated on 92,000 studies included in Cochrane intervention reviews

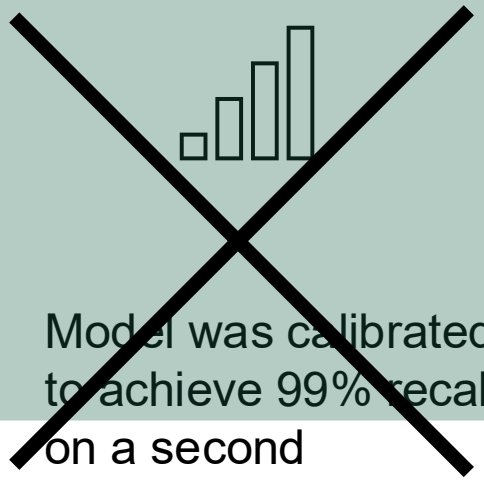
Model was deployed for live use in Cochrane review workflows

Why zero-shot learning is a gamechanger

Development and evaluation of the Cochrane RCT Classifier



Conventional machine learning model trained on 280,000 records from Cochrane Crowd



Model was calibrated to achieve 99% recall on a second ('Hedges') dataset (~50,000 records)

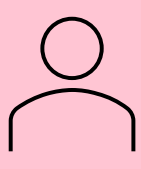
Model was validated on 92,000 studies included in Cochrane intervention reviews

Model was deployed for live use in Cochrane review workflows

With the new AI tools there's no need to create (expensive / hard to find) training data

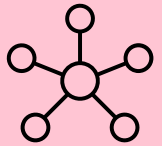
Why zero-shot learning is a gamechanger

Development and evaluation of a classification task using a language model



Instead, a human writes some prompts for a large language model in their normal language

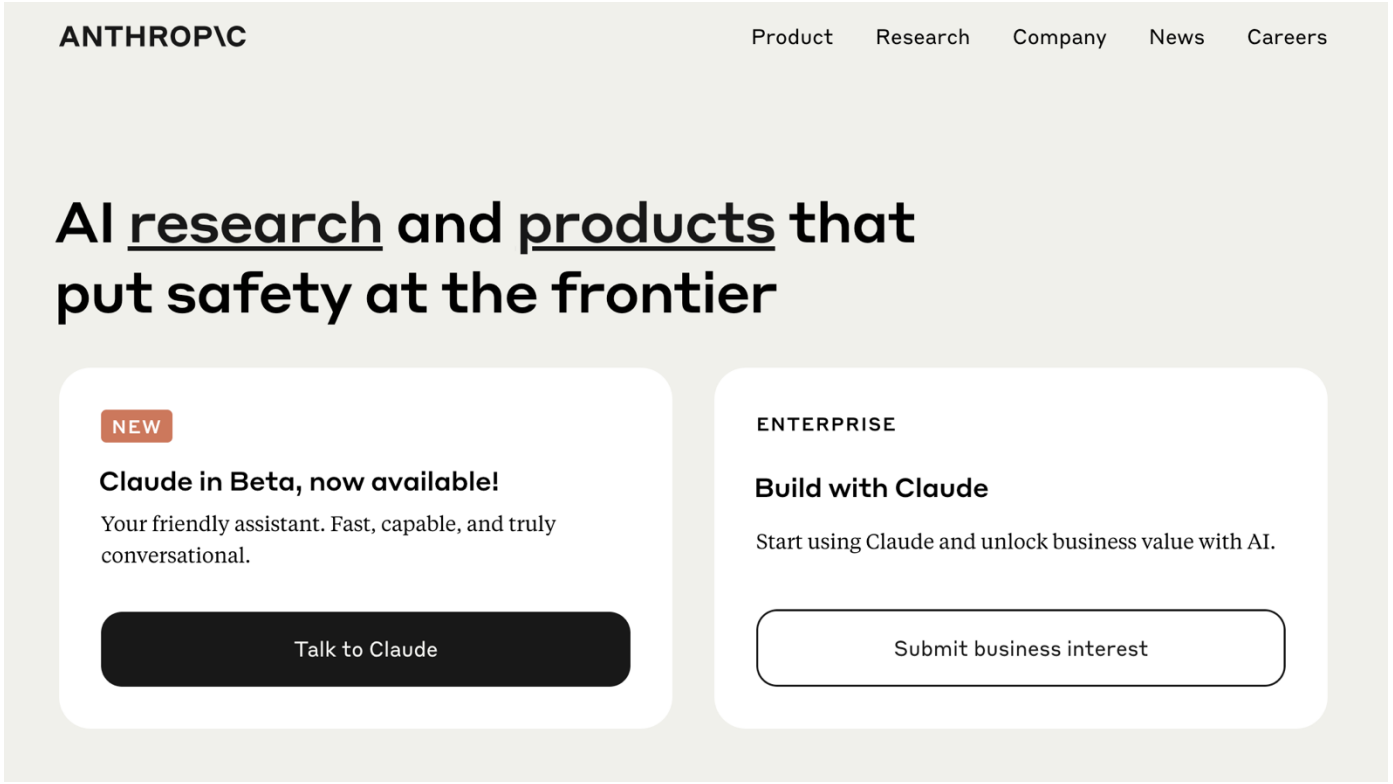
They check they work on their data



The language model can then apply the prompts to the remaining data

Data (information) extraction

- Earlier language models lacked precision & limited context ‘window’
- Newer models have larger windows and offer impressive early results
- E.g. Claude2, published by Anthropic



Smoking Cessation Internet Intervention for Low and High SES Groups

PDF 198.93 KB

Here is a report of a randomized trial. State: the population recruited; the intervention(s) evaluated; the main outcome(s) measured; and the statistical results of the primary outcome

JT

According to the paper, this randomized controlled trial had the following key details:

Population:

- 4613 participants aged 18+ years who smoked daily and were willing to make a serious quit attempt in the UK.
- Recruited online, mainly via the English Department of Health website.
- Stratified into low socioeconomic status (n=2142) and high socioeconomic status (n=2471) based on occupation.

Interventions:

- Interactive internet-based smoking cessation website called StopAdvisor.
- Control group received a 1-page static website with brief standard smoking cessation advice.

Main outcome:

- Primary outcome was biochemically verified 6-month sustained abstinence from smoking (self-report of smoking ≤5 cigarettes in the past 6 months plus saliva cotinine <15 ng/mL).

Results:

- No overall difference between groups in 6-month abstinence (StopAdvisor 10% vs control 10%, RR 1.06, 95% CI 0.89-1.27, p=0.49).
- In the low SES subgroup, 6-month abstinence was higher in StopAdvisor vs control (8% vs 6%, RR 1.36, 95% CI 1.00-1.86, p=0.0499).
- No difference between groups in the high SES subgroup (StopAdvisor 12% vs control 13%, RR 0.95, 95% CI 0.77-1.17, p=0.61).

UCL

Internet-based intervention for smoking cessation (StopAdvisor) in people with low and high socioeconomic status: a randomised controlled trial

Jamie Brown, Susan Michie, Adam WA Geraghty, Lucy Yardley, Benjamin Gardner, Lion Shahab, John A Stapleton, Robert West

Summary

Background Internet-based interventions for smoking cessation could help millions of people stop smoking at very low unit costs; however, long-term biochemically verified evidence is scarce and such interventions might be less effective for smokers with low socioeconomic status than for those with high status because of lower online literacy to engage with websites. We aimed to assess a new interactive internet-based intervention (StopAdvisor) for smoking cessation that was designed with particular attention directed to people with low socioeconomic status.

Methods We did this online randomised controlled trial between Dec 6, 2011, and Oct 11, 2013, in the UK. Participants aged 18 years and older who smoked every day were randomly assigned (1:1) to receive treatment with StopAdvisor or an information-only website. Randomisation was automated with an unseen random number function embedded in the website to establish which treatment was revealed after the online baseline assessment. Recruitment continued until the required sample size had been achieved from both high and low socioeconomic status subpopulations. Participants, and researchers who obtained data and did laboratory analyses, were masked to treatment allocation. The primary outcome was 6 month sustained, biochemically verified abstinence. The main secondary outcome was 6 month, 7 day biochemically verified point prevalence. Analysis was by intention to treat. Homogeneity of intervention effect across the socioeconomic subsamples was first assessed to establish whether overall or separate subsample analyses were appropriate. The study is registered as an International Standard Randomised Controlled Trial, number ISRCTN99820519.

Findings We randomly assigned 4613 participants to the StopAdvisor group (n=2321) or the control group (n=2292); 2142 participants were of low socioeconomic status and 2471 participants were of high status. The overall rate of smoking cessation was similar between participants in the StopAdvisor and control groups for the primary (237 [10%] *vs* 220 [10%] participants; relative risk [RR] 1·06, 95% CI 0·89–1·27; p=0·49) and the secondary (358 [15%] *vs* 332 [15%] participants; 1·06, 0·93–1·22; p=0·37) outcomes; however, the intervention effect differed across socioeconomic status subsamples (1·44, 0·99–2·09; p=0·0562 and 1·37, 1·02–1·84; p=0·0360, respectively). StopAdvisor helped participants with low socioeconomic status stop smoking compared with the information-only website (primary outcome: 90 [8%] of 1088 *vs* 64 [6%] of 1054 participants; RR 1·36, 95% CI 1·00–1·86; p=0·0499; secondary outcome: 136 [13%] *vs* 100 [10%] participants; 1·32, 1·03–1·68, p=0·0267), but did not improve cessation rates in those with high socioeconomic status (147 [12%] of 1233 *vs* 156 [13%] of 1238 participants; 0·95, 0·77–1·17; p=0·61 and 222 [18%] *vs* 232 [19%] participants; 0·96, 0·81–1·13, p=0·64, respectively).

Lancet Respir Med 2014

Published Online
September 25, 2014
http://dx.doi.org/10.1016/S2213-2600(14)70195-X

See Online/Comment
http://dx.doi.org/10.1016/S2213-2600(14)70214-0

Cancer Research UK Health Behaviour Research Centre, Department of Epidemiology and Public Health (J Brown PhD, B Gardner DPhil, L Shahab PhD, Prof R West PhD) **and Department of Clinical, Educational, and Health Psychology** (Prof S Michie DPhil), **University College London, London, UK; National Centre for Smoking Cessation and Training, London, UK** (Prof S Michie, Prof R West); **Primary Care and Population Sciences** (A W A Geraghty PhD) **and School of Psychology** (Prof L Yardley PhD), **University of Southampton, Southampton, UK; Addictions Department, Institute of Psychiatry, Kings College London, London, UK** (J A Stapleton MSc)

Correspondence to:
Dr Jamie Brown, Health

The overall rate of smoking cessation was similar between participants in the StopAdvisor and control groups for both the primary (237 [10%] *vs* 220 [10%] participants; relative risk [RR] 1·06, 95% CI 0·89–1·27; p=0·49) and the secondary (358 [15%] *vs* 332 [15%] participants; 1·06, 0·93–1·22; p=0·37) outcomes. However, analysis of the interaction between intervention and socioeconomic status showed clear evidence of non-ignorable heterogeneity of intervention effect by both primary (RR 1·44, 95% CI 0·99–2·09; p=0·0562) and secondary (1·37, 1·02–1·84; p=0·0360) cessation measures. This finding was evident before and after adjustment for all other baseline characteristics (adjusted data not shown). Consequently, the analysis of outcome was done separately within each of the two socioeconomic status subsamples.

In the subsample of participants with low socioeconomic status, a benefit of StopAdvisor was evident for both primary and secondary measures compared with the information-only website, whereas in those with high socioeconomic status, no evidence of a difference was shown (table 2). Adjustment for all baseline characteristics had a negligible effect on these comparisons (table 2). In a post-hoc sensitivity analysis, we re-examined the effect of StopAdvisor on biochemically verified smoking cessation in participants with low socioeconomic status, after exclusion of those in full-time education from the classification (n=237). The benefit of StopAdvisor compared with the information-only website remained in both unadjusted (primary outcome 87 [9%] of 967 *vs* 60 [6%] of 938 participants; RR 1·41, 95% CI 1·03–1·93; p=0·0346; secondary

who did not have post-16 education. In this smaller subsample (n=1687), the results were consistent with the primary analyses for StopAdvisor versus information only, but were non-significant in both the unadjusted (primary outcome 818 participants; R secondary outcome 1·21, 0·93–1·58; p= 1·27, 0·92–1·75; 0·92–1·58; p=0·17) analysis, we re-exa self-reported rather of smoking cessati basis of similar verification criteria new analyses showe reported in table 2 (socioeconomic sta compared with inf not significant (141 [13%] of 1088 unadjusted RR 1·20 RR 1·23, 0·97–1 prevalence 227 [21% RR 1·13, 0·95–1·34 p=0·07). That sta in participants with failed to reach sign decreased power between the two gro rates.

	StopAdvisor	Control	Relative risk (95% CI)	Odds ratio (95% CI)*	Percentage-point difference (95% CI)	p value†
Primary outcome (abstinence for 6 months)						
High SES	147/1233 (12%)	156/1238 (13%)	0·95 (0·77 to 1·17)	0·94 (0·74 to 1·19)	−0·68 (−3·27 to 1·91)	0·61
Adjusted	..	..	0·97 (0·78 to 1·19)‡	0·95 (0·75 to 1·22)‡	..	0·75
Low SES	90/1088 (8%)	64/1054 (6%)	1·36 (1·00 to 1·86)	1·39 (1·00 to 1·94)	2·20 (0·02 to 4·38)	0·0499
Adjusted	..	..	1·43 (1·05 to 1·96)‡	1·46 (1·04 to 2·05)‡	..	0·0238
Secondary outcome (point prevalence at 6 months)						

State, in tabular form, the outcomes of the sutdy

JT

Here is a table summarizing the key outcomes of the study:

	StopAdvisor	Control	Relative Risk (95% CI)	p-value
Overall				
6-month abstinence	10%	10%	1.06 (0.89-1.27)	0.49
Low SES				
6-month abstinence	8%	6%	1.36 (1.00-1.86)	0.0499
High SES				
6-month abstinence	12%	13%	0.95 (0.77-1.17)	0.61

Copy

What were the risk of bias as assessed by the Cochrane RoB tool?

JT

The study authors did not report a risk of bias assessment using the Cochrane RoB tool. However, here are some observations about potential risks of bias based on the details provided in the paper:

- Random sequence generation: Participants were randomized using an automated unseen random number function, so this should have been adequate.
- Allocation concealment: The randomization was automated with no experimenter involvement, so allocation was likely adequately concealed.
- Blinding of participants/personnel: Not reported, but blinding of participants may not have been feasible given the difference in intervention. Personnel were likely not blinded.
- Blinding of outcome assessment: Outcome assessors were reported to be blinded to treatment allocation.
- Incomplete outcome data: Follow-up rates were 77% overall and similar between groups, with missing data treated as smoking, so the risk appears low.
- Selective reporting: The outcomes specified in the protocol appear to be reported, suggesting low risk of bias.
- Other biases: The study appears to have been adequately powered and analyzed as pre-specified, with low risks of other biases.

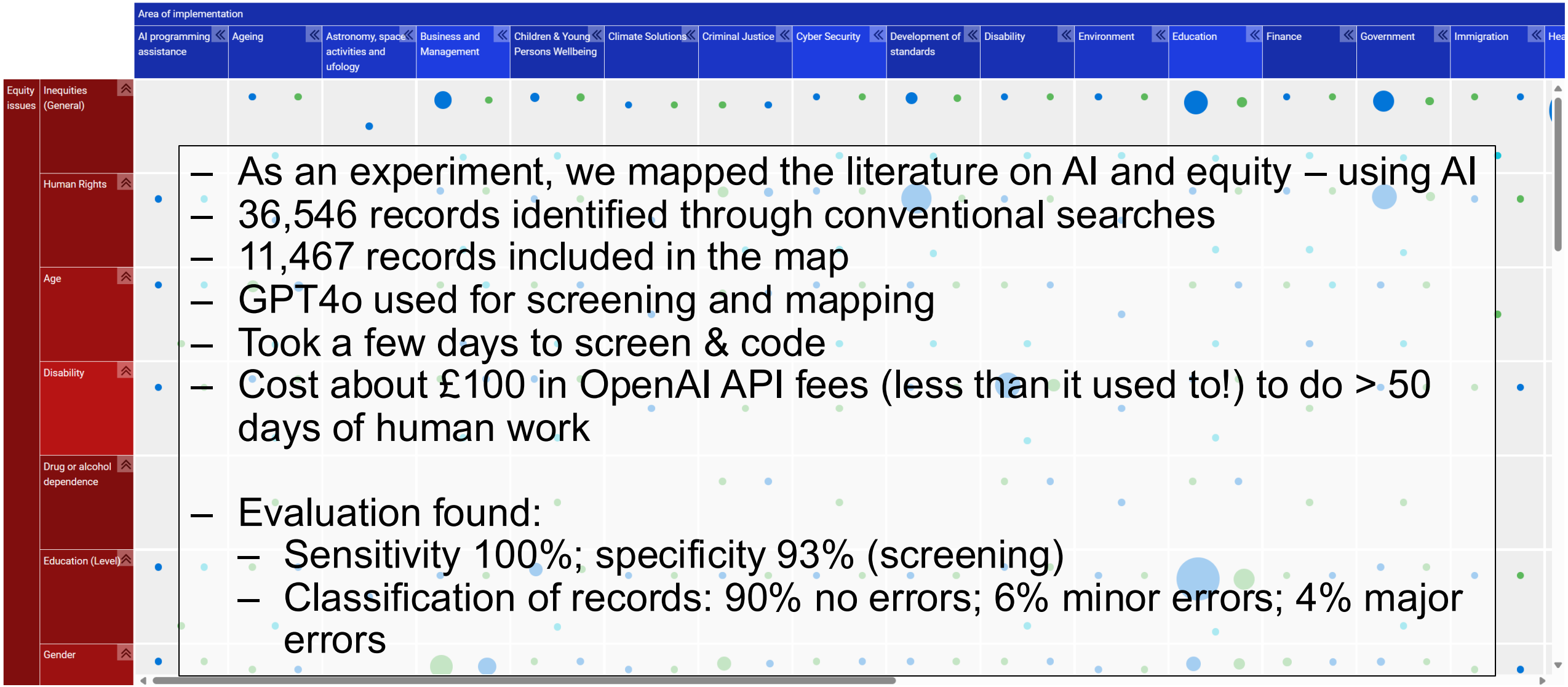
So in summary, the lack of blinding of participants and personnel, as well as the lack of formal risk of bias assessment by the authors, suggest some uncertainties but overall the study seems reasonably well conducted.

Copy

EPPI Centre
Evidence for Policy & Practice

AI and equity: what are the benefits and harms associated with this new generation of decision-making tools? (filtered map)

Campbell Collaboration



- As an experiment, we mapped the literature on AI and equity – using AI
- 36,546 records identified through conventional searches
- 11,467 records included in the map
- GPT4o used for screening and mapping
- Took a few days to screen & code
- Cost about £100 in OpenAI API fees (less than it used to!) to do > 50 days of human work

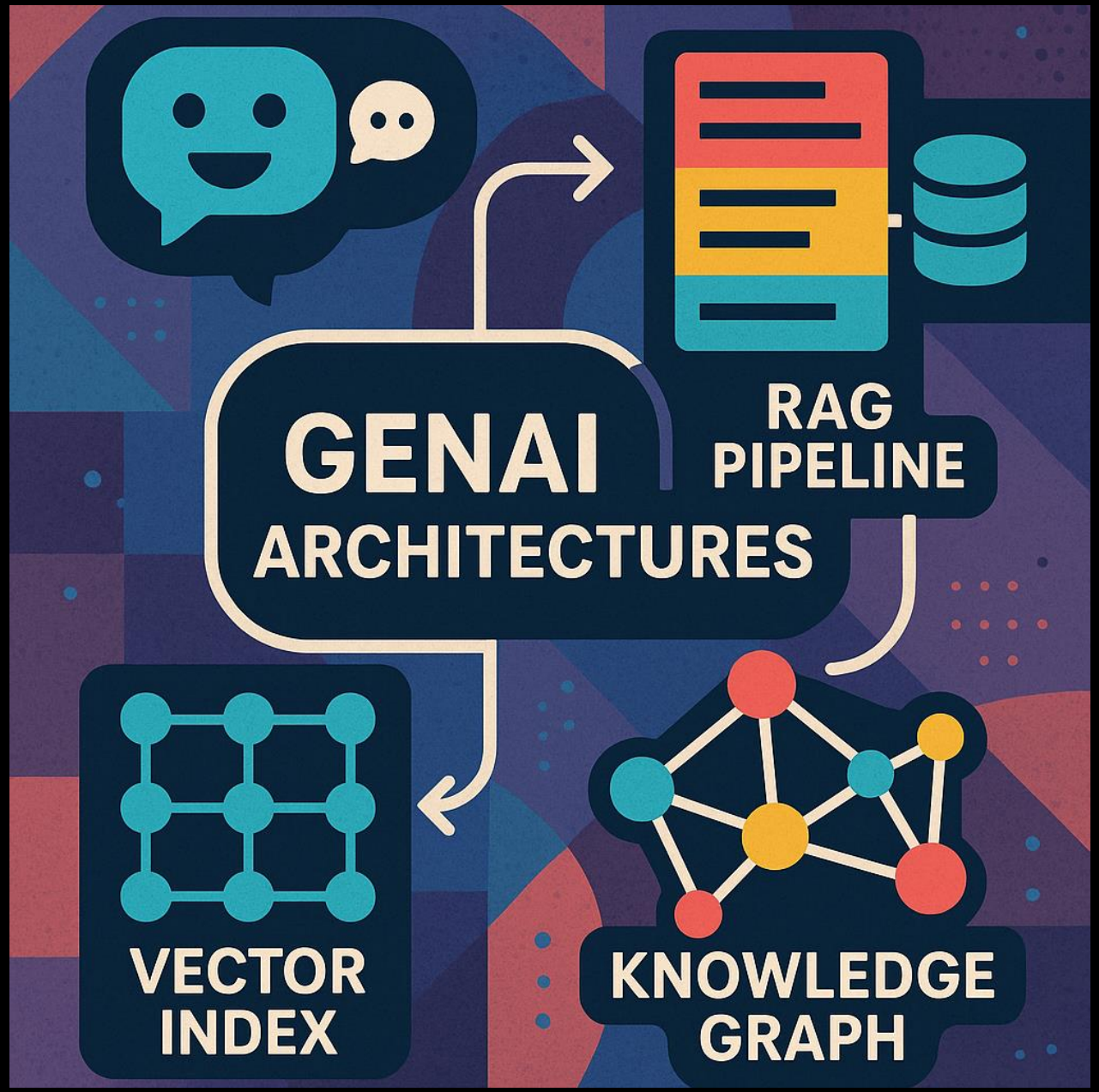
- Evaluation found:
  - Sensitivity 100%; specificity 93% (screening)
  - Classification of records: 90% no errors; 6% minor errors; 4% major errors

Sounds amazing, but…

- Concerns about bias
- Appropriation of content for training without permission
- Environmental impact
- Lack of transparency
- Currently a lack of an evidence base
- A great deal of hype (and money) means claims are currently running ahead of capabilities

In summary

Rule-based	Unsupervised	Supervised	Generative
• Not fashionable • Potentially powerful • Very demanding in time • Rules can be fragile	• Very little time effort required to create rules or training data • No control over classifications	• Makes use of data created for other purposes • Does not break as easily as rule-based approaches • Can predict specific classification terms (unlike unsupervised) • Can require large quantities of training data which may be unavailable / uneconomic to create	• Considered current 'state of the art' • Huge research focus • Sometimes beats simpler models (though sometimes only marginally) • Concerns about bias and other negative outcomes



Understanding their architectures is critical for understanding the strengths and limitations of these new tools

- ChatBot
- ‘Semantic’ search
- Retrieval Augmented Generation (RAG)

ChatGPT

Should I used generative AI tools to help with evidence synthesis?

Yes, you *can* and *should consider* using generative AI tools to support evidence synthesis — but **with clear boundaries** and **critical oversight**. Here's a breakdown of how they can help, and what you should be careful about:

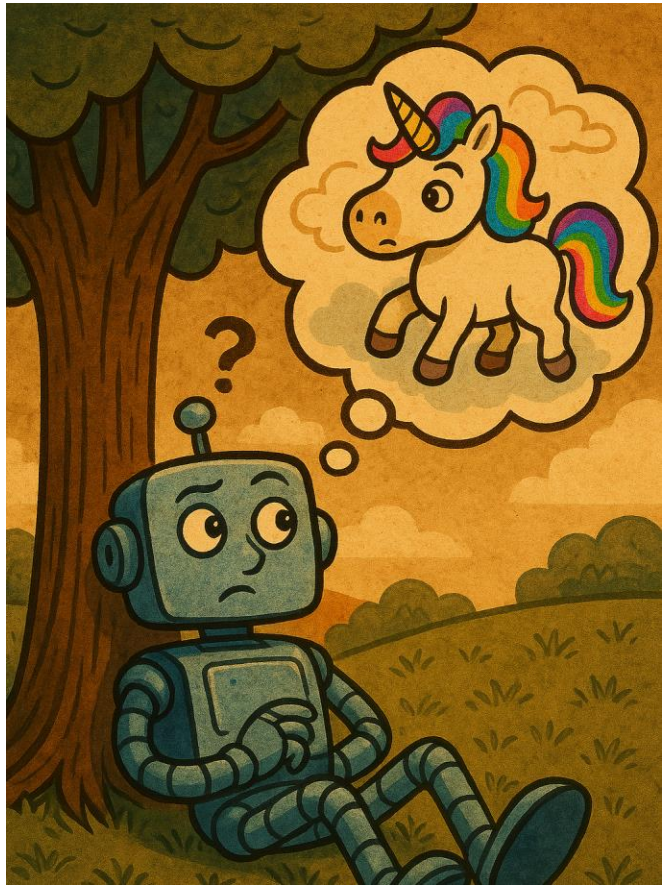
✅ Ways Generative AI Can Help with Evidence Synthesis

1. **Search and Screening Support**
   - Summarize abstracts quickly.
   - Assist in relevance screening by generating concise overviews.
   - Classify studies by population, intervention, outcome, etc.
2. **Data Extraction Aid**
   - Pull structured information (e.g., study design, sample size) from articles.
   - Convert narrative findings into standardized formats.
   - Help identify outcome measures and comparator groups.
3. **Thematic Analysis and Coding**
   - Suggest initial codes for qualitative synthesis (e.g., for a meta-ethnography).
   - Identify emerging themes across studies.

Ask anything

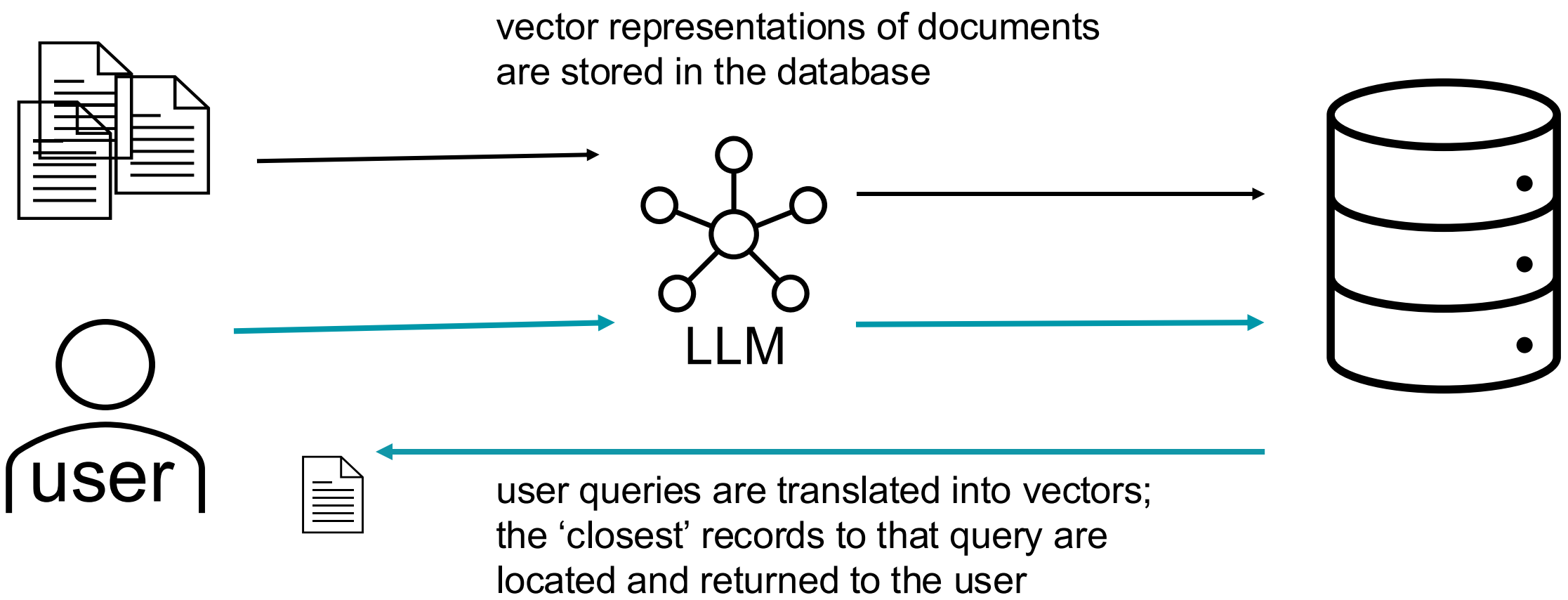
Search Reason Deep research Create image

Strengths and limitations: chatbot



- Can be asked questions in standard prose
- Can provide accurate answers quickly
- **But**
- Frequency biased
- ‘Hallucinate’
- Sounds confident, but is often wrong

‘Semantic search’

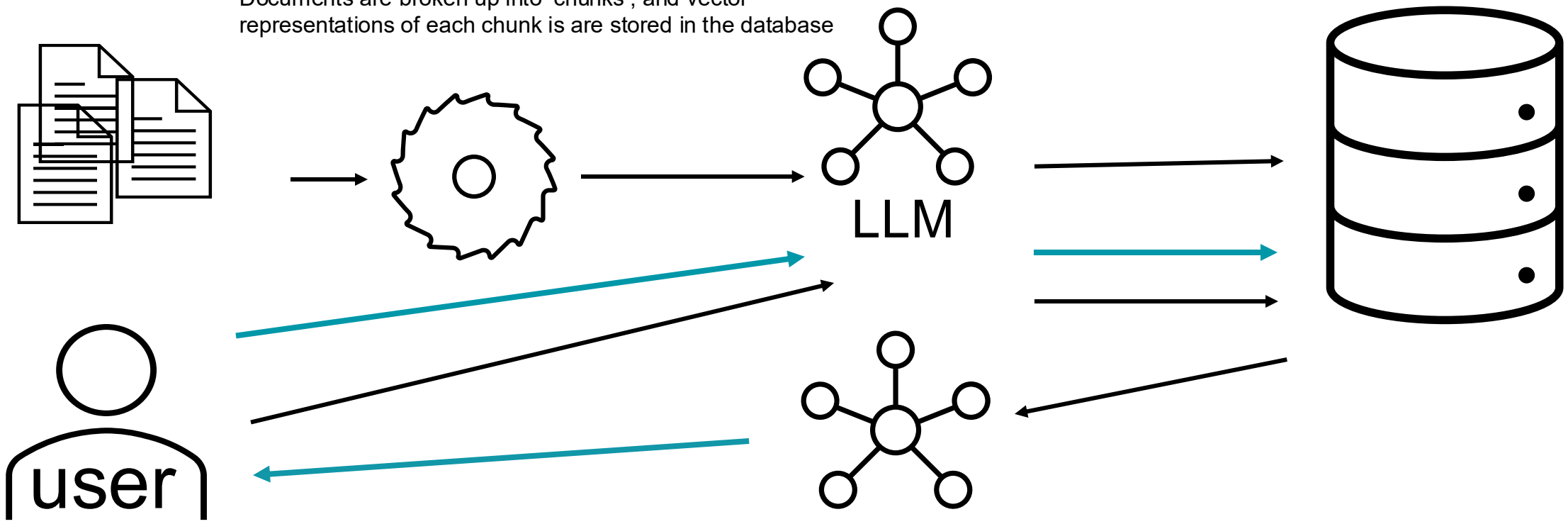


Strengths and limitations: vector indexes

- Can provide more semantically powerful searches
- Less ‘fragile’ than a Boolean search (and not necessary to know all relevant terms in advance)
- BUT
- Dependent on the right documents being available for indexing
- Dependent on the query being sufficiently ‘similar’ to the documents being retrieved
- Little in the way of an evidence base to support their use in evidence synthesis

Retrieval Augmented Generation

Documents are broken up into ‘chunks’, and vector representations of each chunk is are stored in the database



User queries are translated into vectors; the ‘closest’ chunks of documents to that query are located; the LLM then generates an answer to the user’s query, based on the chunks of text returned

Strengths and limitations: retrieval augmented generation

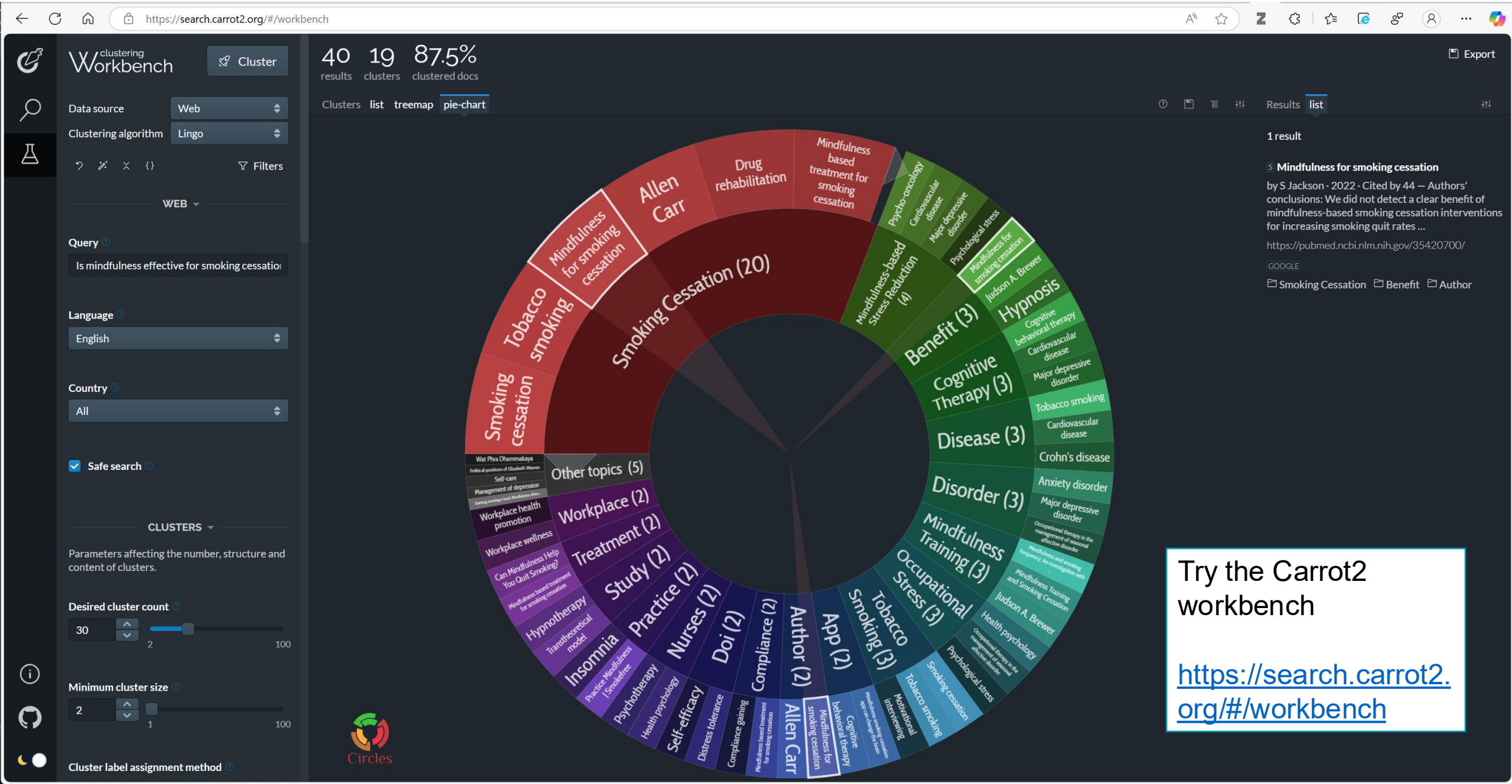
- Can provide a powerful interactive experience where users can ‘chat’ to their documents using standard prose
- BUT
- Has many of the limitations of BOTH chatbots and vector indexes:
  - Can hallucinate
  - Requires good translation from query to retrieval AND question to the LLM
  - What if all the relevant documents are not retrieved?
  - What if irrelevant documents *are* retrieved?

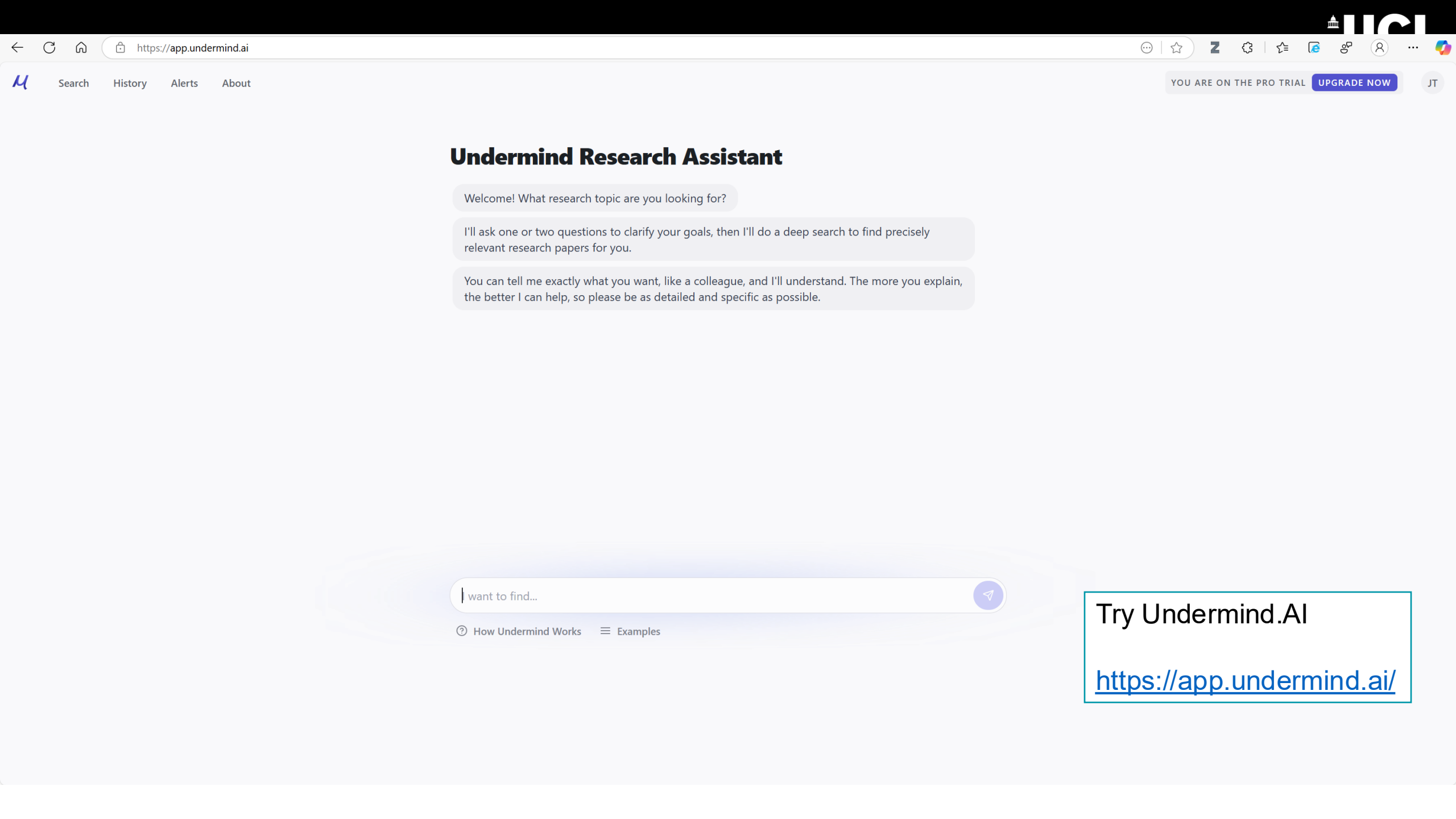
Important questions to ask of LLM-based evidence synthesis tools

- For chatbots:
  - Can I verify its accuracy?
  - (Does it matter if not?)
- For search tools:
  - Are the records I need indexed?
  - How can I check that its retrieved everything it should?
- For ‘RAG’-based approaches:
  - Are the right documents indexed?
  - Are the right documents retrieved?
  - Are incorrect documents avoided?
  - If present, does the summariser check that the research is reliable / that combining them is a valid thing to do?

Now it’s your turn!

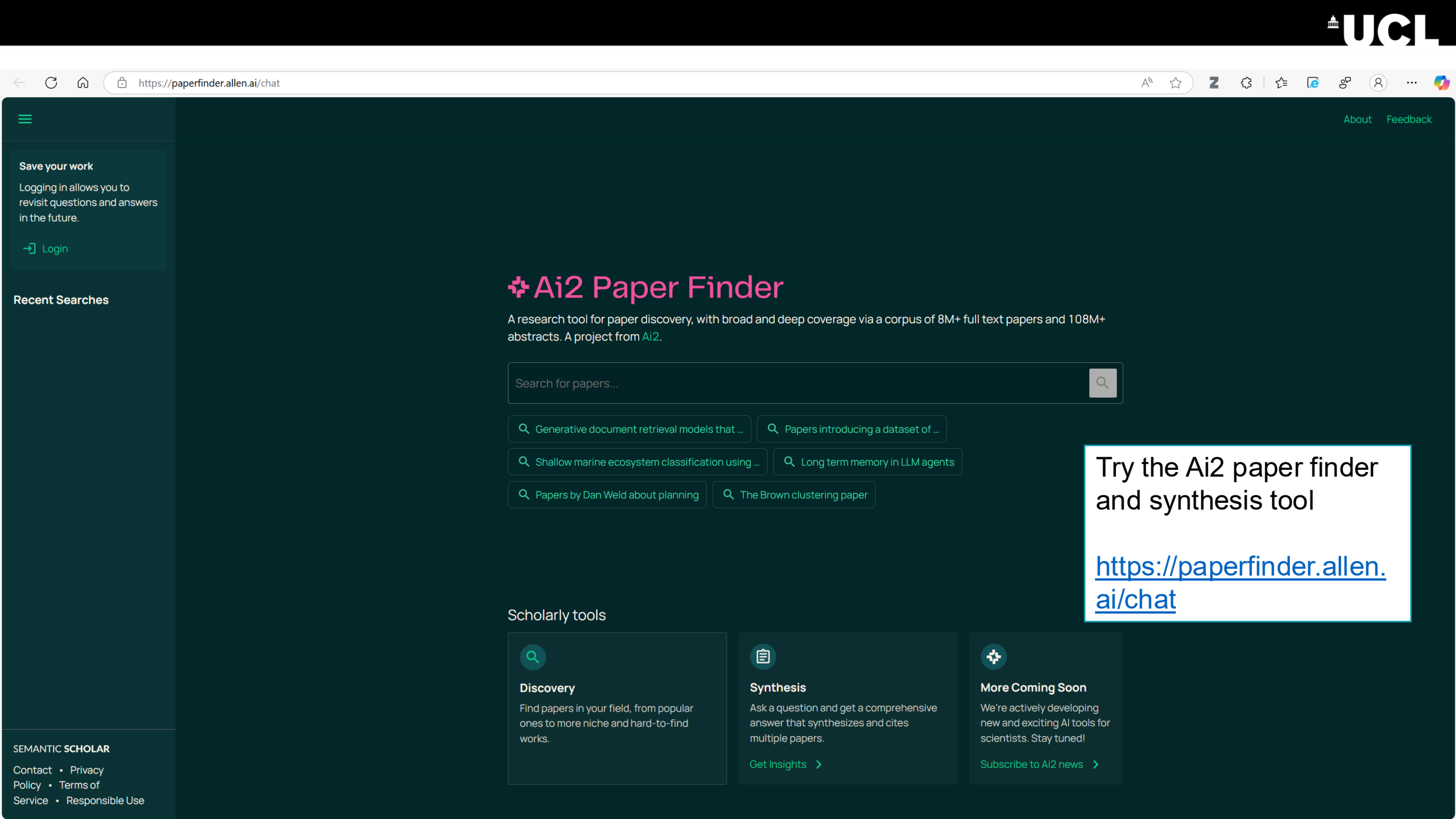
Try out some tools


https://search.carrot2.org/#/workbench
clustering Workbench
Cluster
Data source
Web
Clustering algorithm
Lingo
Filters
WEB
Query
Is mindfulness effective for smoking cessatio
Language
English
Country
All
Safe search
CLUSTERS
Parameters affecting the number, structure and content of clusters.
Desired cluster count
30
2
100
Minimum cluster size
2
1
100
Cluster label assignment method
40 results
19 clusters
87.5% clustered docs
Export
Clusters list treemap pie-chart
Results list
1 result
5 Mindfulness for smoking cessation
by S Jackson · 2022 · Cited by 44 — Authors' conclusions: We did not detect a clear benefit of mindfulness-based smoking cessation interventions for increasing smoking quit rates ...
https://pubmed.ncbi.nlm.nih.gov/35420700/
GOOGLE
Smoking Cessation
Benefit
Author
Smoking Cessation (20)
Tobacco smoking
Smoking cessation
Mindfulness for smoking cessation
Allen Carr
Drug rehabilitation
Mindfulness based treatment for smoking cessation
Other topics (5)
Workplace (2)
Treatment (2)
Study (2)
Practice (2)
Nurses (2)
Doi (2)
Compliance (2)
Author (2)
App (2)
Tobacco Smoking (3)
Occupational Stress (3)
Mindfulness Training (3)
Disorder (3)
Disease (3)
Cognitive Therapy (3)
Benefit (3)
Mindfulness-based Stress Reduction (4)
Wat Phra Dhammakaya
Political positions of Elizabeth Warren
Self-care
Management of depression
Workplace health promotion
Workplace wellness
Can Mindfulness Help You Quit Smoking?
Mindfulness based treatment for smoking cessation
Hypnotherapy
Transtheoretical model
Insomnia
Practice Mindfulness | Smokefree
Psychotherapy
Health psychology
Self-efficacy
Distress tolerance
Compliance gaining
Mindfulness based treatment for smoking cessation
Allen Carr
Mindfulness for smoking cessation
Cognitive behavioral therapy
Mindfulness smoking cessation app | can change the brain
Motivational interviewing
Tobacco smoking
Smoking cessation
Psychological stress
Occupational therapy in the management of seasonal affective disorder
Health psychology
Judson A. Brewer
Mindfulness Training and Smoking Cessation
Mindfulness and smoking frequency: An investigation with ...
Occupational therapy in the management of seasonal affective disorder
Major depressive disorder
Anxiety disorder
Crohn's disease
Cardiovascular disease
Tobacco smoking
Major depressive disorder
Cardiovascular disease
Cognitive behavioral therapy
Hypnosis
Judson A. Brewer
Mindfulness for smoking cessation
Psychological stress
Major depressive disorder
Cardiovascular disease
Psycho-oncology
Circles
Try the Carrot2 workbench
https://search.carrot2.org/#/workbench



Try Undermind.AI

https://app.undermind.ai/


https://paperfinder.allen.ai/chat
About
Feedback
Save your work
Logging in allows you to revisit questions and answers in the future.
Login
Recent Searches
Ai2 Paper Finder
A research tool for paper discovery, with broad and deep coverage via a corpus of 8M+ full text papers and 108M+ abstracts. A project from Ai2.
Search for papers...
Generative document retrieval models that ...
Papers introducing a dataset of ...
Shallow marine ecosystem classification using ...
Long term memory in LLM agents
Papers by Dan Weld about planning
The Brown clustering paper
Scholarly tools
Discovery
Find papers in your field, from popular ones to more niche and hard-to-find works.
Synthesis
Ask a question and get a comprehensive answer that synthesizes and cites multiple papers.
Get Insights
More Coming Soon
We're actively developing new and exciting AI tools for scientists. Stay tuned!
Subscribe to Ai2 news
SEMANTIC SCHOLAR
Contact • Privacy Policy • Terms of Service • Responsible Use
Try the Ai2 paper finder and synthesis tool
https://paperfinder.allen.ai/chat

https://elicit.com/review/5a76519c-ad0e-4ae6-b176-d900eb222b2b

Elicit | Recent | Library | Upgrade | Help | James Thomas

Optimal Papers for Meta-Regression Covaria... | Research report View only | Share

MAY 13, 2025

How many papers do I need per covariate in a meta regression?

Meta-regression paper requirements range from 8-40 studies, with higher numbers needed for complex models or high heterogeneity.

ABSTRACT

This synthesis of simulation, theoretical, and methodological studies reveals no single formula for the number of papers (or effect sizes) needed per covariate in meta-regression. Fang and Zhang (2020) report that at least 20 effect sizes are required for parameter estimation, while Hedges et al. (2010) suggest a range of 20–40 studies ensures robust variance estimation. Jenkins and Quintana-Ascencio (2020) illustrate that when heterogeneity is low, as few as 8 studies may suffice, whereas high variance may require up to 25 studies. Mathur and VanderWeele (2020) indicate that, depending on the metric, 10–20 studies can be adequate. In addition, Tipton (2015) and Higgins and Thompson (2004) note that degrees of freedom depend on both the number and type of covariates, implying that more complex models may demand higher overall sample sizes. Thus, while explicit guidance on papers per covariate is rarely provided, the available evidence shows that requirements range from 8 to 40 studies overall, with adjustments based on model complexity and heterogeneity.

METHODS

We analyzed 10 papers from an initial pool of 50, using 5 screening criteria. Each paper was reviewed for 5 key aspects that mattered most to the research question. More on methods

RESULTS

Characteristics of Included Studies

Study	Study Design	Primary Focus	Statistical Approach	Sample Size Recommendations	Full text retrieve

Report

Status

- Gather papers — 50 papers found — Details
- Screen papers — 10 papers included — Details
- Extract data — 50 data points extracted — Details
- Generate report — Save PDF

Chat

Ask anything about the report or its underlying data

Try Elicit

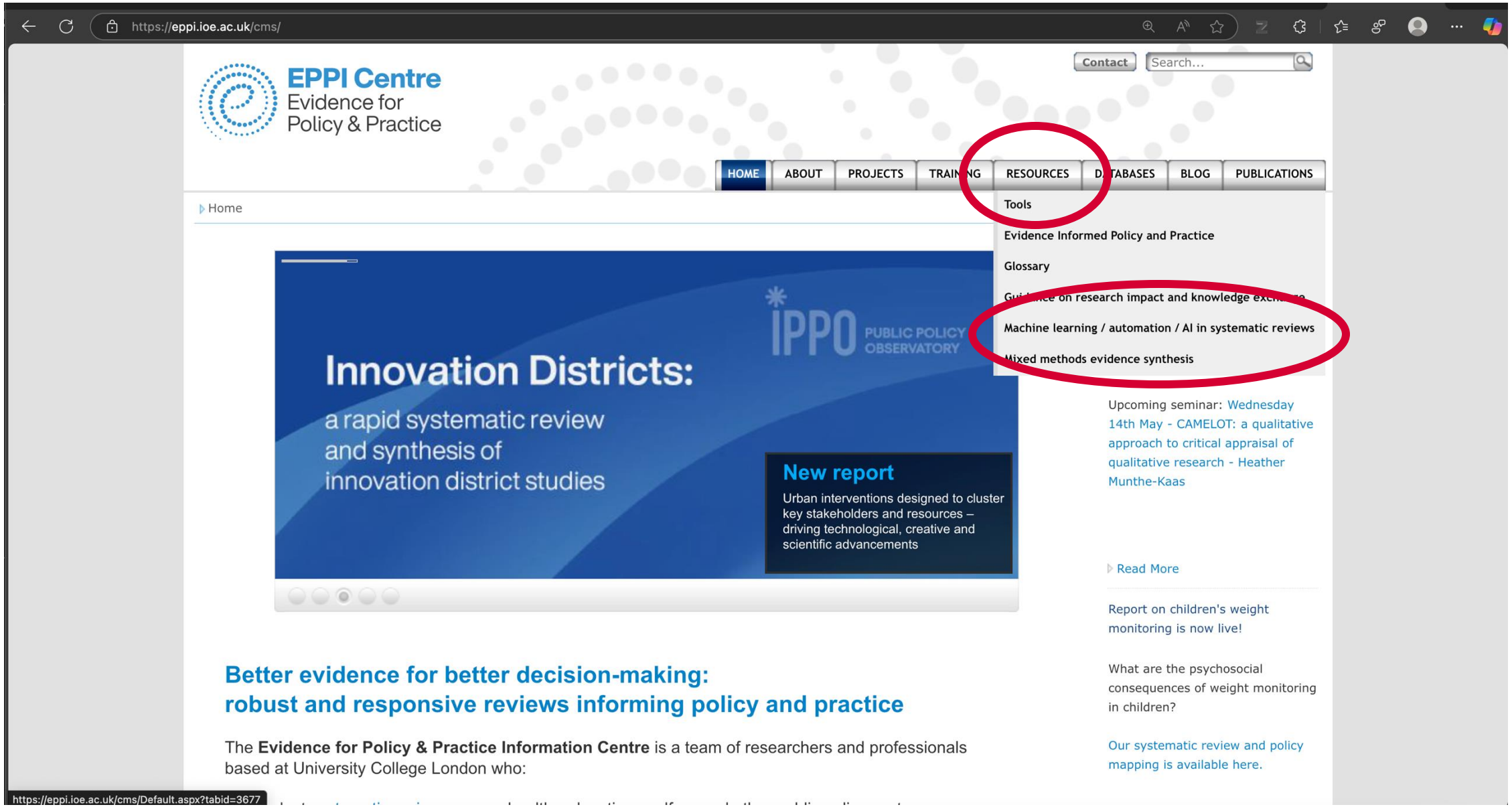
https://elicit.com/

Try out one or more tools

- Unsupervised machine learning
  - Carrot2 Workbench
- Generative machine learning
  - Ai2 paper finder and synthesis tools
  - Undermind.AI
  - Elicit
  - RobotReviewer
- One of the tools for search strategy development

- Ask yourself
  - Is it clear how the tool works?
  - Can I tell whether it can find all the material that is relevant to my query?
  - How much do I trust its output?

https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677



https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3677

Now it’s your turn!

Discussion

Conclusions

– More research is needed?

We were asked to write some guidance…

- … about which tool to use, and when
- But found we couldn't!
- The evidence base on which to base our advice was very limited
- AI tools were being developed that were not engineered to be fit-for-purpose

Vision: RAISE guidance for the responsible use of AI in evidence synthesis

- A draft of the guidance and recommendations is now online for consultation
- Our vision is for it to be a 'living' set of guidelines, that is updated through community input and helps to define roles & responsibilities within the ecosystem
- Should the ecosystem develop in this well-organized way, we hope to see the development of AI tools that adhere to the principles of research integrity, and so enable evidence accessibility in equitable and rigorous ways

Roles-based ecosystem

- We need to support the wider adoption of AI to overcome the increasing burden of doing timely and cost-effective evidence synthesis
- We need cross-field standards to support the development of appropriate and responsible AI
- We anticipate an ecosystem made up of individuals, collaborations, and organisations which each have a role to play in developing and using AI in a responsible way
- (one person / organisation may play multiple roles)

How you can get involved (1)

- The link : https://osf.io/fwaud/
- Timetable for development
  - A new version will be published in the next few weeks
- Three documents:
  - Roles-based recommendations for practice
  - Guidance on building and evaluating AI tools
  - Guidance on selecting and using AI tools
- Do take a look and let us know what you think!

How you can get involved (2): ‘Studies Within A Review’ (SWARs)

Received: 18 September 2022 | Accepted: 29 November 2022
DOI: 10.1111/jebm.12505

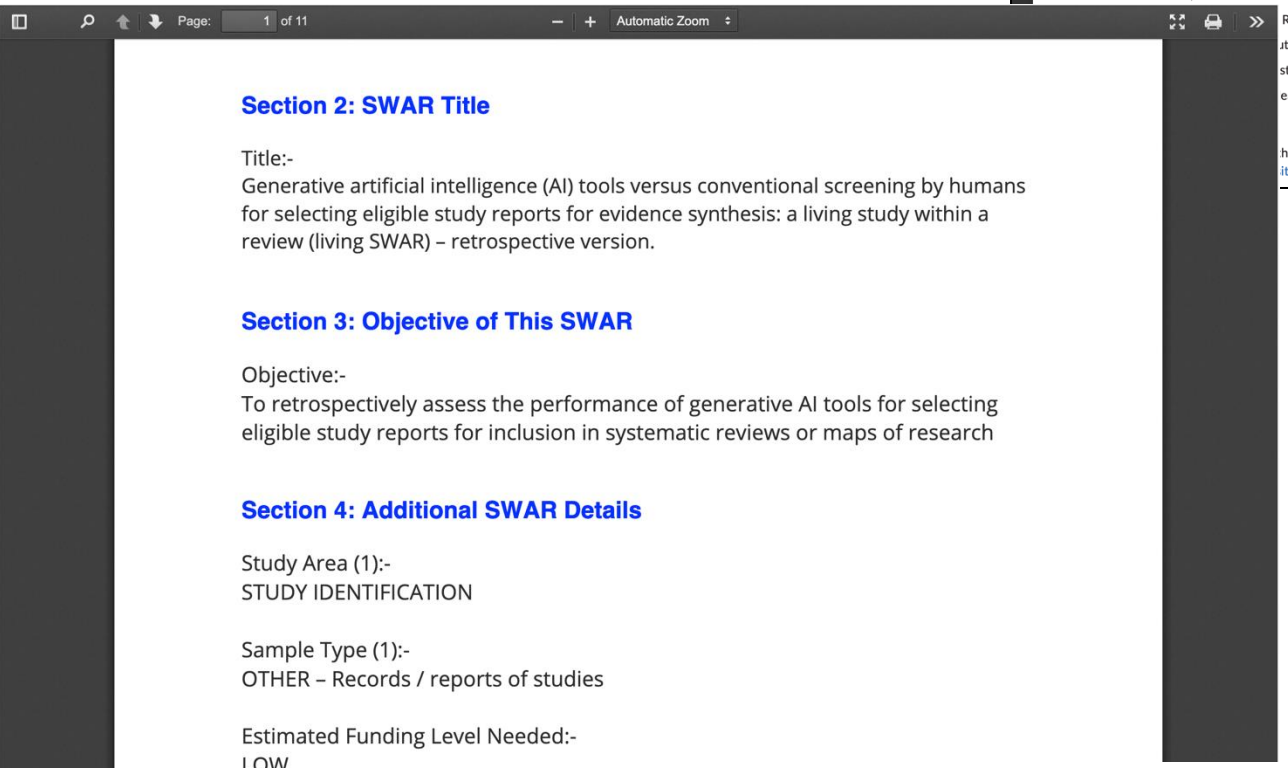
PERSPECTIVE

WILEY

Study within a review (SWAR)

Declan Devane[1,2,3] | Nikita N. Burke[1,2] | Shaun Treweek[4] | Mike Clarke[5] | James Thomas[6] | Andrew Booth[7] | Andrea C. Tricco[8,9,10] | K. M. Saif-Ur-Rahman[1,2]

[1]Evidence Synthesis Ireland and Cochrane Ireland, University of Galway, Galway, Ireland
[2]School of Nursing and Midwifery, University of Galway, Galway, Ireland
[3]HRB-Trials Methodology Research Network, University of Galway, Galway, Ireland
[4]Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[5]Northern Ireland Methodology Hub, Queen's University Belfast, Belfast, UK
[6]EPPI-Centre, UCL Social Research Institute, University College London, London, UK



Page: 1 of 11 Automatic Zoom

Section 2: SWAR Title

Title:-
Generative artificial intelligence (AI) tools versus conventional screening by humans for selecting eligible study reports for evidence synthesis: a living study within a review (living SWAR) – retrospective version.

Section 3: Objective of This SWAR

Objective:-
To retrospectively assess the performance of generative AI tools for selecting eligible study reports for inclusion in systematic reviews or maps of research

Section 4: Additional SWAR Details

Study Area (1):-
STUDY IDENTIFICATION

Sample Type (1):-
OTHER – Records / reports of studies

Estimated Funding Level Needed:-
LOW

- More consistency in methods, tasks and questions
- Enabling cumulation across studies (which may be small-N)
- Invitation to join a ‘living’ SWAR evaluating the use of LLMs for title & abstract / full text screening
- https://osf.io/g7mkb/

Devane D, Burke NN, Treweek S, Clarke M, Thomas J, Booth A, Tricco AC, Saif-Ur-Rahman KM (2022) Study within a review (SWAR). *J Evid Based Med*; 15: 328-332 https://doi.org/10.1111/jebm.12505

ESIC Stage 2:
What do we need?

Open consultation
12-19 March
2025

Evidence Synthesis
Infrastructure Collaborative

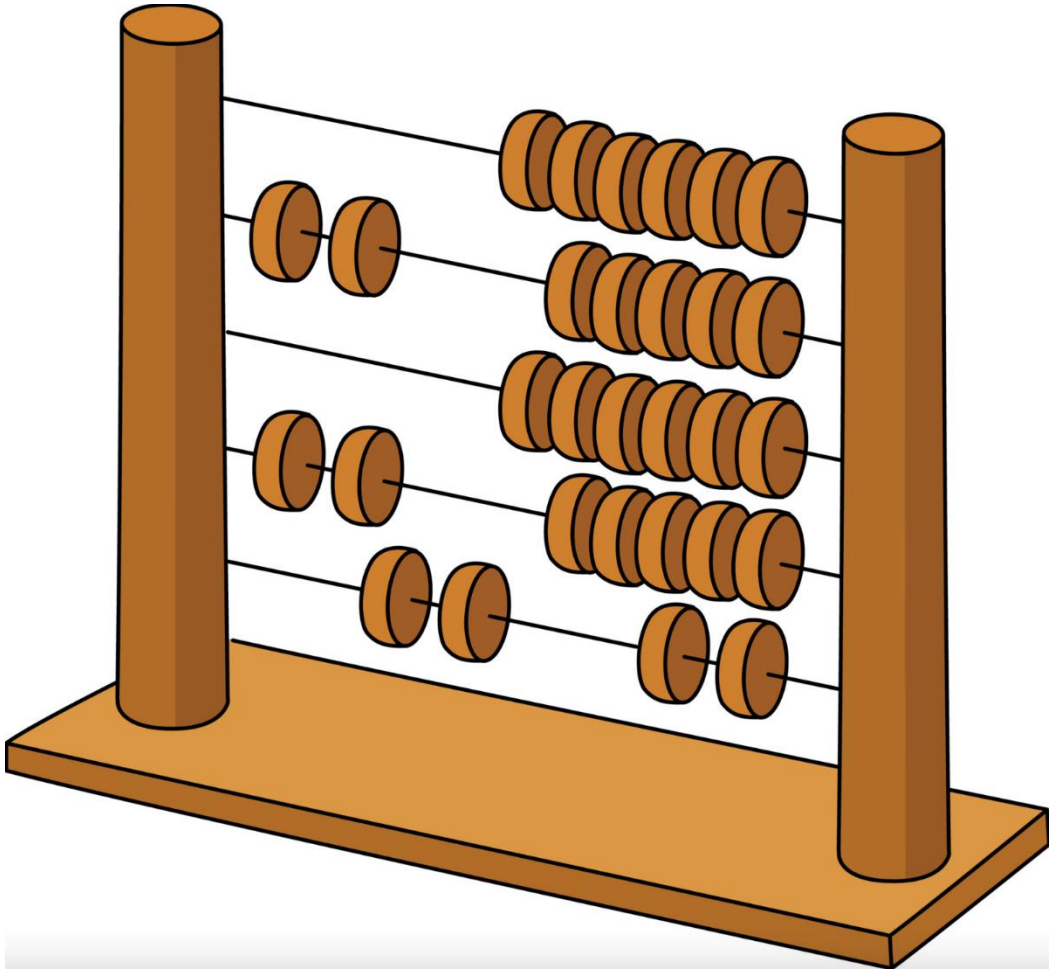
SUPPORTED BY
wellcome

How you can get involved (3)

https://evidencesynthesis.atlassian.net/wiki/spaces/ESE/overview

Summing up

- There are some great tools that may soon be ready for use
- But promise of GenAI will remain a promise until we have a good evidence base
- We need lots of rigorous evaluation before we can see the promise realized
- We need to increase our ‘AI literacy’ across the field to understand when and how to use (and not use) this new generation of tools



Thank you

James Thomas

EPPI-Centre website: http://eppi.ioe.ac.uk
Email
james.thomas@ucl.ac.uk
Twitter
James_M_Thomas

EPPI-Centre
Social Science Research Unit
Institute of Education
University of London
18 Woburn Square
London WC1H 0NR

Tel +44 (0)20 7612 6397
Fax +44 (0)20 7612 6400
Email eppi@ioe.ac.uk
Web eppi.ioe.ac.uk/