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 22, 2025
Word count: 2448

1 Introduction

1.1 Background and Motivation

1.1.1 The framework: AI-assisted systematic reviews

Researchers and practitioners are continually challenged to base their decisions on the latest scientific evidence. To that end, systematic reviews and meta-analyses were developed as rigorous methods of summarizing scientific literature (Chalmers, Hedges, and Cooper 2002; Bastian, Glasziou, and Chalmers 2010). However, systematically reviewing large bodies of literature can be time-consuming, which limits the practical applicability of systematic reviews (Tricco et al. 2020; Nussbaumer-Streit et al. 2021).

Fortunately, recent advances in machine learning have produced tools that enable the systematic screening of scientific literature while greatly reducing the need for manual screening (Van De Schoot et al. 2021). Specifically, active learning models ask users to screen titles and abstracts of papers one by one. Based on the user's decision, the model reassess the probability that each of the remaining papers is relevant, assigns them a 'relevance score' and orders them accordingly. In other words, these models continually reshuffle the papers retrieved from a scientific literature search based on the user's decisions. 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 problem: 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 relevant paper, the model must first have an idea of what constitutes relevance. 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. The cold start problem is particularly problematic when the percentage of relevant papers returned by a systematic search is low, as screening at random until the first relevant abstract is found may take a long time.

1.1.3 The proposed solution: large language models

With the recent advent of large language models (LLMs), a new possible solution to the cold start problem has emerged (Bachmann et al. 2025; Bayer and Reuter 2024). Instead of screening papers at random until relevant and irrelevant abstracts are found, LLMs can generate synthetic examples of both based on the systematic review's inclusion and exclusion criteria. The use of LLMs may add information beyond what is contained in the inclusion and exclusion criteria by generating examples that are more similar to actual abstracts of relevant papers, than the inclusion and exclusion criteria themselves.

In summary, the 'cold start' problem may be overcome and the performance of AI-assisted screening improved by providing the AI-model with examples of relevant abstracts generated by an LLM based on the systematic review's inclusion and exclusion criteria.

1.2 Statistical Framework

1.2.1 Data generating process 1: AI-assisted screening

The starting performance of AI-assisted systematic reviews can be evaluated by simply counting the number of relevant papers successfully retrieved from a given number of screened papers (i.e., X out of n, with $n \leq 100$ trials). In other words, we may discount the fact that the true data generating process (DGP) is a dynamic sequence, and consider starting performance as a discrete proportion of successes from independent trials. This simplifying assumption allows for the application of generalized linear models (GLMs), specifically the binomial model (Nelder and Wedderburn 1972).

The main issue with applying a binomial model to a dynamic sequence is that, in contrast to assumed independent Bernoulli trials, a successful retrieval increases the probability of subsequent successes, and vice versas[1]. This reinforcement mechanism leads to dependence between trials which increases the variance in the number of successes compared to what would be expected under a binomial model, which manifests as overdispersion. One way of dealing with overdispersion is to use a beta-binomial model (Skellam 1948; Harrison 2015; Kim and Lee 2017). The beta-binomial is generalized linear mixed model (GLMM) in which the probability of success is a random variable that follows a beta distribution (Stroup, Ptukhina, and Garai 2024). Formally, the beta-binomial model can be defined as follows:

$$X_i|p_i \sim \text{Binomial}(n_i, p_i) \quad \text{with } p_i \sim \text{Beta}(\alpha_i, \beta_i) \tag{1}$$

For interpretability, we reparamterize α_i and β_i in terms of the beta mean, μ_i, and precision, ϕ_i:

$$p_i \sim \text{Beta}(\mu_i\phi_i, (1-\mu_i)\phi_i) \quad \text{with } \mu_i = \frac{\alpha_i}{\alpha_i + \beta_i} \quad \text{and} \phi_i = \alpha_i + \beta_i \tag{2}$$

This reparametrisation further clarifies how modelling can work using the beta-binomial model (Stroup, Ptukhina, and Garai 2024). For example, the mean parameter μ_i can be modelled using a GLM, while the precision parameter ϕ_i can be used to either simply control for overdispersion (i.e., a single precision parameter for all observations) or to model heterogeneity in overdispersion (i.e., different precision parameters for different groups of observations).

1.2.2 Data generating process 2: random sampling due to a cold start

In reality, two DGP are at play in AI-assisted systematic reviews. The first is AI-assisted screening, which takes place once the AI has been trained using relevant and irrelevant titles and abstracts. However, in the cold-start condition, screening must be done at random until at least one relevant and one irrelevant paper have been identified. This may result in runs where no relevant papers are found. These 'structural zeros' may be considered to come from a different DGP. One way to address this issue is to use zero-inflation models (Lambert 1992; Wagner, Riggs, and Mikulich-Gilbertson 2015).

[1] The DGP of AI-assisted screening can be seen as analogous to the Pólya urn problem, in which each time a ball of a certain colour is drawn, an additional ball of the same colour is added to the urn ('the rich get richer'; Eggenberger and Polya 1923; Kotz, Mahmoud, and Robert 2000). Moreover, the beta-binomial model can be derived from Pólya's urn model (Helfand 2013).

More specifically, we are dealing with a hurdle model, in which the first part of the model (the hurdle) models whether or not any relevant papers are found, while the second part of the model models the number of relevant papers found given that at least one relevant paper has been found (Mullahy 1986). Formally, this can be defined as follows:

1.2.3 Multilevel structure: clustering by dataset

Finally, starting performance may vary systematically between datasets due to differences in topic, percentage of relevant papers return by search, and abstract style, for example (De Bruin et al. 2023). One way of dealing with differences between datasets is to add dataset as a covariate to the model. However, since we are not interested in the effect of specific datasets per se, it may make more sense to treat dataset-specific variations as the results of a random variable. In other words, we may treat dataset as a random effect in a generalized linear mixed model (GLMM) (see Chapter 1 of Stroup, Ptukhina, and Garai 2024 for when GLMMs are appriopriate).

1.3 Objectives

This study aims to investigate whether the starting performance of AI-assisted systematic reviews can be improved by initialising the AI-model with synthetic abstracts generated by an LLM, based on a systematic review's inclusion and exclusion criteria. This was done by simulating the screening of previously conducted systematic reviews.

2 Methodology

2.1 Conditions

2.1.1 Experimental condition: synthetic-abstracts

In the LLM condition, the AI-model is provided with at least one set of titles and abstracts comprising a 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 *(100, 500, 900 words)*,

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

To instruct the LLM, a DSPY module (i.e., a module for dynamic prompt generation within Python) 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 DSPY module then generates the desired number of abstracts using OpenAI's gpt-4o-mini model. For the exact code, see: (Ommeren 2025).

2.1.2 Control conditions

There are three control conditions. The first control condition is the *cold-start* condition, in which no papers are provided prior to the start of screening. In this condition, papers must therefore be screened at random until at least one relevant and one irrelevant paper are found. This condition serves as a baseline against which to compare the effect of LLM initialisation on starting performance.

The second control condition is the *true-example* condition, in which one relevant paper is sampled from the set of actually relevant papers in the systematic review (with replacement between runs). This condition serves to compare the use of LLM-generated examples to the use of actual examples of relevant papers.

The third control condition is the *inclusion-criteria* condition, also known as the 'no-LLM' condition. It is is essentially the same as the experimental condition, except that the inclusion and exclusion criteria of the systematic review function directly as an example of an abstract of a relevant paper with which to initialise the LLM, rather than using the LLM to generate an example. This condition serves to investigate whether the LLM adds any value beyond the direct use of the inclusion and exclusion criteria.

2.2 Outcome variable

2.2.1 Starting performance

Starting performance was operationalized as the number of relevant records found within the first 100 records screened X/n, where $n <= 100$ (the number of papers screened may be less than 100 if all relevant records are found before reaching 100 screened papers). The first 100 is based on the idea that approximately 100 papers can be screened in an hour (Nussbaumer-Streit et al. 2021).

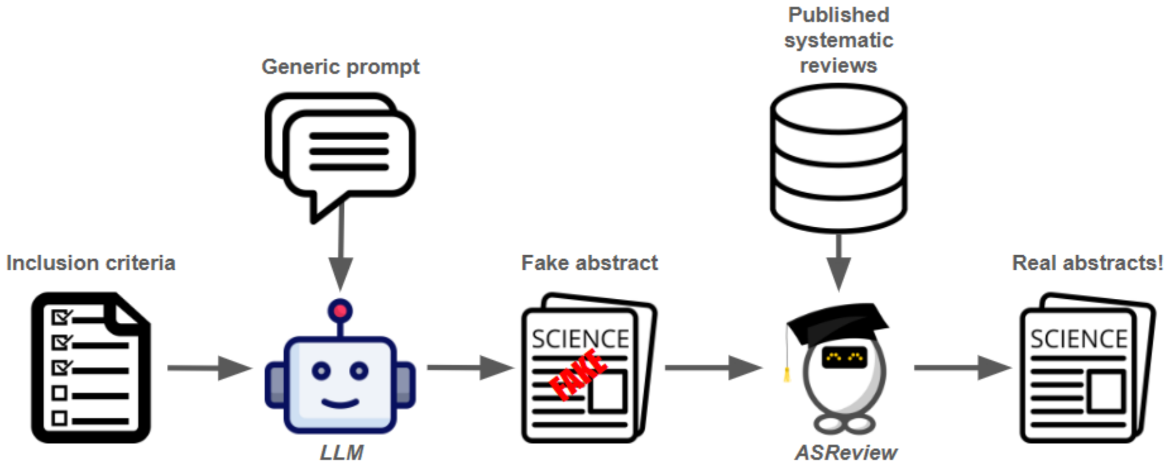

Figure 1: Simulation pipeline (for more detailed schematic of workings within ASReview see (J. d. Bruin et al. 2025))

2.3 Simulation set-up

The effect of using LLM-generated examples to initialize the AI-model on starting performance compared to starting performance in the three control conditions was investigated by simulating the abstract-title based screening of the SYNERGY datasets using ASReview's AI-model (Van De Schoot et al. 2021; De Bruin et al. 2023). Futhermore, the effect of the number of abstracts, length of abstracts and temperature setting of the LLM on starting performance was investigated in the LLM condition using a factorial design (i.e., $3 \times 3 \times 3 = 27$ combinations) (Morris, White, and Crowther 2019). This was repeated 5 times for a total of 135 runs per condition.

ASReview is an open-source software package for AI-assisted systematic reviews (Van De Schoot et al. 2021). To simulate the screening process, ASReview was accessed via its Python API. For all simulation runs, AI-model was set to the recommended U4 configuration in ASReview (see J. d. Bruin et al. 2025 for U4 configuration)[2].

The SYNERGY datasets comprise 24 sets of previously screened and labelled papers. Notably, this provides access to the decisions made by screeners for each paper 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 (Ommeren 2025). See appendix B for a detailed description of how the simulation results were exported. Note that to ensure that the simulation results accurately reflect the screening process, padding was applied to the simulation results (see appendix C for full description).

2.4 Analysis

2.4.1 Analysis

Starting performance was operationalized as the number of relevant records found within the first 100 records screened and was treated as discrete proportion data, X/n, where $n \leq 100$. Prior to the analysis, the predictors in the experimental condition (number of abstracts, abstract length, and temperature) were transformed using a dummy-variable approach (Dziak and Henry 2017), and mean-centred for interpretability. Next, a historgram of starting performance was plotted split by dataset and coloured by condition. Furthermore, the number of unique outcome values (X/n) per dataset and condition was inspected to assess possible convergence problems for more complex GLMMs.

Modelling was done in successive stages, starting with a simple binomial model and iteratively adding complexity before arriving at the final model (Stroup, Ptukhina, and Garai 2024). Only condition was included during model building, as this was the main effect of interest. Each model was fit using the *glmmTMB* R package (Brooks et al. 2017), using maximum likelhood estimation (MLE). At each stage, model fit was evaluated using goodness-of-fit statistics (AIC, BIC), likelihood ratio tests (LRTs) where appriopriate, and by inspecting the randomized quantile residuals using the *DHARMa* R package (Hartig 2016).

[2] Note that technically two configurations are sequentially used: first papers at retrieved at random until at least one relevant and one irrelevant paper have been found and the AI-model can be trained; followed by the U4 configuration.

The *glmmTMB* package does not offer restricted maximum likelihood estimation (REML), but rather uses MLE using the Laplace approximation (Maestrini, Hui, and Welsh 2024). For the final model (at the time of writing a simplified version thereof), a simulation study was performed to investigate the bias and variance in the coefficients, under the assumption that the model was correct, using MLE and REML (Morris, White, and Crowther 2019). No bias was found in the estimated coefficients but considerable variance was found in β_0. Consequently, while the exact estimate of starting performance in the reference condition (i.e., experimental LLM conditin) should be treated with caution, the main interest of this study – the *difference* in starting performance across conditions – does not require such caution. Empirical bayes methods are possible using *glmmTMB* too.

The final model was a beta-binomial mixed model with a logit-link, random intercepts for dataset, one overdispersion parameter for all observations, and one zero-inflation to account for runs in which no relevant papers were found. Condition-specific overdispersion parameters were considered, but led to convergence problems, possibly due to the limited number of unique outcome values per dataset and condition. The inter-class correlation (ICC) was computed for the null-version of the final model (Nakagawa, Johnson, and Schielzeth 2017). Covariates were then added in an exploratory bottom-up modelling approach (Hox, Moerbeek, and Van de Schoot 2017). The analysis was performed in R version 1.1.36.

3 Results (note: preliminary!)

3.1 Descriptives

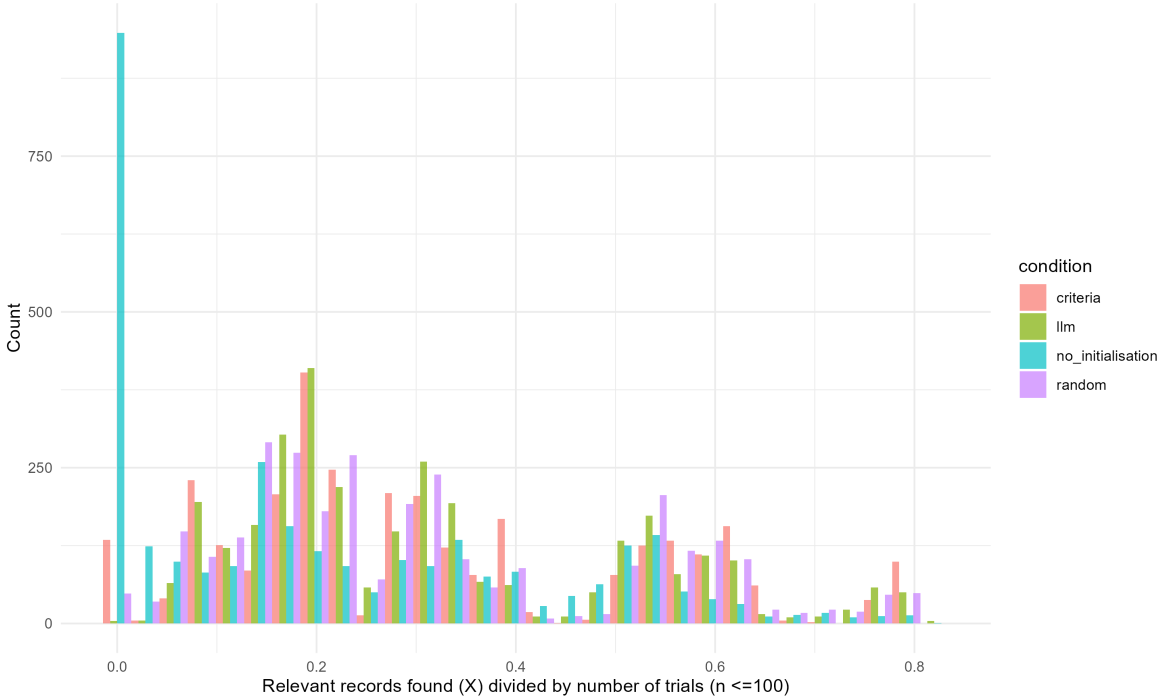

Figure 2: Starting performance: the number of relevant records found divided by the number of trials

The frequency of starting performance outcomes across all datasets and coloured by condition is shown in figure 2 (for the dataset-split histograms see Appendix ...). A few notable patterns emerge from these historgams.

The first is that while the aggregate histogram shows strong multi-modality and overdispersion, the dataset-specific histograms generally do not show these patterns. This supports the use of a random effect for dataset to account for between-dataset heterogeneity.

That said, some dataset-condition combinations continue to show multi-modality. This is particurly true for the cold start condition but also for some others (e.g. the true-example condition in Walker 2018). The remaining multi-modality may be due to subgroups of runs with different characteristics (e.g., different initializations or random seeds) that lead to different performance levels. This supports the use of modelling succes probability as a random variable with a beta distribution (i.e., the beta-binomial model).

The last notable pattern is that the cold-start condition shows a peak at zero, indicating that in some runs no relevant records were found within the first 100 screened records. This supports the use of a zero-inflation component to account for these 'structural zeros'.

Finally, some datasets show limited variability in starting performance: the Donners 2021 and Leenaars 2019 dataset both had two or less unique runs in 3/4 conditions (the cold start conditions being the exception in both cases). Furthermore, the inclusion-criteria condition resulted in 2 or less unique runs for Radjenovic 2013, Smid 2020, Wolters 2018 and van der Waal 2022.

3.2 Main results

4 Conclusion

The results of this simulation study provide a proof of concept that providing LLM-generated abstracts of relevant papers to active learning models for systematic reviews can improve starting performance compared to no initialisation (i.e., a cold-start). However, exploratory analyses suggest that the exact instructions given to the LLM (i.e., prompt engineering) does not affect starting performance. In line with these results, providing the AI-model with LLM-generated abstracts did not improve starting performance over and above merely directly providing the inclusion and exclusion criteria. It thus seems that LLMs simply repackage existing knowledge rather than adding new knowledge.

The key lesson is thus that when it comes to ALM for systematic reviews, something is better than nothing when it comes to initialisation. We thus recommend that researchers and practitioners consider using LLM-generated examples 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).

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.

Table 1: Effect of initialisation condition on starting performance

	(1) Binom	(2) BB	(3) BB+ZI	(4) BB+ZI
Fixed effects (log-odds)				
Criteria (LLM)	0.002	0.028$^{\cdot}$	-0.054***	-0.054***
	(0.006)	(0.015)	(0.011)	(0.011)
No initialisation	-0.678***	-0.944***	-0.487***	-0.487***
	(0.006)	(0.017)	(0.012)	(0.012)
Random initialisation	-0.019**	-0.018	-0.071***	-0.071***
	(0.006)	(0.015)	(0.011)	(0.011)
Intercept	-0.925***	-0.882***	-0.826***	-0.826***
	(0.200)	(0.193)	(0.187)	(0.187)
Random effects and dispersion				
SD(dataset intercept)	0.958	0.926	0.895	0.895
Beta-binomial dispersion (ϕ)	–	15.6	41.1	41.1
Zero-inflation intercept	–	–	-2.497***	-2.497***
			(0.035)	(0.035)
Model fit				
Log-likelihood	-62 302.7	-45 409.4	-42 890.3	-42 890.3
AIC	124 615.3	90 830.8	85 794.6	85 794.6
Marginal R^2	0.082	0.142	0.005	0.005
Conditional R^2	0.968	0.862	0.118	0.118
Observations	12 420	12 420	12 420	12 420
Datasets	23	23	23	23

Notes: All models use a logit link and model the number of relevant records found out of the total number screened. Model (1) is a binomial GLMM estimated via `lme4`. Models (2)–(4) are beta-binomial GLMMs estimated via `glmmTMB`. Dispersion (ϕ) is reported only for beta-binomial models. Zero-inflation is intercept-only. Marginal R^2 reflects variance explained by fixed effects; conditional R^2 reflects variance explained by fixed and random effects. $^{\cdot}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

5 Discussion

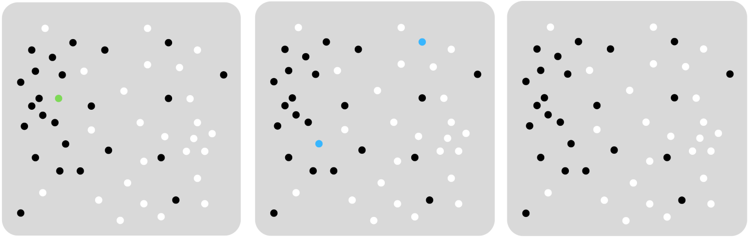

Figure 3: Ideal starting point for systematic reviews using active learning

5.0.1 Simulation design

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.

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.

5.0.2 Statistical Analysis

AI-assistant screening is actually sequence dependent. This is ignored by simply the summed successes and number of retrievals. Future work may consider modelling the sequence directly.

The response variable makes sense to be viewed as a proportion and not a count because we know the true number of total relevant records. Future work modelling the retrieval of actual retrievals may more aptly consider it counts instead (e.g., Poisson).

GLMM assume that the random effects are normally distributed (i.e., that the difference in starting performance between datasets follows a normal distribution). However, based on the results of a previous simulation study comparing the normalized loss performance of AI-assistant screening for different datasets, this seems unlikely (Teijema et al. 2025). Future work may consider specifying a different random-effects distribution using hierarichal generalized linear models (HGLMs) or Bayesian methods.

5.1 Limitations

1. only one LLM used (gpt-4o-mini). Future work may consider other LLMs (e.g., open source LLMs).

2. 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.

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

4. The multilevel model assumption of heterogeneity between conditions may not hold.

5. The effect of the length abstract may be small in part because this parameter had limited effect on the generated abstracts. However, since early analyses showed that the effect of the other LLM-specific parameters was limited (i.e., for temperature and number of abstracts), it is unlikely that the effect of length of abstract would be large either.

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.

C Padding

It is worth noting 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.

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.

Bayer, Markus and Christian Reuter (2024). "Activellm: Large language model-based active learning for textual few-shot scenarios". In: *arXiv preprint arXiv:2405.10808*.

Brooks, Mollie E. et al. (2017). "glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling". In: *The R Journal* 9 (2). https://doi.org/10.32614/RJ-2017-066, pp. 378–400. ISSN: 2073-4859. DOI: 10.32614/RJ-2017-066.

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.

Eggenberger, F and G Polya (1923). "Uber Die Statistik Verketter Vorgage. Zeit". In: *Angew. Math. Mech* 1, pp. 279–289.

Harrison, Xavier A (2015). "A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution". In: *PeerJ* 3, e1114.

Hartig, Florian (2016). "DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models". In: *CRAN: Contributed Packages*.

Helfand, NORA (2013). "Polya's Urn and the Beta-Bernoulli process". In: *University of Chicago REU*.

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*.

Kim, Jongphil and Ji-Hyun Lee (2017). “The validation of a beta-binomial model for overdispersed binomial data”. In: *Communications in Statistics-Simulation and Computation* 46.2, pp. 807–814.

Kotz, Samuel, Hosam Mahmoud, and Philippe Robert (2000). “On generalized Pólya urn models”. In: *Statistics & Probability Letters* 49.2, pp. 163–173.

Lambert, Diane (1992). “Zero-inflated Poisson regression, with an application to defects in manufacturing”. In: *Technometrics* 34.1, pp. 1–14.

Maestrini, Luca, Francis KC Hui, and Alan H Welsh (2024). “Restricted maximum likelihood estimation in generalized linear mixed models”. In: *arXiv preprint arXiv:2402.12719*.

Morris, Tim P, Ian R White, and Michael J Crowther (2019). “Using simulation studies to evaluate statistical methods”. In: *Statistics in medicine* 38.11, pp. 2074–2102.

Mullahy, John (1986). “Specification and testing of some modified count data models”. In: *Journal of econometrics* 33.3, pp. 341–365.

Nakagawa, Shinichi, Paul CD Johnson, and Holger Schielzeth (2017). “The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded”. In: *Journal of the Royal Society Interface* 14.134, p. 20170213.

Nelder, John Ashworth and Robert WM Wedderburn (1972). “Generalized linear models”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3, pp. 370–384.

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.

Ommeren, Timo Q. van (2025). *JUMP STARTING EVIDENCE SYNTHESIS*. Version 0.1.0. URL: `https://github.com/timovanommeren/jump_starting_evidence_synthesis`.

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.

Skellam, John Gordon (1948). “A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10.2, pp. 257–261.

Stroup, Walter W, Marina Ptukhina, and Julie Garai (2024). *Generalized linear mixed models: modern concepts, methods and applications*. Chapman and Hall/CRC.

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.

Wagner, Brandie, Paula Riggs, and Susan Mikulich-Gilbertson (2015). “The importance of distribution-choice in modeling substance use data: a comparison of negative binomial, beta binomial, and zero-inflated distributions”. In: *The American journal of drug and alcohol abuse* 41.6, pp. 489–497.

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.