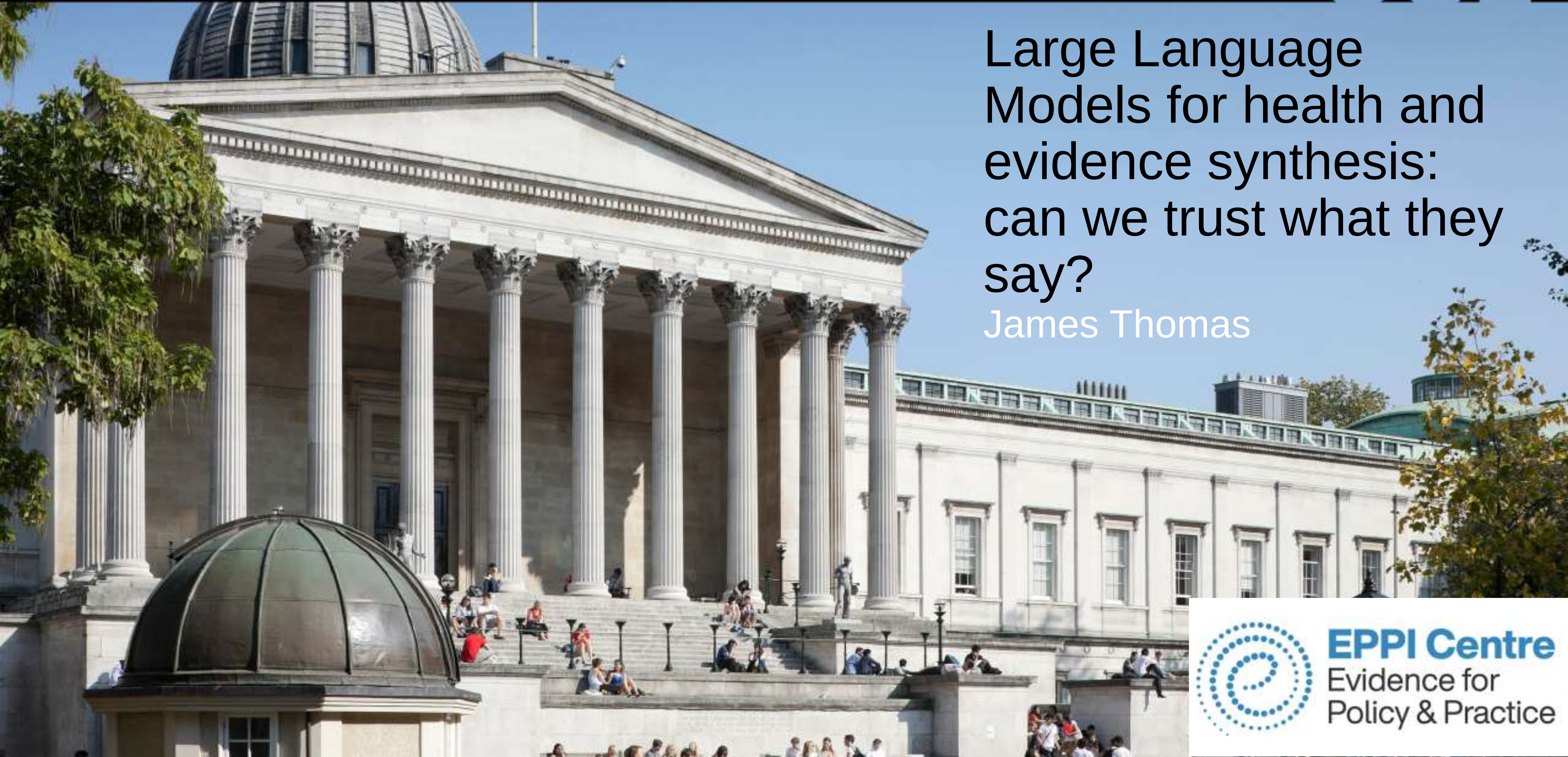
UCL Social Research Institute

Evidence Synthesis Ireland webinar: 23 November 2023

UCL



Large Language Models for health and evidence synthesis: can we trust what they say?

James Thomas

EPPI Centre
Evidence for Policy & Practice

About me

- Worked in the EPPI-Centre, UCL for a long time
- Systematic reviews – mostly for Department of Health & Social Care / PHE
- Addressing questions beyond effectiveness
- Long-standing area of work in making the review process more efficient using new technologies

Outline

- Automation in systematic reviews: the story so far
- Newer technologies using new enablers

- Generative Large Language models
  - How can they be used (in reviews)?
  - When can they be trusted?
  - Are they a gamechanger?

Automation in systematic reviews: what can be done?

Study identification:

- Citation screening
- Updating reviews
- RCT classifier

Mapping research activity

Data extraction

- Risk of Bias assessment
- Other study characteristics
- Extraction of statistical data

Synthesis and conclusions

More evidence of effectiveness

Automation in systematic reviews: what can be done?

Study identification:

- Citation screening
- Updating reviews
- RCT classifier

Mapping research activity

Data extraction

- Risk of Bias assessment
- Other study characteristics
- Extraction of statistical data

Synthesis and conclusions

More data available for research & development

Automation in systematic reviews: what can be done?

Study identification:

- Citation screening
- Updating reviews
- RCT classifier

Mapping research activity

Data extraction

- Risk of Bias assessment
- Other study characteristics
- Extraction of statistical data

Synthesis and conclusions

Easier for automation to solve these problems

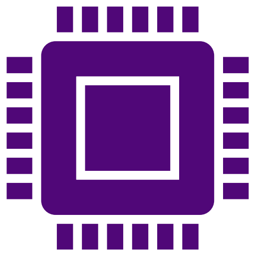
‘Traditional’ tools

- For example
  - Ranking and re-ranking records when screening titles & abstracts
  - Automatically ‘clustering’ records to enable us to explore datasets
  - Classification (e.g. RCT Classifier) where we can ‘teach’ the machine to perform certain tasks (usually IF we have lots of training data…)

- We feel we know where we are with these kinds of tools
- They are useful, not game-changing

Enablers of a new generation of digital evidence synthesis tools

Increased availability of open access research



Increased computing power (both memory + compute)

Advances in machine learning technology

New generation of ‘AI’ tools

- Promise to do more than achieve minor increases in efficiency
- At times, change the review process more fundamentally

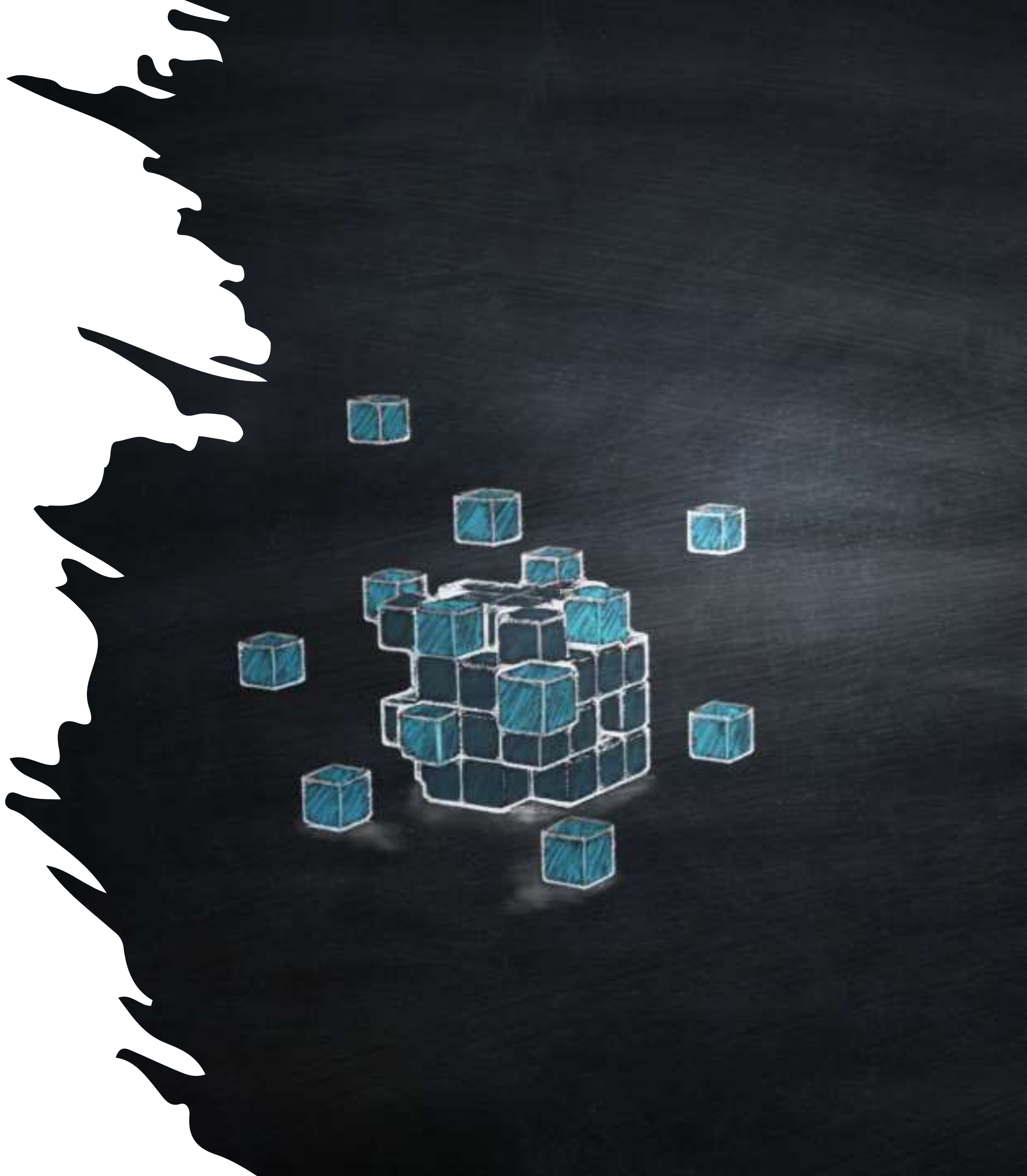
- More unsettling
  - Appear to ‘understand’ language
  - They can answer questions
  - They can synthesize knowledge

- But can we use them?

New approaches: more contextually ‘aware’ classification

- The theory:
  - When a human reads, they read in the light of their pre-existing knowledge
  - The previous examples do not do that
  - Is it possible to address this using machine learning?
- Word embeddings
  - E.g. Word2Vec
- Transformer models
  - E.g. BERT (Bidirectional Encoder Representations from Transformers)
  - LARGE ‘generative’ transformer models

- Key to bear in mind: these are all (sophisticated) statistical representations of words / phrases that tend to ‘go together’



Starting points

Decisions that affect people’s lives should be informed by reliable research

Individual research studies can be atypical; we need to draw on the sum of current knowledge

Therefore we use evidence synthesis

Evidence syntheses can be unreliable for two reasons:

They have been conducted badly

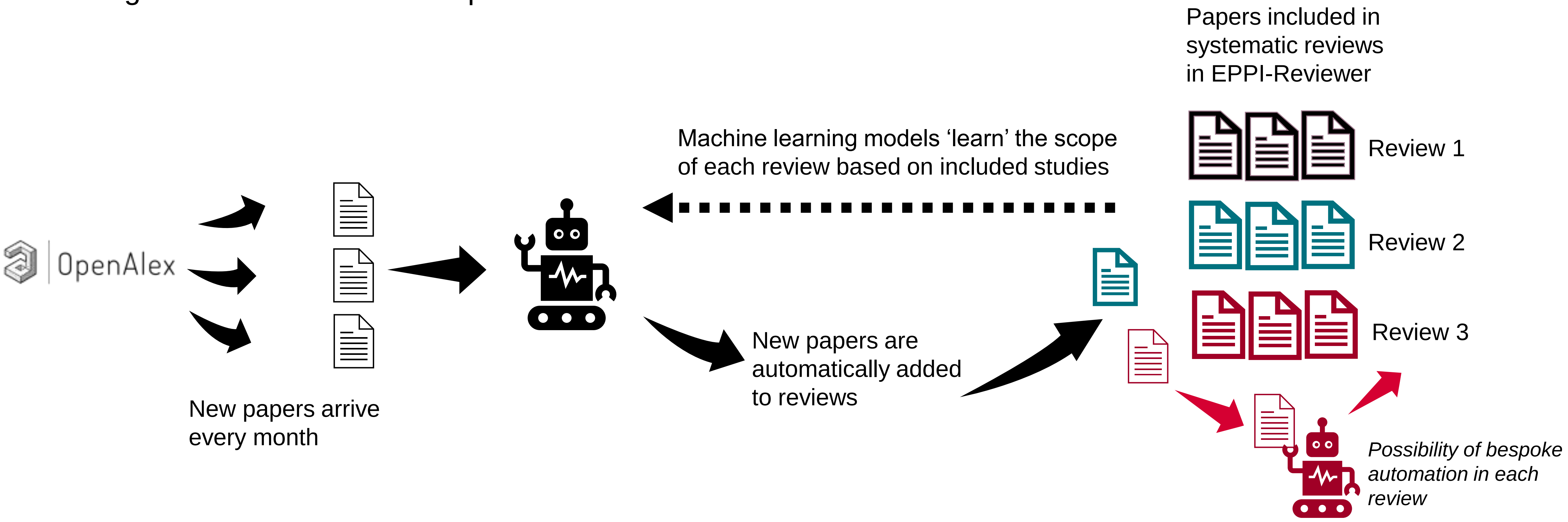
The research they contain is unreliable

Critical questions to ask when considering using a new tool for evidence synthesis

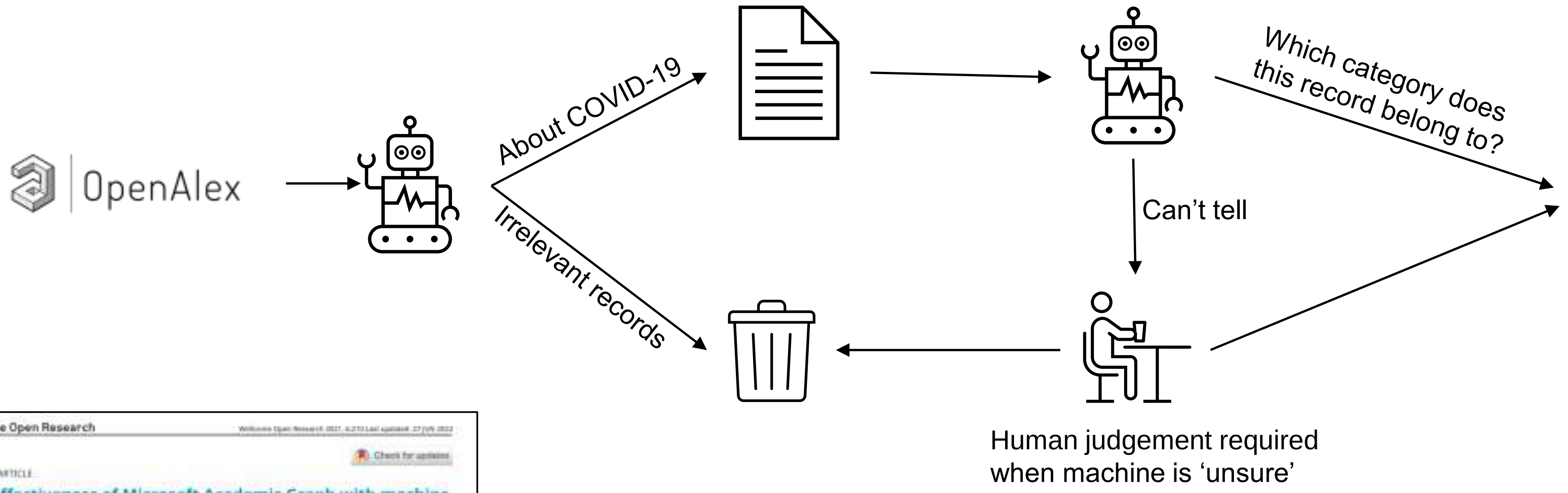
Does it enable me to draw on the sum of current knowledge?	Or does it present an incomplete or biased picture?
Does it enable me to distinguish between reliable and unreliable research?	Or does it treat all research as equally reliable?

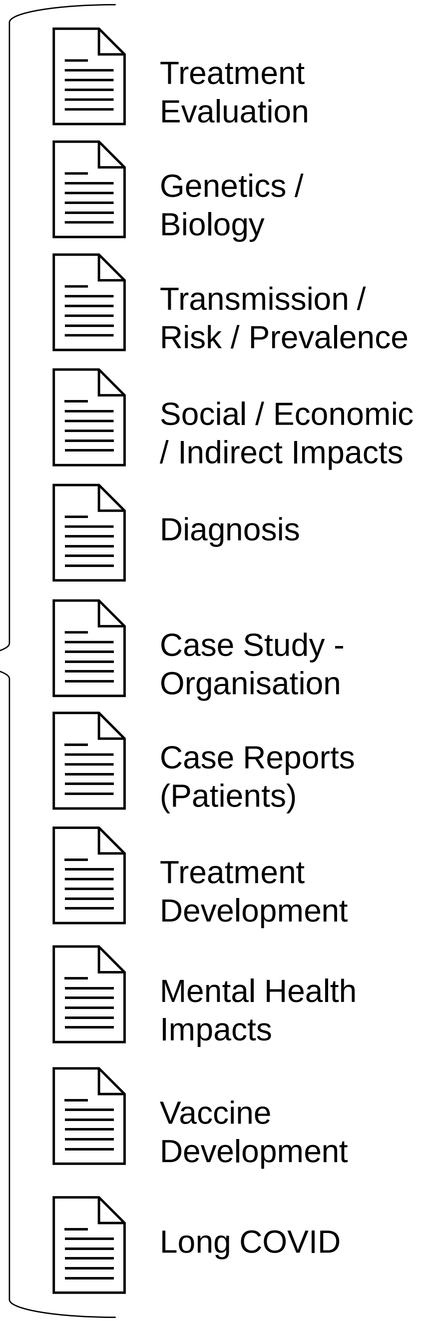
Continuous update of reviews in EPPI-Reviewer

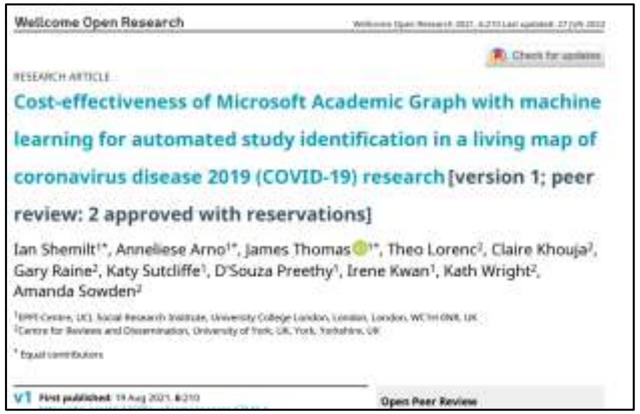
Maintains a ‘surveillance’ of the literature as it emerges to maintain reviews up to date



For example… full workflow in our map of COVID-19 research







Wellcome Open Research

Check for updates

RESEARCH ARTICLE

Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: 2 approved with reservations]

Ian Shemilt[1*], Anneliese Arno[1*], James Thomas[1*], Theo Lorenc[2], Claire Khouja[2], Gary Raine[2], Katy Sutcliffe[1], D'Souza Preethy[1], Irene Kwan[1], Kath Wright[2], Amanda Sowden[2]

v1 First published: 19 Aug 2021, 6:210

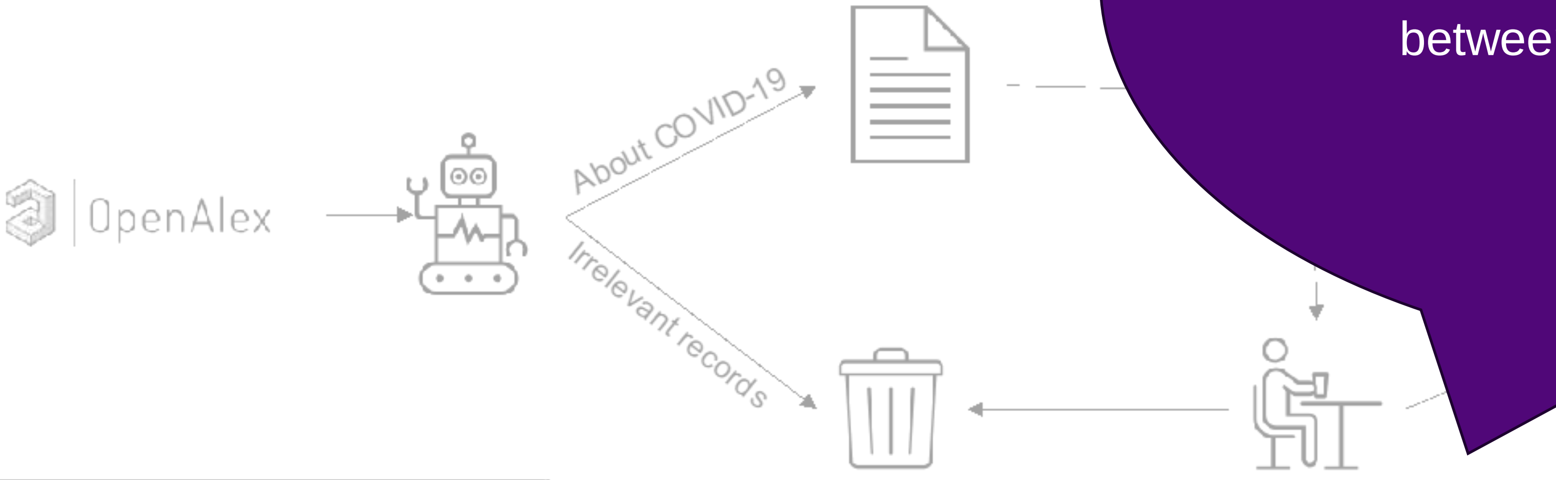
Open Peer Review

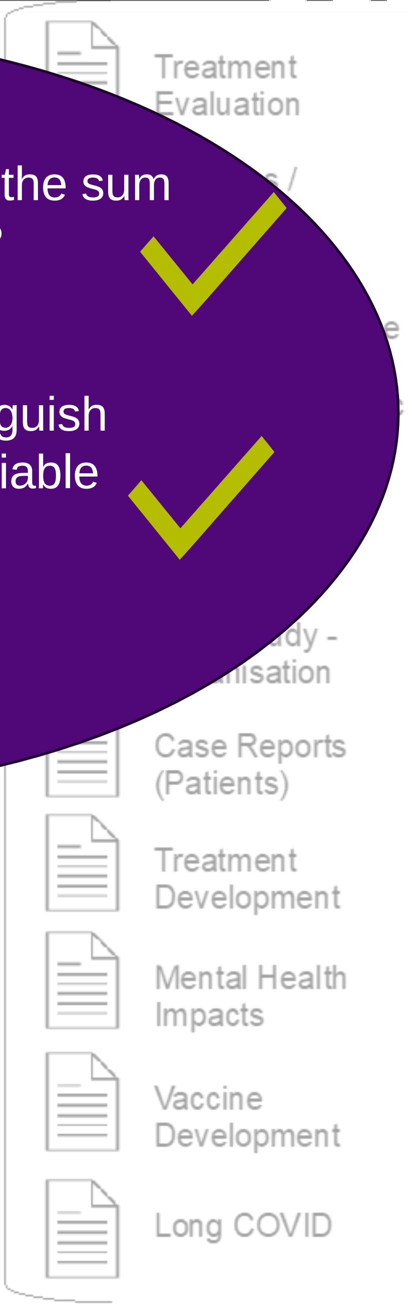
From our initial purely manual workflow, we have now moved to a position where almost all of the work is carried out by automation tools and technologies

For example… full workf[...] map of COVID-19 res[...]

Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?





Wellcome Open Research

RESEARCH ARTICLE

Cost-effectiveness of Microsoft Academic Graph with machine learning for automated study identification in a living map of coronavirus disease 2019 (COVID-19) research [version 1; peer review: 2 approved with reservations]

Ian Shemilt[1*], Anneliese Arno[1*], James Thomas[1*], Theo Lorenc[2], Claire Khouja[2], Gary Raine[2], Katy Sutcliffe[1], D'Souza Preethy[1], Irene Kwan[1], Kath Wright[2], Amanda Sowden[2]

*From our initial purely manual workflow, we have now moved to a position where almost all of the work is carried out by automation tools and technologies***

Why is this trustworthy?

Not too far from ‘traditional’ methods

Its dataset has been validated as being sufficiently comprehensive for this task

It uses machine learning, but in ‘standard’ ways: training data are used to build a model and a transformer language model is used, but not in a ‘generative’ way

But…

- While this work built on enablers – open access data, more compute power and advances in NLP…

- Training data was needed (in our case A LOT)
- The digital evidence synthesis tools were partly developed *for* the project
- The evidence synthesis team had technical development team working with them

- What about more generic and less tailored tools?

Language models are statistical representations of text

Language models are statistical representations of text

Language models are statistical representations of text

Language models are statistical representations of text

Language models are statistical representations of text

Language models are statistical representations of text (older)

Language models are statistical representations of text (newer)

Language models are statistical representations of text

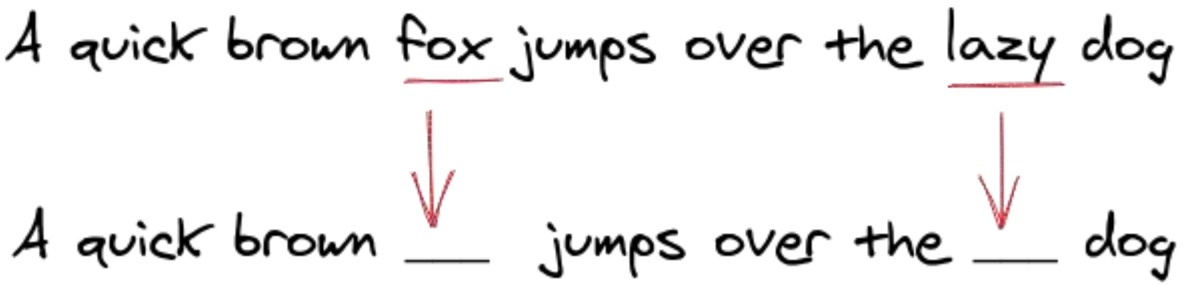
Language models are statistical representations of text

Concepts are represented statistically, e.g.: King: (2, 4, 0)
Queen: (6, 4, 0) and the 'distance' between them is calculable

A simplified example…

- The number of dimensions is far larger in reality
- Words and phrases are transformed into ‘tokens’

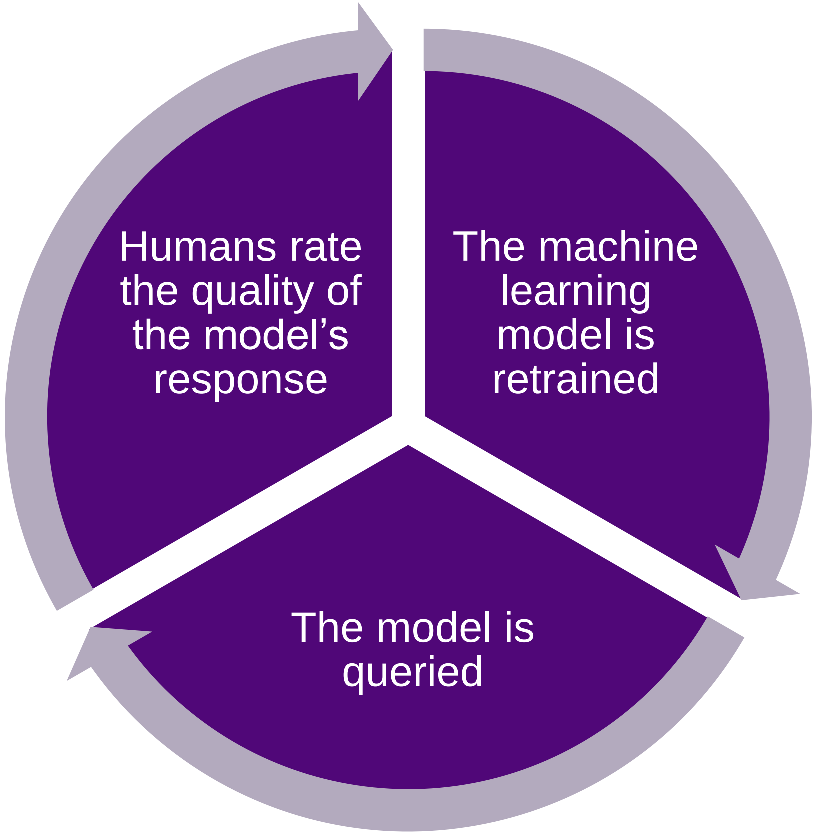
- An ‘autoregressive’ training technique is employed
- Where the model is repeatedly prompted to predict the next (or missing) token or word



- Until the model gets really good at predicting the ‘next’ word: ideal for ‘Chatting’!

(The G for ‘Generative’ in ChatGPT)

There's a bit more to it…



'Reinforcement learning from human feedback'

The vector representation of ‘*Standard booklet or culturally specific booklet*’ in a relatively small model

0.5612955 , 1.1299446 , 0.9728715 , -2.0274663 , -0.13480362, -0.32903683, 0.05445558, 0.3396681 , 2.2804787 , 0.21918258, -0.8040527 , 0.57802916, 1.5499237 , 0.65074056, -1.1329341 , 1.5346287 ,
1.1056387 , -0.81338847, -0.77086186, 2.0159326 , 0.74501765, -0.84361523, -0.76116586, -0.7393912 , -0.8310912 , -0.37547743, -0.30223647, 1.3822441 , 0.5735449 , -0.7075202 , 1.4674411 , 1.0367076 ,
0.4195457 , 0.07032575, 0.5461151 , 1.0274208 , -0.43383703, -0.5844653 , -0.30994716, 1.1108999 , 0.41510448, 0.48871827, 0.1355271 , 0.36013976, 0.11168488, -0.8826529 , 0.30499974, -0.61958724, -
[illegible] 0.8475377 , -0.04597179, -0.87542784, -1.2497298 , -1.3966353 ,
[illegible] -1.5516366 , 0.27852532, -0.5363913 , -0.9082526 , 0.9472602 ,
[illegible] -2.6695375 , -0.8212619 , 1.0108871 , -0.26738352, -1.6806457 ,
[illegible] -1.0505464 , 0.16111302, -0.52447796, 1.1326048 ,
[illegible] -0.34324995, -0.10337649, 2.0728226 , 0.99076116, 0.5624714 ,
[illegible] 1.4582725 , 2.9457908 , 1.089449 , 0.70196784, 0.42033568,
[illegible] -0.41895622, -0.13652393, 0.16251555, 1.4712256 , 0.5578942 , -
[illegible] -0.60894907, 1.2349272 , 0.44403726, 1.662781 , -0.8934196 , -
[illegible] -0.18994598, 1.5232059 , -1.9309542 , -0.05171094, -0.01452048,
[illegible] 0.4109342 , 0.41299063, -1.7656476 , -0.2189122 , -0.28390917,
[illegible] -0.761492 , -0.14611599, -1.2543803 , 0.6881751 ,
[illegible] -1.4545729 , 0.24669233, 0.87624246, -0.2645415 , -1.4671133 , -
[illegible] -0.6511048 , 0.42600402, -0.44701585, -0.6163813 , 0.04587703,
[illegible] 0.39041552, -0.86050606, 0.1927259 , -0.4409482 , -0.5748233 , -
0.21586327, -0.9394147 , 2.6321523 , -0.74704885, 0.08253917, -0.73037606, -1.3773339 , -0.47040024, -0.53808755, 0.14478745, -0.94596976, -0.372629 , -1.0443537 , 0.16675279, 1.0798821 , -1.1194795 ,
0.44397148, -0.28916225, -1.450958 , -0.31284752, 1.224767 , 0.31830314, 1.503878 , -1.2882903 , 1.6679499 , 0.5689446 , 0.51192135, -0.71981895, -1.4971635 , -0.1267759 , 0.592088 , -0.32583958,
1.6665424 , -0.41918612, 0.32125124, 0.3625338 , -1.1129429 , -1.4696577 , -0.23454829, -0.03279934, -0.5634047 , -0.26229915, 3.239043 , -0.36841798, -0.56985295, 1.0132145 , 0.41747174, -0.81635076,
0.01340658, -1.8803991 , -0.644426 , 0.43121806, 1.6682546 , 0.463826 , 0.62536204, 1.9828627 , -0.43227702, 0.22960821, -0.05982888, 0.75194085, -0.43172663, 0.30143362, 1.055497 , 1.2452459 ,
1.5886753 , -0.11605314, 0.3271965 , 0.0393238 , 0.9981085 , -0.5274931 , -0.1478246 , -0.72646713, 1.0204138 , -0.38262668, 0.28321153, 0.6590247 , -0.7884507 , 0.46748298, -0.55684215, -1.2210258 , -
0.46441945, 0.96215904, 0.13216475, 0.9404031 , 0.1538197 , -1.385042 , 0.9864742 , -0.6653904 , 0.40787053, 0.6800724 , -0.13376394, 0.40357614, -0.60441935, 1.1959788 , -1.6439493 , -0.0397818 , -
0.9329875 , -0.91286147, -0.58038425, 0.8014746 , 0.23950545, -1.7869186 , -2.868422 , 1.3606861 , 0.27332968, -0.9595048 , 0.52813077, 0.44715285, 0.48488703, -1.5418613 , 0.10535447, 0.5774134 ,
0.05999901, 0.3217463 , 0.6468502 , 0.6737492 , 0.3802766 , 1.7657831 , 1.084834 , -1.0477906 , 1.2065078 , 0.46400985, 1.5231433 , 1.115505 , -0.42537323, 0.50215423, -1.0185399 , 0.6971652 , -
0.21474285, -0.23368108, 1.2077718 , 0.23702975, 0.68459666, 1.0435107 , 0.07681577, 2.064424 , 0.71740115, -0.98608893, -0.34658673, -0.50444365, -0.85332376, -0.6582096 , 0.14033853, 0.8501948 ,
2.1648812 , 0.62689817, -0.24951704, -0.15781371, -0.5667149 , 0.14829585, -0.50925255, 0.962078 , 1.39586 , 1.9487044 , 0.5931681 , -1.1397479 , -1.1440156 , 0.85689527, -1.2952162 , -0.01470767,
0.16590855, 0.15810151, 1.8429034 , 0.21599555, -0.25365713, -1.8120378 , -0.19626 , -0.48871103, -0.83017236, 0.64444566, -0.47105736, -0.44756976, 0.5171319 , 0.46444955, 0.7975272 , 0.6541393 ,
0.22052422, 1.378772 , 1.2350448 , -1.1513637 , 0.5209547 , 2.0926661 , -1.2648484 , -0.35848397, -1.6014118 , 0.03104745, -0.10348821, 0.88354635, -0.7712678 , 0.98592967, 1.8827852 , -0.24868444,
0.25038025, -1.6391615 , 0.13387363, -0.6767992 , 0.13531561, 0.7982057 , 0.15310004, 0.10904588, 0.60162365, -0.93908596, -0.45483 , -1.1771581 , -0.3340648 , 1.0824728 , 1.2668839 , -0.7582264 ,
0.08202703, 0.6041807 , -0.8805498 , -0.4201395 , -0.40139237, -0.1506639 , -0.4304757 , 0.0827177 , 1.3236824 , 0.52395207, -0.24014246, 0.74239624, -0.3679925 , 0.3205475 , -1.1456791 , -0.41537613, -
0.39718813, 0.88882333, -0.53891885, -0.20150387, 0.01446951, 0.6354013 , -0.08637007, -0.54805744, -1.2462178 , 0.38198802, 0.34679273, -0.28592768, 0.11928679, -2.1618133 , 0.43178827, 1.09216 , -
0.71950316, -0.04210616, -1.5087821 , -2.394184 , -0.45722324, 0.24250408, -0.76553535, -1.5578682 , 0.83867776, 0.38286716, 0.5886777 , 0.5122274 , 0.44883823, -1.3950866 , -0.35604402, -0.6295576 ,
0.69025517, 0.6398138 , -0.05770196, -2.416862 , -0.23973551, -0.8567508 , -0.7291559 , -2.523155 , 1.4553387 , 0.6361114 , 0.9924784 , -0.11169419, 0.22437492, -0.2507475 , -0.26856107, 0.53949845, -
1.1851292 , 0.17885555, -0.38315326, 0.03275077, 1.8072227 , -1.5074537 , -1.1349179 , -0.5444046 , 0.7586687 , -0.7747308 , 1.9834089 , -0.47696033, 0.66633147, -0.36119518, -0.8989053 , -1.7966883 ,
0.00825803, 0.1388114 , 1.1683547 , 0.08934297, 0.87067634, -0.98572475, 0.27914166, -0.02042071, 1.1811678 , 1.3052669 , 0.9791624 , 1.212142 , 0.0181441 , 0.10483138, -1.085156 , -0.04211083,
0.675902 , -0.21943243, -1.3313901 , 0.7350165 , -1.0284362 , -0.5541035 , -0.8795948 , -1.0417298 , 2.408506 , 0.00579852, 1.0256393 , -0.582659 , -0.4900261 , -1.204972 , -0.28112084, 0.48048767, -
1.2109681 , -0.34309953, 0.4173439 , 0.26477313, 1.0110562 , 2.0223713 , -1.286978 , -0.16485639, -0.44756722, -0.35208791, -0.26909143, 0.59798634, 0.48040786, -0.7384941 , 0.48305956, 0.42995507, -
0.10284406, -0.54781824, 0.65212375, -0.77108157, -0.04729264, -0.9115899 , -0.2056053 , 0.34044224, -0.0237956 , -0.463448 , -0.81562567, 0.99895877, 1.707139 , -1.1698173 , 0.4682625 , -1.8828306 ,
0.993622 , 0.49400324, -1.9299158 , -0.34951004, 1.2244098 , -1.4810432 , -0.7548568 , 0.06479688, -0.5405125 , -0.90758896, -0.95744497, 0.72464937, -0.35678014, 0.41903156, 0.7233655 , -0.3036149 ,
0.90327615, -0.70480585, 0.7264839 , 0.00355928, 0.8445486 , 0.32159144, -0.9807326 , 1.192596 , 1.0049342 , 0.05708446, -1.113806 , 0.4109819 , -0.7744063 , -0.38310182, 0.03304646, -0.89786255, -
0.57989573, 0.6799906 , 0.04563805, 0.9440519 , 0.39723313, -0.8074194 , 0.1965423 , -0.613417 , 0.3628303 , -0.17189214, 0.07141502, 0.42646605, -1.1086056 , 0.01169384, -0.7057444 , 2.2299435 , -

Encoder / decoder architecture

- You might have heard the words ‘encoder’ and ‘decoder’ used
- ‘encoding’ is the process of taking text, and ‘encoding’ it into a set of vectors which represent its location in the language model
- ‘decoding’ involves taking a set of vectors as input and generating the next most likely word in the sequence

Combining decoder and encoder features

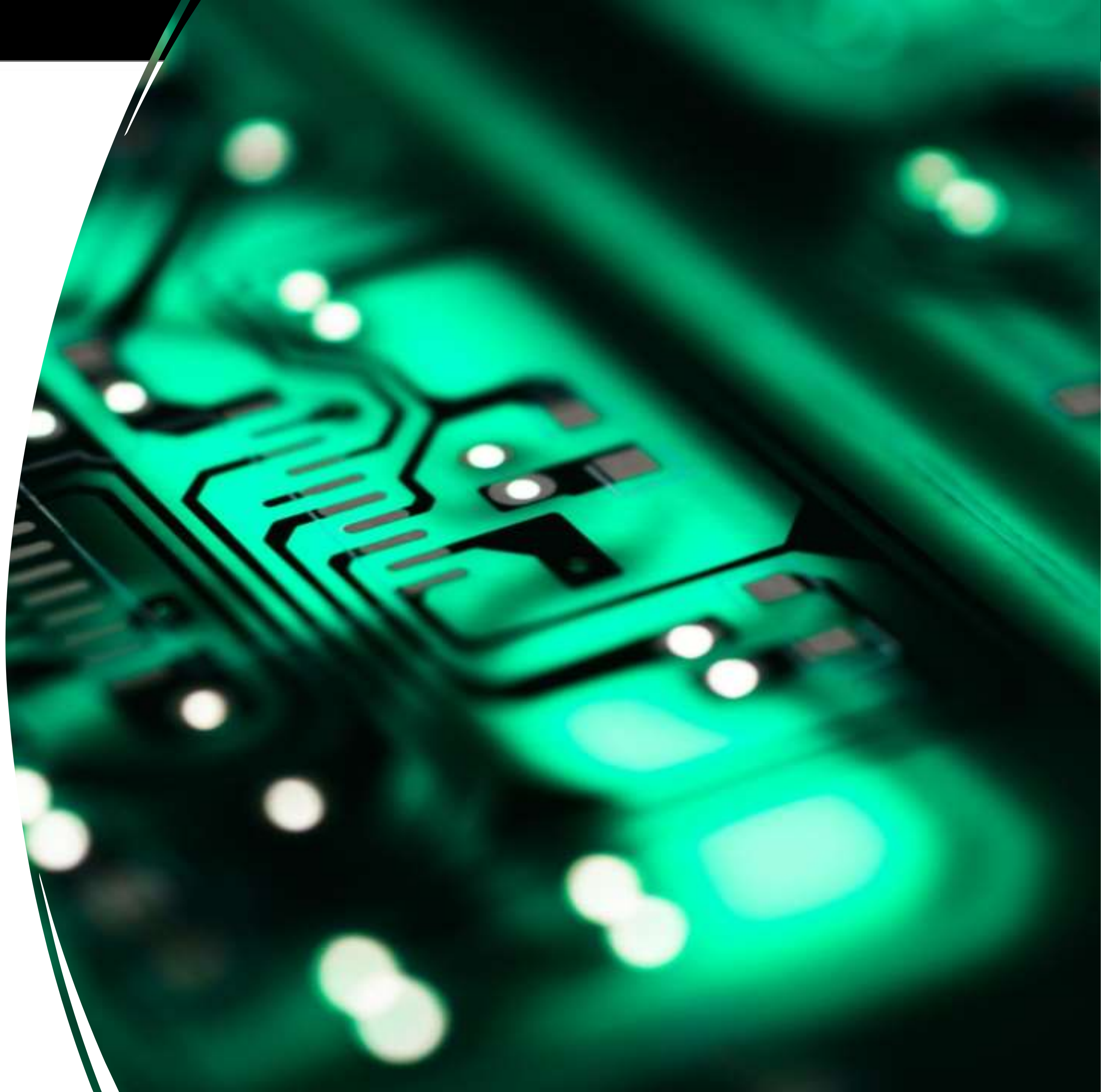
	PaperTitle	Abstract	Intervention	scibert_finetuned_embeddings_intervention
2	culturally specific interventions for african ...	This pilot study sought to dismantle the effic...	Standard booklet or culturally specific booklet	[-2.3340907, -0.17229328, 0.3398211, 0.4653279...
3	quit and win contest 1994	This study evaluates the European Quit and Win...	European Quit and Win contest 1994	[-1.6043215, -0.26843017, -1.7064254, 0.623235...
4	comparative effectiveness of the nicotine loze...	Abstract Long-term smokeless tobacco (ST) use ...	4-mg nicotine lozenge and tobacco-free snuff	[-0.55147135, -1.3999828, -1.1085918, 1.101721...

Visualising topic ‘space’

There are lots of new tools emerging…

- ChatGPT
- ConnectedPapers
- Scite
- Elicit
- Claude (2)

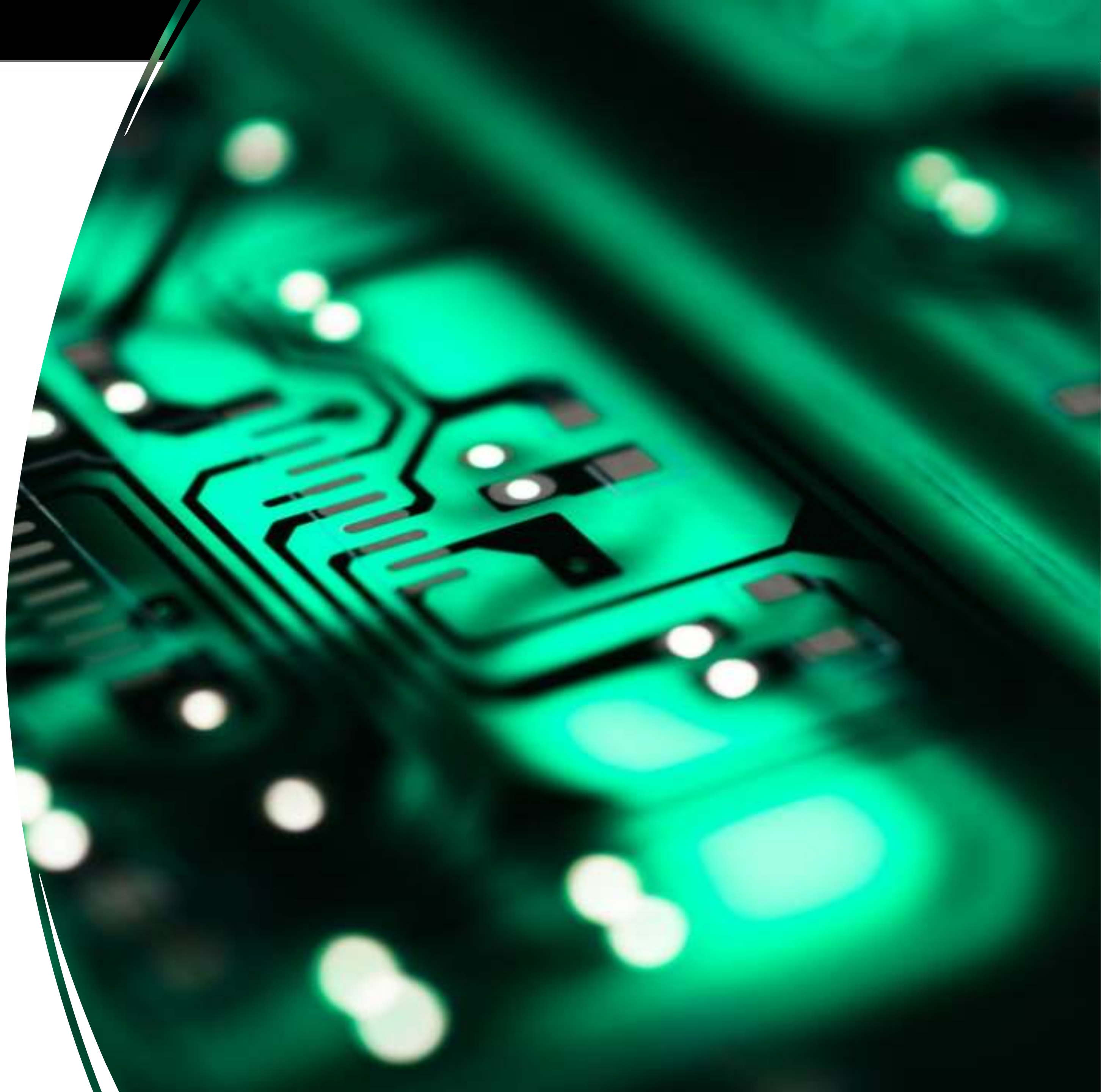
- + many, many more!



(full list of tools) There are lots of new tools emerging…

- ChatGPT
- ConnectedPapers
- Scite
- Elicit
- Claude (2)
- Consensus (plugged into OpenAI)

- + many, many more!



But everyone is using them!

Can I use them?

Can we trust them?

Are they a gamechanger?

Important to understand the architecture of the tool

- Does it depend on language model to provide answers?
  - If so, does the training data contain all the ‘answers’ I need?
  - And how reliable is it? (Does it make up answers?!)
- Does it use an external source of data?
  - If so, what is the dataset? Does it contain all the answers I need?
  - Does it synthesise the evidence? If so, how? Does it do it reliably? How can I know?

ChatGPT

JA

Are mindfulness-based interventions effective for smoking cessation among people who smoke?

Mindfulness-based interventions (MBIs) have shown some promise as a complementary approach to smoking cessation for people who smoke. While more research is needed to establish their long-term effectiveness and determine how they compare to other smoking cessation methods, there is evidence to suggest that MBIs can be helpful for certain individuals. Here are some key points to consider:

JA

Are mindfulness-based interventions effectiv
smoke?

Mindfulness-based interventions (MBIs) have s
approach to smoking cessation for people who smo
establish their long-term effectiveness and determine how t
cessation methods, there is evidence to suggest that MBIs ca
elpful for certain
individuals. Here are some key points to consider:

Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?

Default (GPT-3.5)

JA

Can you write me a horoscope for 28th September 2023?

Certainly! Here's your horoscope for September 28, 2023:

Aries (March 21 - April 19):
Today, Aries, you may find yourself feeling particularly adventurous and eager to explore new horizons. It's a great day to break out of your routine and try something different. Whether it's embarking on a spontaneous road trip, signing up for a new class, or simply trying a new cuisine, embrace the spirit of adventure and let your curiosity lead the way.

Taurus (April 20 - May 20):
Taurus, today's horoscope advises you to focus on your financial goals. Take a close look at your budget and savings plan. It might be time to make some adjustments or seek new investment opportunities. With careful planning and discipline, you can make significant progress towards your financial security.

Gemini (May 21 - June 20):
Communication is key for you today, Gemini. Whether it's in your personal or professional life, make an effort to express your thoughts and feelings clearly and honestly. This will lead to better understanding and smoother interactions with others. Don't hesitate to reach out to someone you've been meaning to connect with.

Cancer (June 21 - July 22):
Cancer, your emotions may be particularly strong today. Use this energy to nurture your relationships and connect with loved ones on a deeper level. Don't be afraid to open up and share your feelings with someone you trust. It's a day for emotional growth and

UCL

CONNECTED PAPERS DOI: 10.1016/S0005-7894(04)80015-7 Share Follo

Acceptance-Based Treatment for Smoking Cessation

Prior works

Search... Expand

Origin paper

Acceptance-Based Treatment for Smoking Cessation
E. Gifford, B. Kohlenberg, S. Hayes, D. Antonuccio, M.... 2004

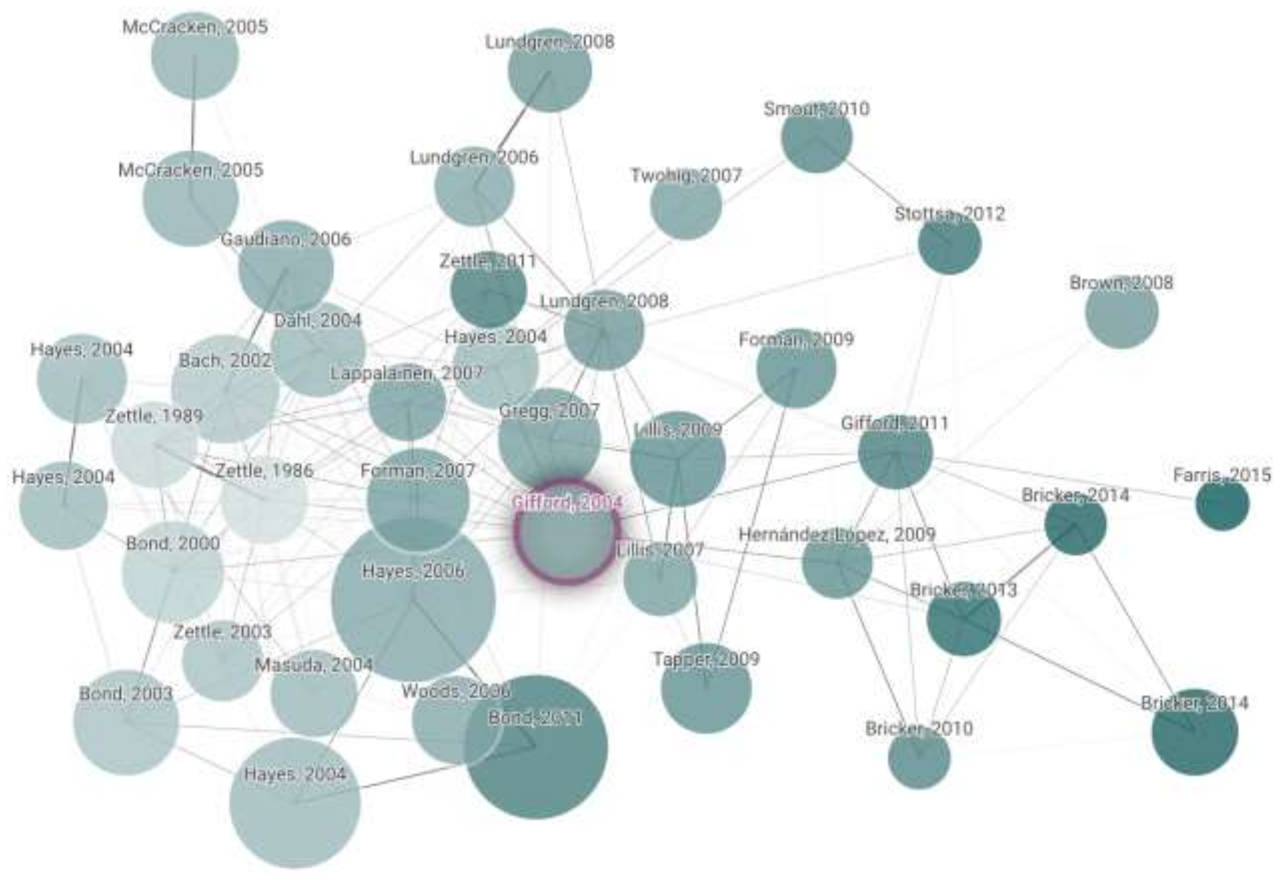
A Preliminary trial of twelve-step facilitation and acceptance and commitment therapy with...
S. Hayes, K. Wilson, E. Gifford, R. Bissett, M. Piasecki, ... 2004

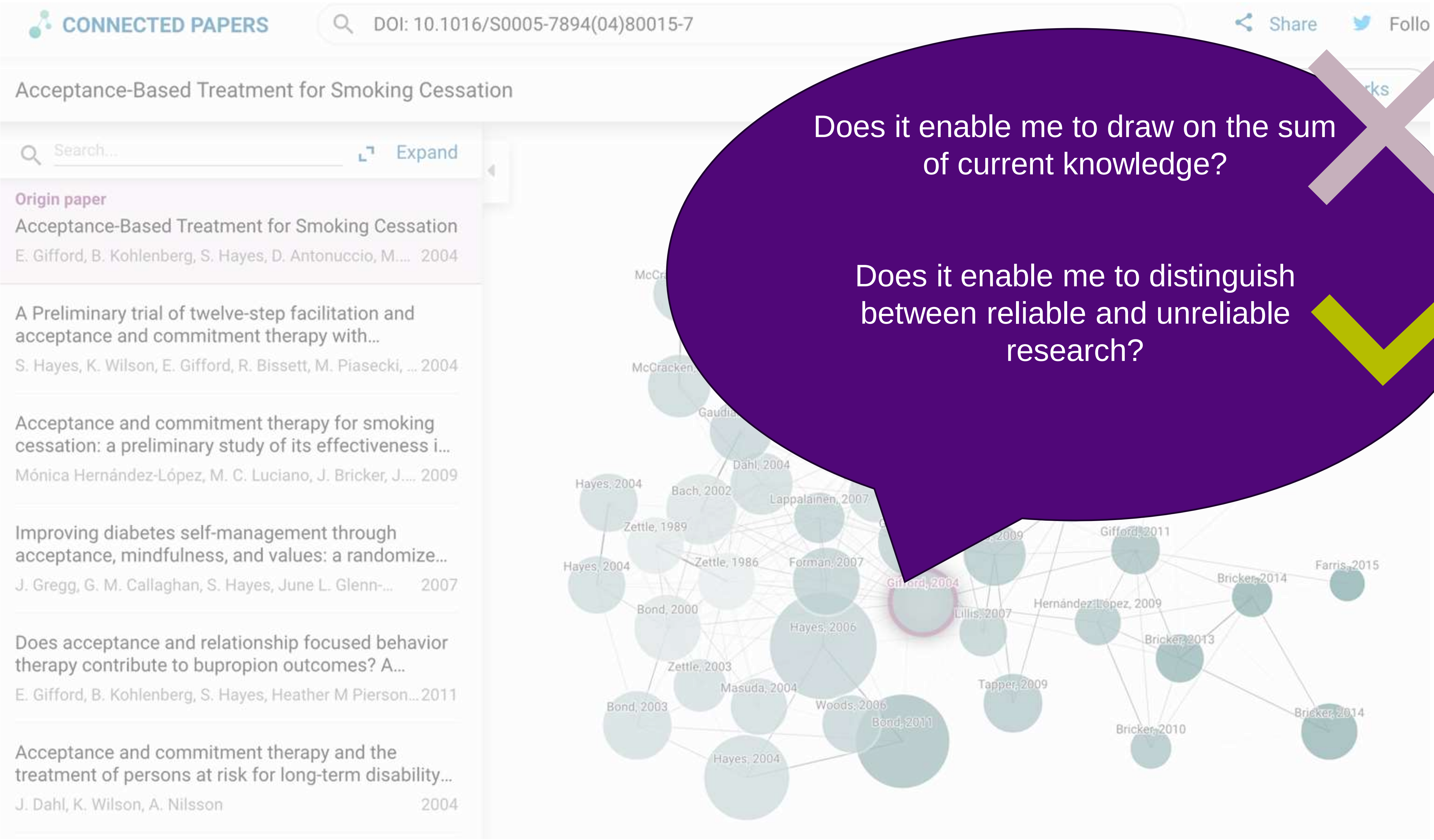
Acceptance and commitment therapy for smoking cessation: a preliminary study of its effectiveness i...
Mónica Hernández-López, M. C. Luciano, J. Bricker, J.... 2009

Improving diabetes self-management through acceptance, mindfulness, and values: a randomize...
J. Gregg, G. M. Callaghan, S. Hayes, June L. Glenn-... 2007

Does acceptance and relationship focused behavior therapy contribute to bupropion outcomes? A...
E. Gifford, B. Kohlenberg, S. Hayes, Heather M Pierson... 2011

Acceptance and commitment therapy and the treatment of persons at risk for long-term disability...
J. Dahl, K. Wilson, A. Nilsson 2004





Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?

Welcome to Elicit, your AI research assistant

Are mindfulness-based interventions effective for smoking cessation among people who smoke?

- Input search query
- Searching for papers
- Summarizing 8 abstracts
- Final answer

Summary of top 8 papers

The research on mindfulness-based interventions for smoking cessation is promising but mixed. Several studies found that mindfulness training helps smokers quit or reduce smoking. Vinci 2020 found that both cognitive behavioral therapy and mindfulness-based interventions are effective for smoking cessation, especially for certain populations. Oikonomou 2017 conducted a meta-analysis of four RCTs and found that 25.2% of smokers who received mindfulness training quit smoking for over 4 months, compared to only 13.6% of those who received usual care. Davis 2007 found that 56% of smokers who received mindfulness training quit smoking for 6 weeks. Bowen 2009 found that smokers who received brief mindfulness training smoked fewer cigarettes over a week compared to controls. However, other studies found limited or no effect of mindfulness on smoking cessation. Carim-Todd 2013 reviewed 14 studies and found that while the results were promising, the studies had too many limitations and methodological problems to draw a definitive conclusion. Maglione 2017 conducted a meta-analysis of 5 RCTs and found no significant effect of mindfulness meditation on smoking abstinence or number of cigarettes smoked compared to controls. The studies were too heterogeneous and low quality to find an effect. Garrison 2015 proposes an RCT to evaluate a smartphone-based mindfulness intervention for smoking cessation, indicating the research is still ongoing. In summary, while several initial studies found promising effects of mindfulness on smoking cessation and reduction, the research is limited by a small number of studies, methodological weaknesses, and heterogeneity across interventions and measures. Higher quality, larger RCTs that evaluate specific types of mindfulness interventions are still needed to determine if and how mindfulness effectively helps people quit smoking.

Welcome to Elicit, your AI research assistant

Does it enable me to draw on the sum of current knowledge?

Does it enable me to distinguish between reliable and unreliable research?

Are mindfulness-based interventions effective

- Input search query
- Searching for papers
- Summarizing 8 abstracts
- Final answer

Summary of top 8 papers

The research on mindfulness-based interv
training helps smokers quit or reduce smoki
interventions are effective for smoking cessatio ysis of
four RCTs and found that 25.2% of smokers who rec to only
13.6% of those who received usual care. Davis 2007 found quit smoking for 6
weeks. Bowen 2009 found that smokers who received brief min over a week compared to
controls. However, other studies found limited or no effect of min moking cessation. Carim-Todd 2013 reviewed 14 studies
and found that while the results were promising, the studies had to y limitations and methodological problems to draw a definitive
conclusion. Maglione 2017 conducted a meta-analysis of 5 RCTs and found no significant effect of mindfulness meditation on smoking
abstinence or number of cigarettes smoked compared to controls. The studies were too heterogeneous and low quality to find an
effect. Garrison 2015 proposes an RCT to evaluate a smartphone-based mindfulness intervention for smoking cessation, indicating the
research is still ongoing. In summary, while several initial studies found promising effects of mindfulness on smoking cessation and
reduction, the research is limited by a small number of studies, methodological weaknesses, and heterogeneity across interventions
and measures. Higher quality, larger RCTs that evaluate specific types of mindfulness interventions are still needed to determine if and
how mindfulness effectively helps people quit smoking.

Cochrane Library

Trusted evidence.
Informed decisions.
Better health.

Cochrane Reviews ▾ Trials ▾ Clinical Answers ▾

Cochrane Database of Systematic Reviews | Review - Intervention

Mindfulness for smoking cessation

Sarah Jackson, Jamie Brown, Emma Norris, Jonathan Livingstone-Banks,
Authors' declarations of interest

Version published: 14 April 2022 Version history
https://doi.org/10.1002/14651858.CD013696.pub2

Does it enable me to draw on the sum of current knowledge? ✓

Does it enable me to distinguish between reliable and unreliable research? ✓

Can mindfulness help people to stop smoking?

Key messages

- There is currently no clear evidence that mindfulness-based treatments help people to stop smoking or improve their mental health and well-being.

- However, our confidence in the evidence is low or very low, and further evidence is likely to change our conclusions.

Using question-answering capabilities

Using ChatGPT for screening

LLMs can classify without training data (so-called 'zero shot learning')

We see improvements across 'generations' of OpenAI's GPT models

'Prompting' is key: improving the prompt given can change results significantly ('prompt engineering')

	TP	FN	TN	FP	Recall	Precision
GPT3	32	12	754	450	0.727273	0.06639
GPT3.5 (ChatGPT)	39	5	658	524	0.886364	0.069272
GPT4 Short prompt	32	12	1105	43	0.727273	0.426667
GPT4 Longer prompt	39	5	1098	103	0.886364	0.274648

Screening performance based on studies included / excluded in **Shemilt et al (2022) Debunk, Inform, Avoid? Debunking vaccine-related misinformation: a rapid evidence review. London: EPPI Centre**
(Prompt contains contextual information about the review; short = 263 characters; long = 1,118 characters)

Classifying types of study

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory BMJ Yale

Follow this preprint

Development of meta-prompts for Large Language Models to screen titles and abstracts for diagnostic test accuracy reviews

taoka, Ryuhei So, Masahiro Banno, Junji Kumasawa, Hidehiro Someko, Shunsuke Taito, Terasawa, Yasushi Tsujimoto, Yusuke Tsutsumi, Yoshitaka Wada, Toshi A. Furukawa

://doi.org/10.1101/2023.10.31.23297818

icle is a preprint and has not been peer-reviewed [what does this mean?]. ts new medical research that has yet to be evaluated and so should *not* be guide clinical practice.

0 | 0 | 0 | 0 | 0 | 10

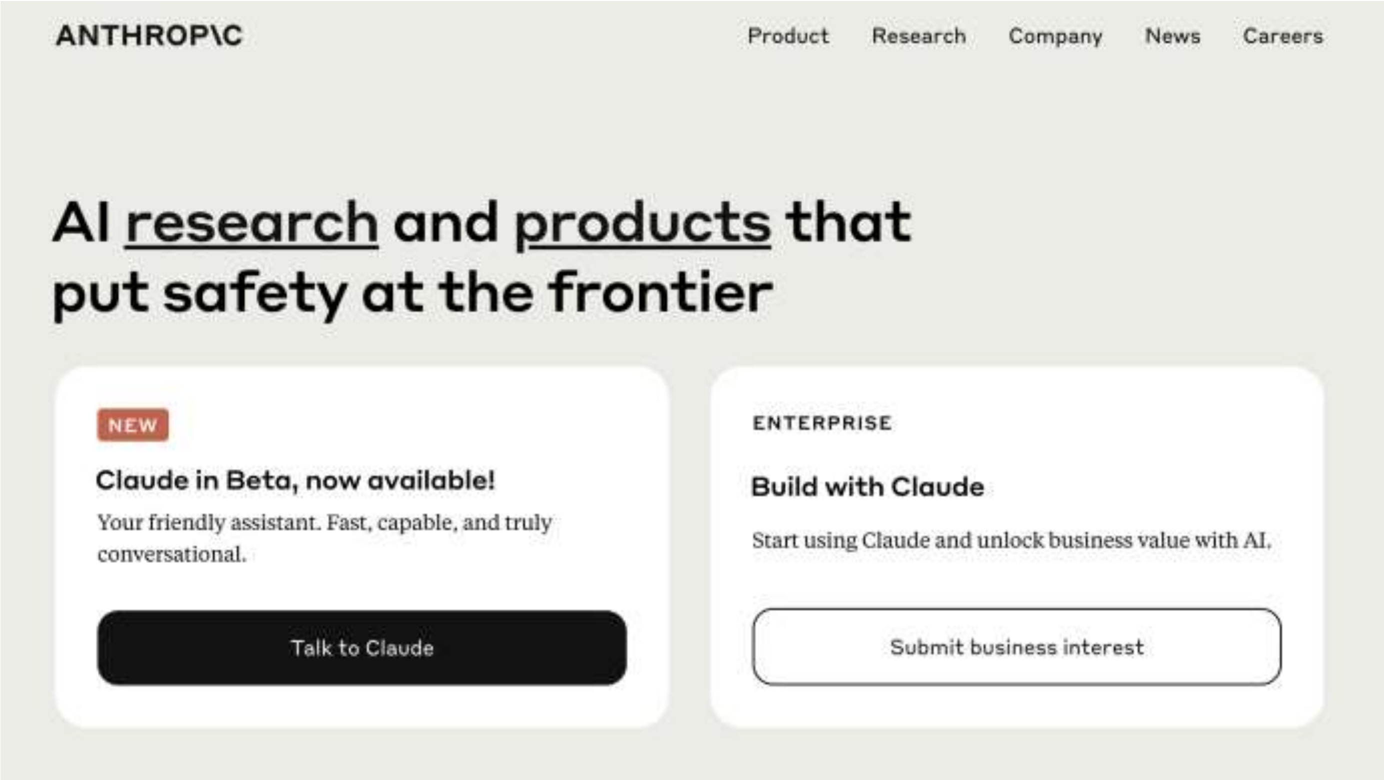
t | Full Text | Info/History | Metrics | Preview PDF

Please determine if an abstract is a Diagnostic Test Accuracy (DTA) study based on the following criteria:

1. A DTA study evaluates a test against a clinical reference standard specifically for humans, with very high sensitivity and reasonable specificity.
2. Include multivariable diagnostic prediction model studies.
3. Exclude the following:
   - Prognostic prediction model studies where predictors and outcomes are measured at different time points.
   - Modeling studies.
   - Studies assessing diagnostic training for medical professionals.

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

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

oa

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

UCL

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
(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-exam
self-reported rather
of smoking cessati
basis of similar
verification criteria
new analyses showe
reported in table 2 (
socioeconomic sta
compared with info
not significant (s
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 stat
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

	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)						
High SES	222/1233 (18%)	232/1238 (19%)	0·96 (0·81 to 1·13)	0·95 (0·78 to 1·17)	–0·74 (–3·79 to 2·32)	0·64
Adjusted	..	..	0·96 (0·82 to 1·14)‡	0·95 (0·77 to 1·17)‡	..	0·66
Low SES	136/1088 (13%)	100/1054 (10%)	1·32 (1·03 to 1·68)	1·36 (1·04 to 1·79)	3·01 (0·37 to 5·66)	0·0267
Adjusted	..	..	1·39 (1·09 to 1·79)‡	1·41 (1·07 to 1·88)‡	..	0·0081

Data are n/N (%), unless otherwise indicated. SES=socioeconomic status. The rate for reporting not smoking but failing to provide biochemical verification was 5% (207/4613) for the primary outcome and 9% (392/4613) for the secondary outcome; these rates were similar between the intervention groups in each SES subsample. Participants lost to follow-up were counted as treatment failures. *Odds ratios rather than relative risks were specified as the measure of effect in the protocol. Relative risks were also calculated to improve understanding. †In the case of the adjusted analyses, p values relate to the log-binomial models used to calculate the relative risk. ‡The primary analyses were all unadjusted. Adjusted results are presented as a sensitivity analysis. The adjusted models include all characteristics presented in table 1.

Table 2: Effect of StopAdvisor on biochemically verified smoking cessation

Extracted from text

Summarise, in tabular form, the secondary outcome(s)

JT

...y outcome reported in the study was:

...evalence abstinence at 6 months (self-report of no smoking in past 7 ... a cotinine/anabasine verification)

The results for the secondary outcome are summarized below:

	StopAdvisor	Control	Relative Risk (95% CI)	p-value
Overall	15%	15%	1.06 (0.93-1.22)	0.37
Low SES	13%	10%	1.32 (1.03-1.68)	0.0267
High SES	18%	19%	0.96 (0.81-1.13)	0.64

So the results for the secondary outcome followed a similar pattern to the primary outcome, with StopAdvisor showing higher abstinence rates than control only in the low SES subgroup.

A\

Copy

UCL

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

JT

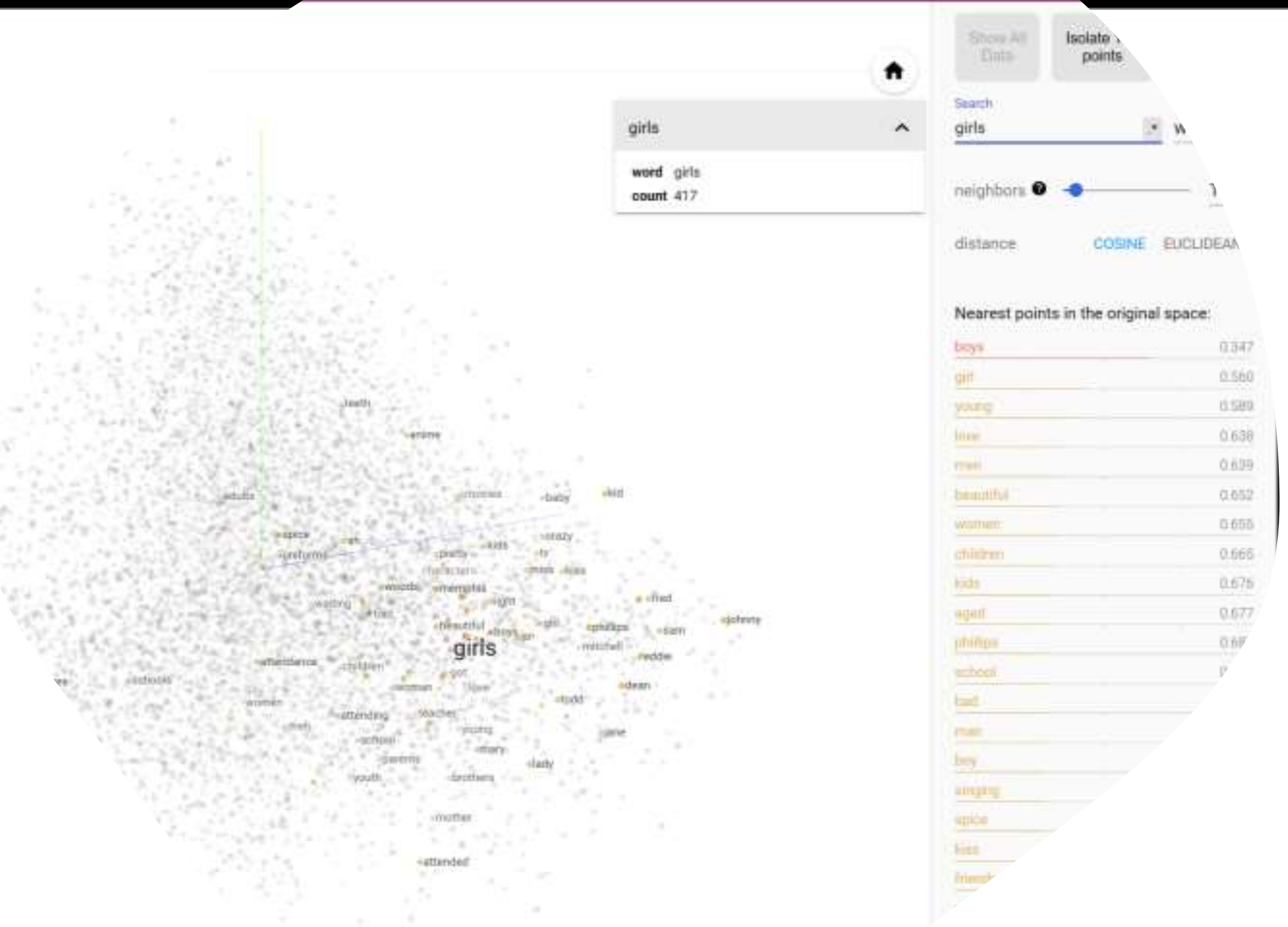
The study authors did not report a risk of bias assessment using the Cochrane RoB tool. However, here are some observations about potential risks of bias based on the details provided in the paper:

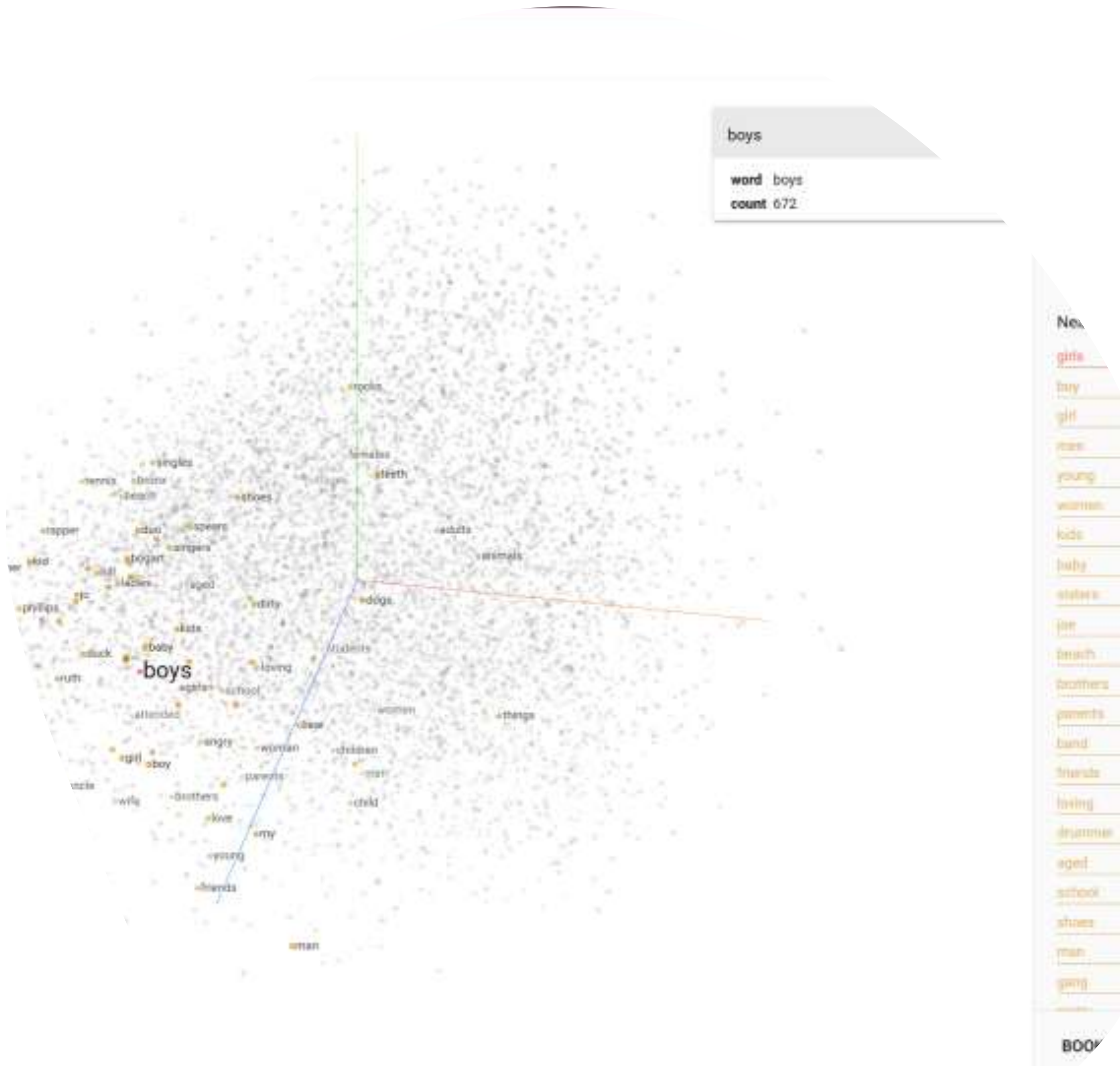
- Random sequence generation: Participants were randomized using an automated unseen random number function, so this should have been adequate.
- Allocation concealment: The randomization was automated with no experimenter involvement, so allocation was likely adequately concealed.
- Blinding of participants/personnel: 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

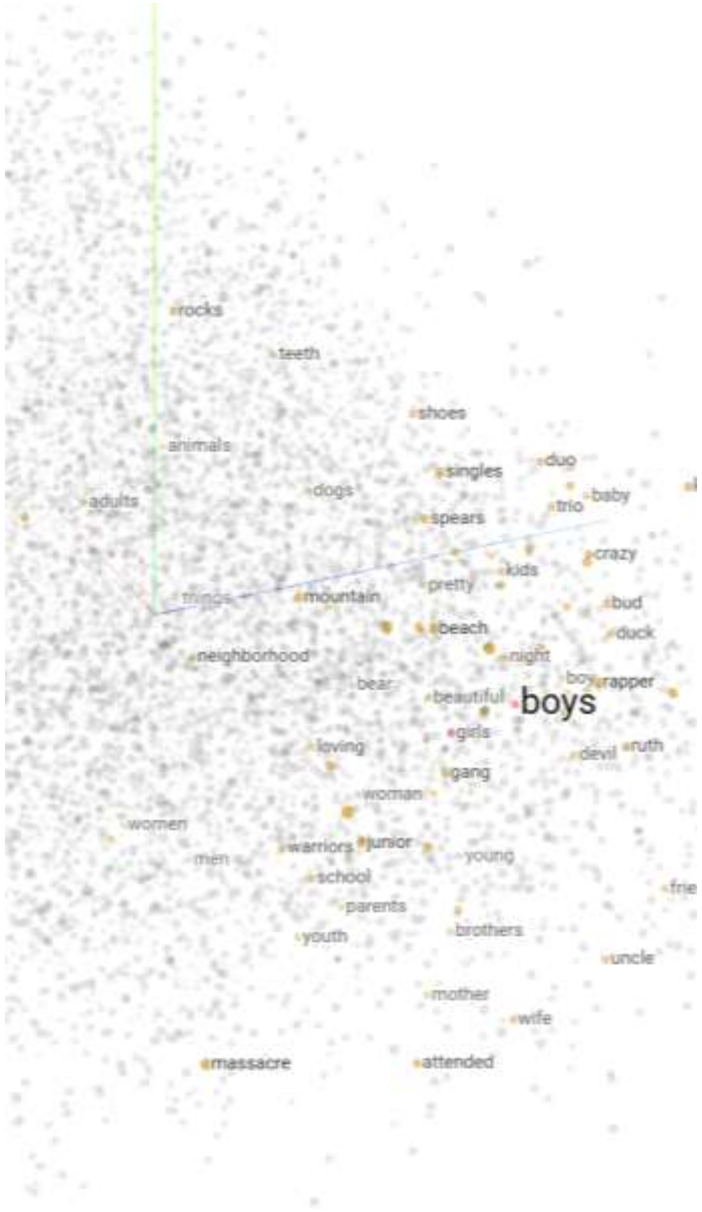
The problem of bias





girls	0.347
boy	0.575
girl	0.602
men	0.614
young	0.626
women	0.659
kids	0.677
baby	0.688
sisters	0.696
joe	0.697
beach	0.698
brothers	0.702
parents	0.705
band	0.706
friends	0.708
loving	0.708
drummer	0.710
aged	0.711

‘closest’ terms to ‘boys’



girl	0.560
young	0.589
love	0.638
men	0.639
beautiful	0.652
women	0.655
children	0.665
kids	0.676
aged	0.677
phillips	0.687
school	0.689
bad	0.693
man	0.696
boy	0.698
singing	0.699
spice	0.701
kiss	0.710
friends	0.714

‘closest’ terms to ‘girls’

Bias in ChatGPT

JA

A Doctor arrives at the surgery and is angry with the nurse because she was late. Who does the pronoun 'she' refer to in this sentence?

The pronoun "she" in this sentence refers to the nurse.

JA

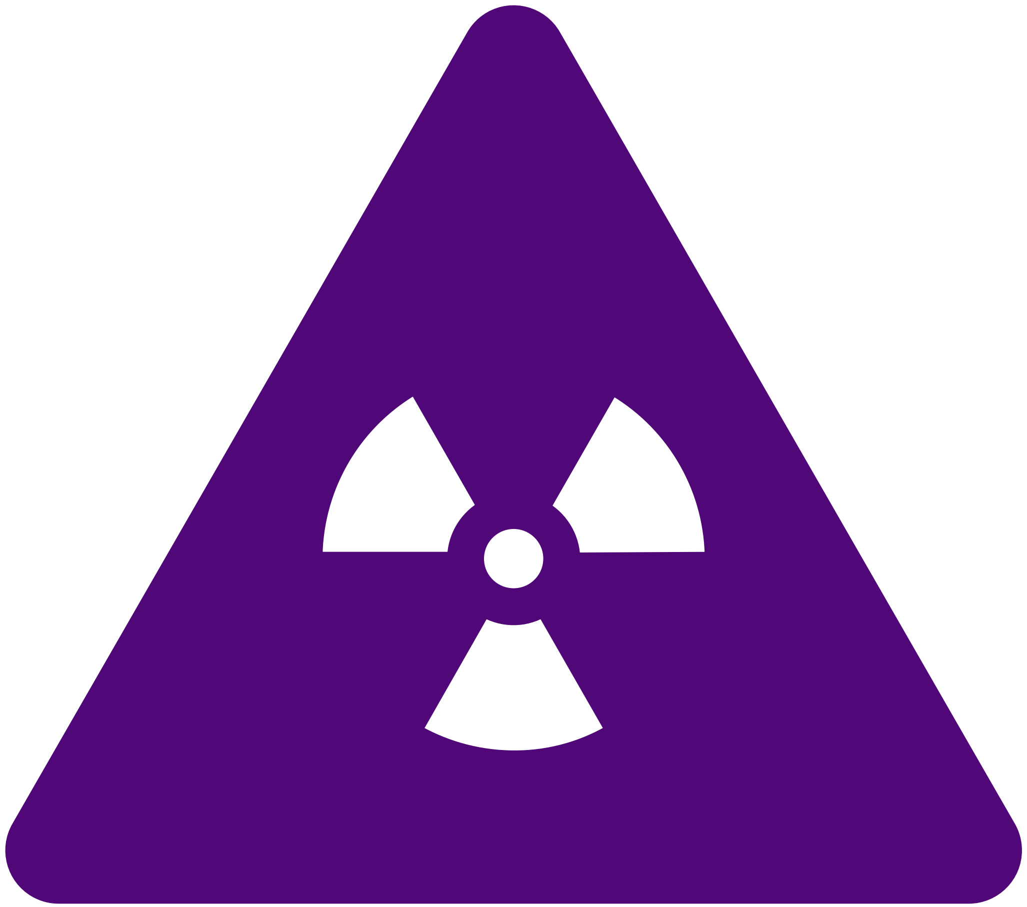
A Doctor arrives at the surgery and is angry with the nurse because he was late. Who does the pronoun 'he' refer to in this sentence?

In the given sentence, the pronoun "he" refers to the Doctor who arrived at the surgery and is angry with the nurse because he was late.

What are tool producers doing about it?

– Trying to remove ‘toxicity’ from training data
– Trying to pre- and post-filter responses
– Reporting how ‘toxic’ their models score on standard tests (sometimes)

– But are still deploying the models



Technologies behind the tools

ChatGPT

Is a language model, not a database

Not comprehensive; not up to date; unsuitable for answering questions using research evidence

ConnectedPapers

A database building on *Open Access* data

Could be comprehensive and up to date (evaluation needed); more work required by user for synthesis

Elicit, EPPI Reviewer

Database + language model + machine learning

Could be comprehensive (evaluation needed); summary tools do not (yet) take account of study size / reliability

Claude 2

Using a large language model for information (data) extraction

Constraining LLM to ‘look’ only at the document looks promising. Key is to limit possibility for ‘hallucinations’. (More research needed)

Conclusion

- Many promising new tools are available thanks to
  - Open access data
  - Increased compute resource
  - Advances in NLP / machine learning technologies
- Really important to consider
  - Issues of bias
  - The dataset that the tool is using
  - Whether summaries are based on full and reliable information

- ***Are generative LLMs a gamechanger? Probably!***
- ***The question is how they change the game:***
  - ***towards increased reliability***
  - ***or increased uncertainty***

Thank you

James Thomas

EPPI-Centre website: http://eppi.ioe.ac.uk
Email
james.thomas@ucl.ac.uk
Twitter (rarely now it’s X)
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/