UNIVERSITEIT UTRECHT

MASTER THESIS

Jump-Starting Evidence Synthesis

Initializing Active Learning Models for Systematic Reviews using LLM-generated Data

Author:

Timo van Ommeren

Supervisors:

Prof. dr. A.G.J. (Rens) van de Schoot, Lauke Stoel

MSc Methodology and Statistics for the Behavioural,
Biomedical and Social Sciences

Methodology Department, Utrecht University

December 8, 2025
Word count: 2193

1 Introduction

500 words

1.1 Background

1.1.1 Active Learning Models for Systematic Reviewing

Researchers and practitioners are continually challenged to base their decisions on the latest scientific evidence. Systematic reviews and meta-analyses were developed to address this need as rigorous methods of summarizing scientific literature (Chalmers, Hedges, and Cooper 2002; Bastian, Glasziou, and Chalmers 2010). However, systematically reviewing the literature can be time-consuming, which limits its practical applicability, especially, in, for example, times of crisis (Tricco et al. 2020; Nussbaumer-Streit et al. 2021).

Fortunately, recent advances in machine learning have produced tools that allow for the systematic screening of scientific literature while greatly reducing the need for manual screening (Van De Schoot et al. 2021). Specifically, active learning models (ALMs) ask users to screen titles and abstracts of papers one by one. Based on the user's decision, the models reassess the probability that the remaining papers are relevant and thus whether to show them to the user. In other words, these models continually reshuffle the papers retrieved from a scientific literature search based on the user's interests. This method reduces the time needed to find as many relevant papers as possible compared to simple index-based screening (Schoot et al. 2025).

1.1.2 The Cold Start Problem

A key challenge to using active learning for systematic reviews is that these models face a ”cold start” (Panda and Ray 2022). For an ALM to query a user with a potentially interesting paper, the model must first have knowledge of the user's interests . One way of overcoming a cold start is to initialize, or 'warm up', the ALM using examples of relevant and irrelevant papers (Teijema et al. 2025). If however no examples are available the user may simply start screening papers at random, until a relevant and an irrelevant paper have been found.

1.1.3 The Advent of Large Language Models

With the recent advent of large language models (LLMs), a new possible solution to overcoming the cold start problem has emerged (Bachmann et al. 2025). Instead of screening papers at random until relevant and irrelevant examples are found, LLMs can generate synthetic examples of both based on the systematic review's inclusion and exclusion criteria. This approach may be particularly advantageous when the percentage of relevant papers returned by a systematic search is low. In this case, screening using even suboptimal examples of relevant and irrelevant papers may be preferable to randomly screening hundreds or thousands of papers. In other words, something may be better than nothing. However, synthetic data may also misdirect the ALM by contaminating the model's training data.

In summary, the 'cold start' problem may be overcome and the performance of ALM-assisted screening improved by initialising the ALM with relevant and irrelevant paper examples generated by LLMs. This approach could be particularly useful if a user has no relevant examples available, as it avoids random screening. A key question, however, is whether LLM-generated examples improve starting performance beyond that achieved through initialisation using the systematic review's inclusion and exclusion criteria directly.

1.2 Objectives

This study aims to investigate the effect of using LLM-generated data to initialise active learning models (ALMs) for systematic reviews, compared to random initialisation, no initialisation or criteria-based initialisation, on starting performance. This will be achieved by simulating the screening of previously published systematic reviews.

2 Methodology

1000 words

2.1 Conditions

This study aims to compare the effect of LLM initialisation with that of three control conditions: random initialisation, no initialisation, and criteria-based initialisation. In all conditions except the 'no initialisation' condition, the LLM is initialised using only relevant papers. This is because systematic reviews contain a large number of irrelevant papers, meaning they no 'cold start' problem.

2.1.1 Experimental condition: LLM initialization

In the LLM initialisation condition, the ALM's classifier is provided with a set of examples comprising at least one relevant paper before screening is simulated. These examples are generated based on the inclusion and exclusion criteria of the given systematic review publication. See figure 1 for a schematic of the simulation pipeline.

Between simulation runs the exact number of abstracts generated as well as their specific contents is varied. More specifically, we aimed to investigate the effect of the following variables on starting performance:

1) number of abstracts generated per simulation run *(1, 4 or 7 abstracts)*,

2) length of abstracts generated per simulation run *(200, 500, 1000 words)*,

3) the temperature setting on the LLM *(0.0, 0.4, 0.8)*,

To instruct the LLM, a DSPY module was created which takes the variables described above as input, along with the inclusion and exclusion criteria of the systematic review and a generic prompt (Khattab et al. 2023). The module then generates the desired number of abstracts using OpenAI's gpt-4o-mini model. For the exact code, see: here

2.1.2 Control conditions

In the *random initialization* condition, one relevant paper was sampled at random prior to the start of screening (with replacement between runs). In the *criteria-based initialisation*, the inclusion and exclusion criteria of the systematic review directly functioned as an example of an abstract of a relevant paper with which to initialise the ALM. In the *no initialisation* condition, no papers were provided prior to the start of screening.

2.2 Outcome variable

2.2.1 Starting performance

Starting performance will be assessed based on the number relevant papers in the first 100 papers screened. This figure is derived from research showing that approximately 100 papers can be screened in an hour (Nussbaumer-Streit et al. 2021). For each simulation, we count the number of relevant records found with a *Time to Discovery* below 100 for each simulation (Ferdinands et al. 2023).

2.3 Simulation set-up

In order to investigate the effect between LLM initialisation and the control conditions on starting performance, the SYNERGY datasets will be screened based on the abstracts and titles of the papers using ASReview's ALM (Van De Schoot et al. 2021; De Bruin et al. 2023).

ASReview is an open-source software package for semi-automated systematic reviewing that implements several ALMs. To simulate the screening process, we will access ASReview via its Python API. For all simulation runs, we will set the ALM to the recommended U4 configuration in ASReview, which combines a support vector machine classifier (seeded with the run number) with a bi-gram term-frequency inverse document frequency (TF-IDF) feature extractor and a querier which always present the next most likely paper [1].

The SYNERGY datasets consist of 24 previously screened and labelled sets of papers. Importantly, this gives us access to the ground truth labels of all the papers in these datasets. See Appendix A for a list of all datasets including their topic, total number of records, number of records included and the percentage of relevant records (De Bruin et al. 2023).

All simulations were done in *Python version 3.10*. For all the code aswell as the full list of the packages and their versions, please see the github repository. See appendix B for a detailed description of how the simulation results were exported.

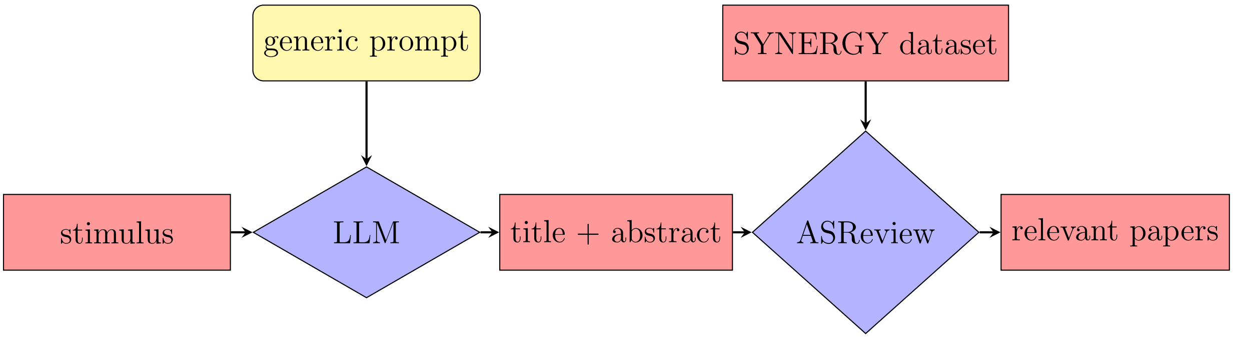

Figure 1: Simulation pipeline (for more detailed schematic of workings within ASReview see (J. d. Bruin et al. 2025))

[1] Note that two ALMs are technically used. A simple random querier is initially used to find at least one relevant and one irrelevant paper, including initialisation examples. After this, the U4 is used.

2.4 Data Generation and Analysis

2.4.1 Padding

It is important to note that the simulation may stop prematurely if all the relevant records are found before the stopping rule is reached (e.g. after screening 100 records). To accurately emulate a researcher who is unaware that all the relevant records have been found and therefore continues screening until the stopping rule is reached, the simulation results were supplemented with rows containing the label 'zero' (i.e. irrelevant records) until the stopping rule would have been reached. This process is referred to as 'padding' and ensures that the final simulation results are accurate.

2.4.2 Analysis

The majority of the variance in the number of relevant records found in the first 100 screened is expected to be explained by differences between datasets. We would therefore like to seperate within-dataset variance from between-dataset variance. Ideal for this purpose are multilevel regression models which allow us to model the effect of condition on starting performance while accounting for between-dataset variance by including dataset as a random effect. To confirm that most of the variance is indeed between datasets, we will visualize the data using boxplots and fit a null model with only dataset as a random effect to calculate the intraclass correlation coefficient (ICC). An exploratory bottom-up modelling approach will then be used for multilevel analysis (for approach see chapter 4 of the book by Hox, Moerbeek, and Van de Schoot 2017) and will be conducted in R using the *lme4* package (Bates et al. 2015).

Note that in the control conditions, the independent variables (i.e., number of abstracts, length of abstracts, temperature) are not applicable, and will therefore be coded as missing. To account for this apparent missingness, a dummy variable indicating whether the condition is the LLM initialisation condition or not will be included in the model using the approach described by (Dziak and Henry 2017)

3 Results (note: preliminary!)

3.1 Descriptives

As expected, most of the variance in starting performance seems to be explained by differences between datasets. Figure 2 shows the number of relevant records found within the first 100 screened per dataset. This may be partly explained by the considerable variation in the number of relevant records present in each dataset. In fact, all relevant records were found within the first 100 records screened, in 6.91% of the runs in the LLM-initialization condition. In the random-initialization condition, this was 0%, in the criteria condition, 8.08%, and in the no-initialization condition, 1.52%.

3.2 Main results

We fit a linear regression to examine whether initialization affected the number of relevant records found within the first 100 screenings. The LLM-initialized model yielded on average $B = 30.11$, $SE = 0.72$, $t = 41.85$. Interestingly, this is approximately equal to number of papers found in the random-initialization condition, at a difference of:, $B =$

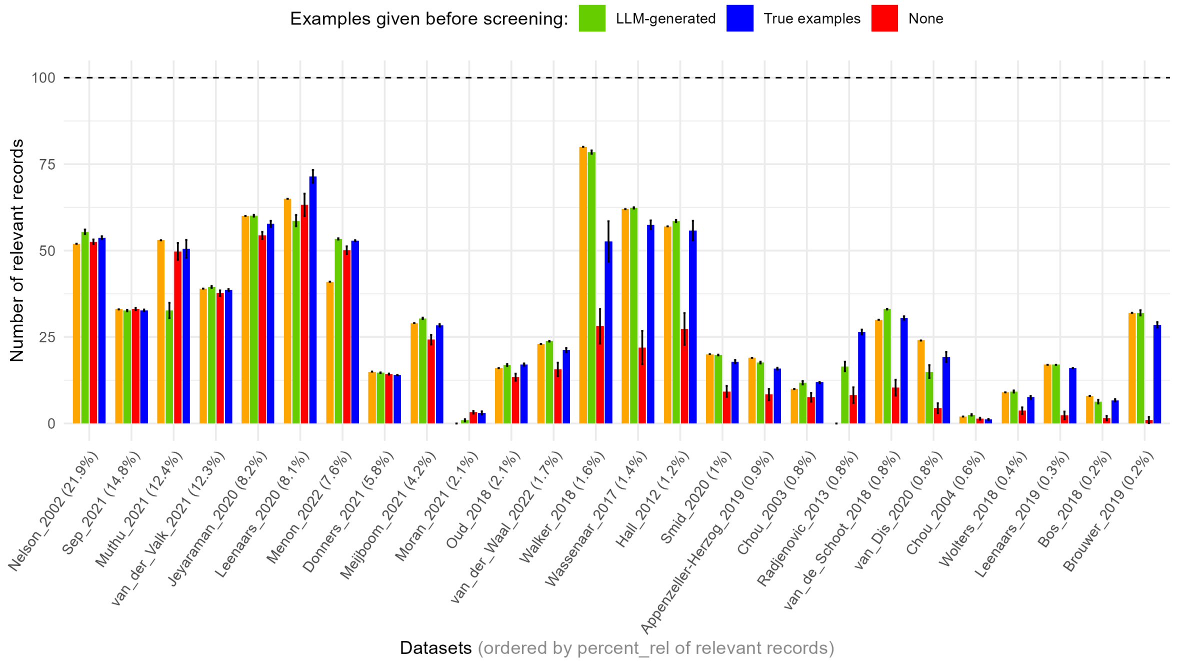

Figure 2: The number of relevant records found per dataset

-0.08, SE = 1.02, t = -0.08, p = 0.934. In contrast, the difference between the LLM-initialization and the no-initialization conditions was statistically significant: B = -9.88, SE = 1.02, t = -9.71, p ¡ .001. Overall model fit was R^2 = 0.05, adjusted R^2 = 0.04, and $F(2, 2613)$ = 62.35, p = p ¡ .001.

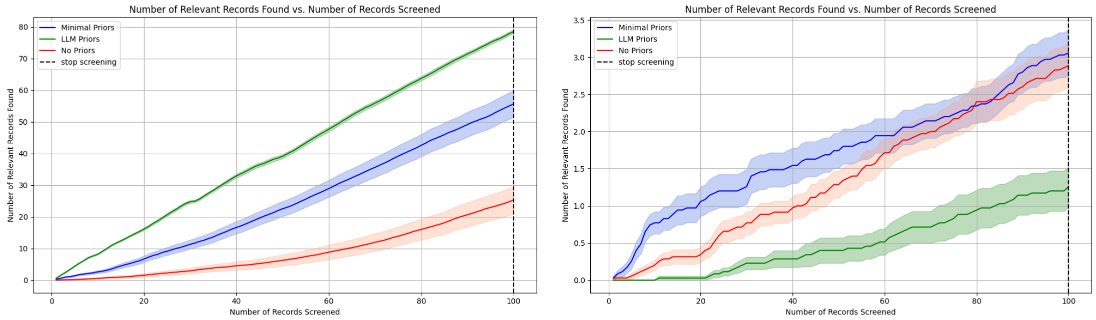

Figure 3: The number of relevant records found per dataset

4 Conclusion

The results of this simulation study provide a proof of concept that LLM-initialization can improve starting performance of active learning models for systematic reviews compared no initialisation. Exploratory analyses suggest that the exact instructions given to the LLM (i.e., prompt engineering) does not seem to have a large effect on starting performance. The key lesson is thus that when it comes to active learning models for systematic reviews, something is better than nothing when it comes to initialisation. We thus recommend that researchers and practitioners consider using LLM-generated exam-

Table 1: Effect of initialisation on the number relevant records found in first 100 screened

	Dependent variable:				
	Outcome: number of records found				
	(1)	(2)	(3)	(4)	
Random initialisation	−0.738	−0.738	−0.738	−0.778	
	(0.445)	(0.445)	(0.445)	(1.859)	
No initialisation	−11.589	−11.589	−11.589	−11.647	
	(0.445)	(0.445)	(0.445)	(2.924)	
Number of abstracts	−0.547	−0.547	−0.547	−0.573	
	(0.445)	(0.445)	(0.445)	(1.302)	
Length abstracts		−0.253			
		(0.126)			
LLM temperature.		−0.009			
		(0.096)			
Dataset size (N records)		0.043			
		(0.966)			
% relevant records			1.802	2.801	
			(0.583)	(0.352)	
% x Random initialisation					
% x No initialisation					
conditionno_initialisation:records					
conditionrandom:records					
Constant	30.931	30.931	23.167	18.896	
	(3.876)	(3.876)	(4.132)	(4.549)	
Var(u_{0j})	358.264	358.264	256.071	440.466	
Var(u_{1j})				80.71	
Var(u_{2j})				202.946	
Var(ε_{ij})	84.556	84.456	84.556	39.48	
$\rho(u_{0j}, u_{1j})$				6.203	
$\rho(u_{0j}, u_{2j})$				-0.18	
$\rho(u_{1j}, u_{2j})$				-0.885	
Δ deviance (df)	6	4.02 (3)	0 (2)	2345.89 (9)***	22
Log Likelihood	−12,503.040	−12,501.030	−12,499.020	−11,326.080	−
Akaike Inf. Crit.	25,018.080	25,020.070	25,012.040	22,684.150	2

ples of relevant and irrelevant papers to initialise active learning models for systematic reviews, especially when no actual examples are available. However, future work should investigate whether initialisation using LLM-generated examples does not negatively impact the overall performance or last-to-find performance of active learning models for systematic reviews by contaminating the training data of these models. Moreover, future work should investigate whether these results generalize to real-world systematic reviews, as the current simulation study may have been affected by data leakage (i.e., LLMs have been trained on synergy datasets).

5 Discussion

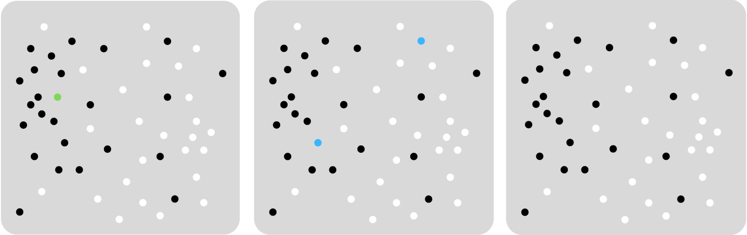

Figure 4: Ideal starting point for systematic reviews using active learning

In other words, whether, in the case of AI-assisted reviewing, prompt engineering actually aids in knowledge discovery or whether LLMs simply repackage existing knowledge.

More LLM-specific variables were considered in advance (such as degree of jargon), however, because early simulations showed difference between the LLM- and criteria-based initialisation conditions, these variables were not further investigated. Future work may consider these variables in more detail.

Since many datasets have less than a 100 relevant records, operationalizing starting performance as the number of relevant records found in the first 100 screened may be limiting, as in some runs all relevant records are found before the stopping rule is reached (). How quickly all relevant records are found is now not considered. Future work may consider alternative operationalizations of starting performance which take this into account.

Most eligibility criteria are written in either conjunctive or disjunctive form. In the latter case, all the inclusion criteria must be met for a paper to be included, whereas in the former, only one must be met. This may affect how useful LLM-generated examples are compared to using the eligibility criteria directly or selecting random examples. This is because, for criteria in conjunctive form, abstracts will be relevant if and only if all the terms from the eligibility criteria are mentioned. In contrast, for criteria in in disjunctive form, the relevant abstracts will likely only contain one of the many terms mentioned in the eligibility criteria.

5.1 Limitations

1. Key limitation simulation study: data leakage (i.e., LLMs have been trained on synergy datasets). The ecological validity of the results are therefore somewhat limited.

   (a) There are two obvious solutions to the problem of data leakage: (1) apply the use of LLm-initialization on a new systematic review, and (2) use an older LLM from hugging face for example.

2. Another possible limitation: no switching of active leaning cycles (could fix contamination of synthetic data issue).

Clusters should not be a problem because we only focus on starting performance (i.e., first 100 screened). Future work may consider the contamination hypothesis in more detail.

A SYNERGY metadata

B Format exported simulation results

Each simulation run is stored in a separate CSV file. Every row represents a screened paper and contains the following information:
1) The paper's record ID,
2) The assigned label (i.e., relevant or irrelevant),
3) The classifier, querier, balancer and feature extractor used,
4) The size of the training set,
5) A time tag.

Furthermore, the following naming convention is used for the CSV files: condition_run_run_IVs_n_abstracts_length_abstracts_llm_temperature.csv. The same naming convention is used for the recall plots of each simulation run and for the generated abstracts in the LLM initialisation condition.

Finally, at the end of each run, the current values of the input parameters, the outcome variables and other relevant metadata are appended to a long format master dataframe for analysis. The columns of this dataframe are as follows:
1) the name of the outcome variable
2) the value of the outcome variable
3) name of the simulated dataset,
4) the condition,
5) the values of the independent variables (NaN for the control conditions):
   (a) number of abstracts
   (b) length of the abstracts
   (c) temperature settings of the llm (i.e., diversity)
6) timestamp
7) the run number

This yields a data-frame containing one observation for each combination of dataset (n=26), condition (n=3), independent variables and their levels (n=3×4×4×5×5 = 1200), and run (n=1), thus with $26 \times 3 \times 1200 \times 1 = 93600$ rows, and the 12 columns enumerated above.

References

Bachmann, Fynn et al. (2025). “Adaptive political surveys and GPT-4: Tackling the cold start problem with simulated user interactions”. In: *PLoS One* 20.5, e0322690.

Bastian, Hilda, Paul Glasziou, and Iain Chalmers (2010). “Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?” In: *PLoS medicine* 7.9, e1000326.

Bates, Douglas et al. (2015). “Fitting linear mixed-effects models using lme4”. In: *Journal of statistical software* 67, pp. 1–48.

Bruin, Jonathan de et al. (2025). “ASReview LAB v2: Open-Source Text Screening with Multiple Agents and Oracles”. In: *Available at SSRN 5136987*.

Chalmers, Iain, Larry V Hedges, and Harris Cooper (2002). “A brief history of research synthesis”. In: *Evaluation & the health professions* 25.1, pp. 12–37.

De Bruin, Jonathan et al. (2023). “SYNERGY-Open machine learning dataset on study selection in systematic reviews”. In: *Version V1*.

Dziak, John J and Kimberly L Henry (2017). “Two-part predictors in regression models”. In: *Multivariate behavioral research* 52.5, pp. 551–561.

Ferdinands, Gerbrich et al. (2023). “Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records”. In: *Systematic Reviews* 12.1, p. 100.

Hox, Joop, Mirjam Moerbeek, and Rens Van de Schoot (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Khattab, Omar et al. (2023). “Dspy: Compiling declarative language model calls into self-improving pipelines”. In: *arXiv preprint arXiv:2310.03714*.

Nussbaumer-Streit, Barbara et al. (2021). “Resource use during systematic review production varies widely: a scoping review”. In: *Journal of clinical epidemiology* 139, pp. 287–296.

Panda, Deepak Kumar and Sanjog Ray (2022). “Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review”. In: *Journal of Intelligent Information Systems* 59.2, pp. 341–366.

Schoot, Rens van de et al. (2025). “The Hunt for the Last Relevant Paper: Blending the best of humans and AI”. In: *European Journal of Psychotraumatology* 16.1, p. 2546214.

Teijema, Jelle Jasper et al. (2025). “Large-scale simulation study of active learning models for systematic reviews”. In: *International Journal of Data Science and Analytics*, pp. 1–22.

Tricco, Andrea C et al. (2020). “Rapid review methods more challenging during COVID-19: commentary with a focus on 8 knowledge synthesis steps”. In: *Journal of clinical epidemiology* 126, pp. 177–183.

Van De Schoot, Rens et al. (2021). “An open source machine learning framework for efficient and transparent systematic reviews”. In: *Nature machine intelligence* 3.2, pp. 125–133.

Table 2: Datasets overview

Nr	Dataset	Topic(s)	Records	Included	%
1	Appenzeller-Herzog_2019	Medicine	2873	26	0.9
2	Bos_2018	Medicine	4878	10	0.2
3	Brouwer_2019	Psychology, Medicine	38114	62	0.2
4	Chou_2003	Medicine	1908	15	0.8
5	Donners_2021	Medicine	258	15	5.8
6	Hall_2012	Computer science	8793	104	1.2
7	Leenaars_2019	Psychology, Chemistry, Medicine	5812	17	0.3
8	Leenaars_2020	Medicine	7216	583	8.1
9	Meijboom_2021	Medicine	882	37	4.2
10	Menon_2022	Medicine	975	74	7.6
11	Moran_2021	Biology, Medicine	5214	111	2.1
12	Muthu_2021	Medicine	2719	336	12.4
13	Nelson_2002	Medicine	366	80	21.9
14	Oud_2018	Psychology, Medicine	952	20	2.1
15	Radjenovic_2013	Computer science	5935	48	0.8
16	Sep_2021	Psychology	271	40	14.8
17	Smid_2020	Computer science, Mathematics	2627	27	1.0
18	van_de_Schoot_2018	Psychology, Medicine	4544	38	0.8
19	van_der_Valk_2021	Medicine, Psychology	725	89	12.3
20	van_der_Waal_2022	Medicine	1970	33	1.7
21	van_Dis_2020	Psychology, Medicine	9128	72	0.8
22	Walker_2018	Biology, Medicine	48375	762	1.6
23	Wassenaar_2017	Medicine, Biology, Chemistry	7668	111	1.4
24	Wolters_2018	Medicine	4280	19	0.4

Table 3: Please note that two of the datasets included in the original SYNERGY dataset were excluded entirely due to data quality issues: Chou (2003) and Jeyaraman (2020). For one dataset (Moran, 2021), an updated version was used due to data quality issues in the original version.