ADVISE: Accelerating the Creation of Evidence Syntheses for Global Development using Natural Language Processing-supported Human-AI Collaboration

Kristen M. Edwards*
Massachusetts Institute of Technology
Cambridge, MA
kme@mit.edu

Binyang Song*
Massachusetts Institute of Technology
Cambridge, MA
binyangs@mit.edu

Jaron Porciello
University of Notre Dame
South Bend, IN
jaron.porciello@nd.edu

Mark Engelbert
International Initiative for
Impact Evaluation, Inc.
mengelbert@3ieimpact.org

Carolyn Huang
International Initiative for
Impact Evaluation, Inc.
chuang@3ieimpact.org

Faez Ahmed
Massachusetts Institute of Technology
Cambridge, MA
faez@mit.edu

When designing evidence-based policies and programs, decision-makers must distill key information from a vast and rapidly growing literature base. Identifying relevant literature from raw search results is time and resource intensive, and is often done by manual screening. In this study, we develop an AI agent based on a bidirectional encoder representations from transformers (BERT) model and incorporate it into a human team designing an evidence synthesis product for global development. We explore the effectiveness of the human-AI hybrid team in accelerating the evidence synthesis process. To further improve team efficiency, we enhance the human-AI hybrid team through active learning (AL). Specifically, we explore different sampling strategies, including random sampling, least confidence (LC) sampling, and highest priority (HP) sampling, to study their influence on the collaborative screening process. Results show that incorporating the BERT-based AI agent into the human team can reduce the human screening effort by 68.5% compared to the case of no AI assistance and by 16.8% compared to the industry standard case of using a frequency-based language model and support vector machine-based classifier for identifying 80% of all relevant documents. When we apply the HP sampling strategy, the human screening effort can be reduced even more: by 78.3% for identifying 80% of all relevant documents compared to no AI assistance. We apply the AL-enhanced human-AI hybrid teaming workflow in the design process of three evidence gap maps for USAID and find it to be highly effective. These findings demonstrate how AI can accelerate the development of evidence synthesis products and promote timely evidence-based decision making in global development.

1 Introduction

In 2011 the U.S. Agency for International Development (USAID) released *Evaluation Policy*, and in doing so made an ambitious commitment to rigorously evaluating evidence in order to make evidence-based policy [1]. Evidence-based policy refers to public policy that is based on, or informed by, evaluated and objective evidence. To emphasize the importance of evidence-based policy within USAID and the U.S. government, the Foundations for Evidence-based Policymaking Act of 2018 required all agencies under the Act to "affirm the agency's commitment to conducting rigorous, relevant, evaluations and to using evidence from evaluations to inform policy and practice" [2]. It is imperative in part because these policies dictate the expenditure of billions of dollars. For example, in 2017, USAID spent $1.01 billion on foreign agricultural assistance alone [3].

However, evaluating all available evidence has been made burdensome by the current information explosion. In 2018 alone, global research output in science and engineering was 2.6 million articles, which grew at a rate of 4% annually from 2008-2018 [4]. A person's capacity to under-

*These authors contributed equally to this work.

stand all available research is limited. Policy-makers have thus turned to evidence synthesis to understand the growing corpus of research available and make informed decisions. Evidence synthesis refers to the process of compiling information and knowledge from many sources and disciplines to inform decisions [5, 6]. However, creating evidence synthesis products like evidence gap maps (EGMs) requires extensive time and effort from human experts. EGMs, as described in the Related Work section, visualize interventions and their associated outcomes [7], and have been shown to provide incredible value to decision-makers in fields ranging from agriculture to public health [5]. For example, Figure 1 represents a portion of an EGM[1] available from the International Initiative for Impact Evaluation (3ie), which is one of the global leaders in generating EGMs for decision-making. We can see there is a research gap between the interventions of "water access & management" and "improved seeds" and the outcomes regarding "profit". Policymakers can plan future investments and research accordingly.

Our goal is to accelerate the design of EGMs in the global development space and alleviate the burden of information filtering. 3ie's current evidence synthesis process includes significant expert screening of documents and moderate use of machine intelligence, often taking nearly six months to complete [6]. Natural language processing (NLP), a form of artificial intelligence (AI), has long been used for text comprehension. Recently, the rule-based NLP models have attracted some attention and been explored to promote evidence-based decision making in the medical, legal, and global development fields [8, 9, 10]. The work that has successfully done so may be improved upon by incorporating the latest transformer- and transfer learning-based NLP models.

1.1 Contributions

Title and abstract (TA) screening is one of the most time-consuming steps in the EGM design process, typically involving comprehending the titles and abstracts of tens or hundreds of thousands of papers for screening. Through collaborating with 3ie, we make the following contributions:

1. We develop an AI agent using bidirectional encoder representations from transformers (BERT) to accelerate the TA screening portion of the EGM design process, and incorporate it into a human team to explore the efficiency gains made through human-AI teaming. With the best combination, our AI agent reduces human effort by 78.3% when identifying 80% of all eligible documents, as compared to no AI assistance.
2. We compare our BERT-based AI agent against the industry standard model, and find that the BERT-based model outperforms the industry standard in both model performance (12% average increase in accuracy for the three EGMs) and saved effort (17% reduction in required effort in the simulated case, and a 46% average reduction in effort for the three deployed EGMs).
3. We identify the optimal training size (5,000 documents) for both model performance and saved effort.
4. We compare active learning (AL) strategies and find that by using high priority (HP) or least confidence (LC) we can decrease human effort by an additional 30% (compared to BERT with no AL) for identifying 80% of all included documents.
5. We support the development of three EGMs: Agriculture [11], Nutrition [12], and Resilience [13] for use by the U.S. Agency for International Development.

2 Related Work

In the following sections we describe related work in the fields of EGMs, natural language processing, and AL, and explore their intersections in the context of human-AI teams.

2.1 Evidence Gap Maps

EGMs are one form of evidence synthesis - the process of compiling information and knowledge from many sources and disciplines to inform decisions [5, 6]. Evidence synthesis provides more reliable information about a topic than a single study by systematically collecting, categorizing, and analyzing a broad range of studies [14]. Evidence synthesis for decision making was largely popularized by the biomedical field, but it provides clear benefits for decision-makers in any field [15, 16, 9]. Thus, evidence synthesis is an incredibly valuable tool for decision-makers in global development seeking to design policies and fund research [5].

3ie has pioneered the use of EGMs, which present a visual overview of completed and ongoing impact evaluations and systematic reviews in a specific sector [17]. 3ie creates these EGMs via the "thematic [collection] of information about impact evaluation and systematic reviews that measure the effects of international development policies and programmes" [7]. The final product is a matrix, organized by "intervention" categories on the vertical axis and "outcome" categories on the horizontal axis. Interventions are the action taken in the study, and outcomes are the result of the action. Each cell of the matrix contains studies that rigorously evaluate the impact of a specific intervention on a specific outcome. An example of this matrix is shown in Figure 1.

3ie sets the global standard for EGMs, and the mapping method has been adapted by organizations including the Campbell Collaboration, the World Bank Independent Evaluation Group, and USAID [17]. Like other forms of evidence synthesis, EGMs begin with an expansive and systematic search of scholarly databases and "grey literature" sources (such as repositories of government documents or websites of think-tanks) to identify potentially relevant studies. EGM teams then screen these search results to identify studies that meet the EGM's criteria for interventions evaluated, outcomes measured, implementation setting, and study design. Once eligible studies are identified, the EGM team extracts information on interventions, outcomes, and other key characteristics of each study to determine its placement in the EGM matrix and to allow for analysis of trends in the

[1] https://developmentevidence.3ieimpact.org/egm/food-systems-and-nutrition-evidence-gap-map

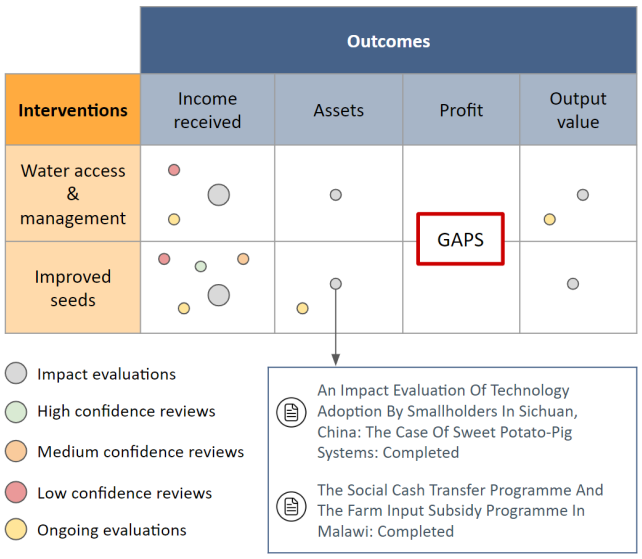

Fig. 1. A representation of a portion of a 3ie EGM showing two interventions and four outcomes. Research gaps exist between the two interventions and the outcome "Profit". Dots of different colors represent different evidence types. Dot sizes indicate how many documents exist in each group.

literature.

3ie uses software called EPPI-Reviewer which aids in the creation of EGMs. While EPPI-Reviewer has some machine learning functions that can accelerate screening [18], most EGM tasks are still performed manually. Thus, each EGM requires significant human effort and expertise, with many EGMs requiring nearly six months to complete [6]. Given that one of the main barriers to evidence use among policymakers is the lack of timely research outputs [19], there is a critical need to reduce the time and effort needed to complete the EGM design and development process.

The high-level steps of designing an EGM are shown in Figure 2. Our work focuses on step three, in which reviewers screen documents for inclusion in an EGM based on their title and abstract. Selected documents will move on to full-text review. We create three transformer-based NLP models that automatically classify documents for inclusion at this step.

2.2 Natural Language Processing in Evidence Synthesis

NLP is a field of machine learning in which computational machines are trained to understand text and spoken language. The earliest language models include frequency-based models like the n-gram model, which were introduced in the 1940s-1950s [20, 21]. During this period, statistical techniques were used for language modeling. These models estimated the probability of a word based on the frequency of its occurrence in a given context.

Researchers have since invented and explored new techniques that utilize more semantic understanding. Semantic understanding aims to bridge the gap between the symbolic representation of language and its actual meaning. It involves extracting and interpreting information at different levels, including word-level semantics, sentence-level semantics, and discourse-level semantics.

Major strides in language modeling include the rule-based approaches which focus on hand-crafted linguistic rules and grammars to process and understand natural language [22, 21]. These systems rely heavily on expert knowledge and manual rule construction. Additionally, in the 1990s machine learning techniques gained popularity in NLP. Hidden Markov Models (HMMs) and Maximum Entropy models were used for tasks like part-of-speech tagging and named entity recognition [23].

Modern NLP models have been largely shaped by the introduction of the transformer in 2017, which allowed text inputs to be fed in parallel and achieved state-of-the-art results over frequency-based language models in many NLP tasks [24, 25, 26, 27]. BERT is among the most well-known transformer-based models and has been extensively explored in NLP tasks such as language translation and question answering [28, 29]. Other such models include GPT and models based off of it [30].

In the medical field, the development of an NLP-based model for automating evidence synthesis, called BioMedICUS, improved the scalability and performance of text analysis and processing of biomedical and clinical reports [8]. The success of NLP in the medical field has led to its use in other fields, with models like LexNLP, which automatically extracts information from legal text [9].

There are several industry-standard NLP tools used to aid human experts when designing evidence synthesis products like EGMs. The most common tools include EPPI-Reviewer, Rayyan, and RobotReviewer [15, 16], and all of these utilize frequency-based language modeling and support vector machine (SVM)-based classifiers as their primary ML model [31, 32, 33].

Furthermore, rule-based NLP models have been explored to promote evidence-based decision making in multiple fields [10]. However, the rule-based models are often case-specific. It requires significant effort to adapt a rule-based model from one EGM to another. Moreover, it is challenging to capture all the subjective criteria used by humans and embed them into the defined rules comprehensively.

We propose that utilizing state of the art transformer based language models will accelerate the creation of EGMs compared to the industry standard frequency-based language model tools. We hypothesize that the increased semantic understanding found in pretrained large language models will lead to better classification performance and lowered human effort when screening documents.

Our work explores BERT-based NLP models as a tool for human-AI teams designing EGMs for global development, which involves more unstructured studies and broader domains than other fields. Recent research that is perhaps most similar to our work is srBERT [34], which explores fine-tuning a BERT model with topic-specific articles in order to accelerate the screening process for a systematic review about "moxibustion for improving cognitive impairment" [34]. Our work, on the other hand, works with much larger and broader datasets in order to create EGMs. Addi-

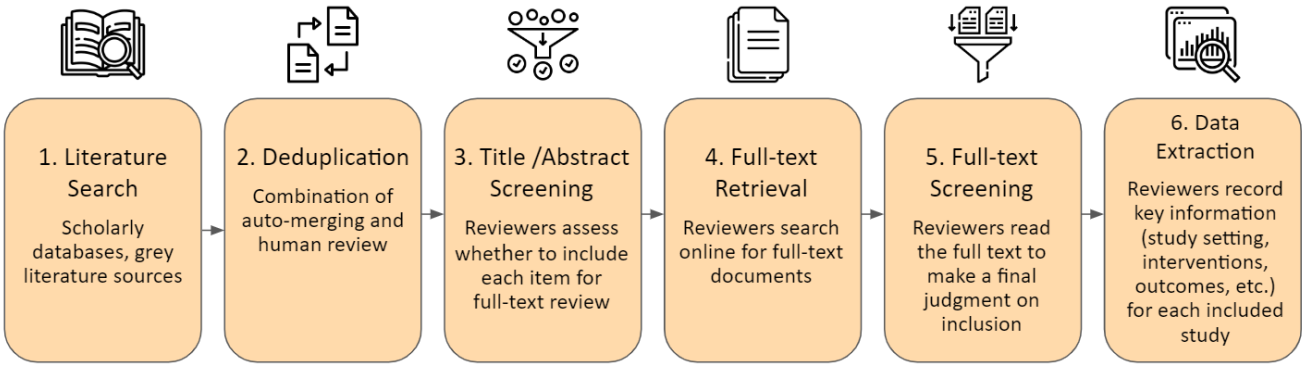

Fig. 2. A high-level view of the current EGM creation process.

tionally, we exhibit the effectiveness of our NLP tool in a real human-AI team and ultimately create three deployed EGMs in Agriculture [11], Nutrition [12], and Resilience [13]. We utilize the experience of creating deployed EGMs to explore the nuances of human-AI teaming in this design process.

2.3 AL and Human-AI Teams

In many AI tasks, obtaining labeled training data is expensive and time-consuming [35]. We are motivated to explore avenues to decrease the size of training data needed by using AL. AL is the concept that an ML algorithm can perform better with less training data if it is allowed to choose the data from which it learns. AL has been applied to deep learning problems such as image classification [36, 37], speech recognition [38], data exfiltration detection [39], and many NLP tasks [40]. There are three main problem setups, or scenarios, in which a learner may be able to ask queries: membership query synthesis [41], stream-based selective sampling [42], and pool-based sampling[43]. The most commonly used pool-based sampling strategies evaluate and rank the entire unlabeled pool in terms of informativeness and then select the best queries [44]. There are also different *query strategies* for choosing which unlabeled instances to query. The most commonly used query strategy is uncertainty sampling [43]. In this strategy, the learner queries the instances for which the learner is least certain how to label [44]. Within the uncertainty sampling category, there are three primary measures that evaluate how uncertain the learner is about each instance: least confidence [45], margin sampling [46], and entropy [20].

During the training, AL tries to optimize the information flow from humans to AI to improve AI performance with less training data. In this study, AI is working collaboratively with humans instead of alone. Accordingly, both the information flow from humans to AI and that from AI to humans are important to the performance of human-AI hybrid teams. Therefore, AL in such a context should consider the bi-directional information flows.

In fact, common barriers stopping human screeners from incorporating AI into their EGM design process include a mismatch in existing workflows, and a steep learning curve [16]. Research shows that human-AI teams using AL for a real life task can lose the human agent's trust if the AI agent makes irrelevant suggestions or predictions during the training process [39]. Therefore, we explore and discuss the effective integration of AI tools into the existing EGM design process. We also use AL-based approaches to maximize the accuracy of the AI classifier, while minimizing the workload put on the human screeners.

3 Methodology

In this work, we utilize a BERT-based NLP model to accelerate the design process of EGMs. We utilize the NLP tool in the workflow of a real human-AI team with members of 3ie. We support the development of three deployed EGMs in the topics of Agriculture, Nutrition, and Resilience.

We analyze how different ML methods and data training sizes affect the human-AI team performance, focusing on the trade-offs between model accuracy and human effort. We compare our fine-tuned BERT model against the industry standard NLP tools. Further, we explore the effect of AL with various query strategies on model accuracy and human effort.

Our work is comprised of two case studies:

1. **Simulated EGM design**: Using a pre-existing, fully labeled dataset we run experiments to determine the most effective classification algorithms (industry standard vs. BERT-based) and AL strategies for EGM creation.
2. **Deployed EGM design**: Utilizing results from the simulated EGM design, our human-AI team creates three EGMs for use by USAID: Agriculture [11], Nutrition [12], and Resilience [13].

These two case studies present different challenges and priorities. In the simulated EGM design, we have the benefit of a fully labeled dataset, which we can practice multiple techniques on. In the deployed EGM design, we are in the real-world situation of creating an EGM from scratch using a human-AI team. Therefore, we only have labels for the documents that we specifically choose to screen.

Further, in the deployed EGM design, we are motivated to design the most comprehensive and informative EGM while efficiently utilizing human resources. Consequently, we want to minimize time that human experts spend screen-

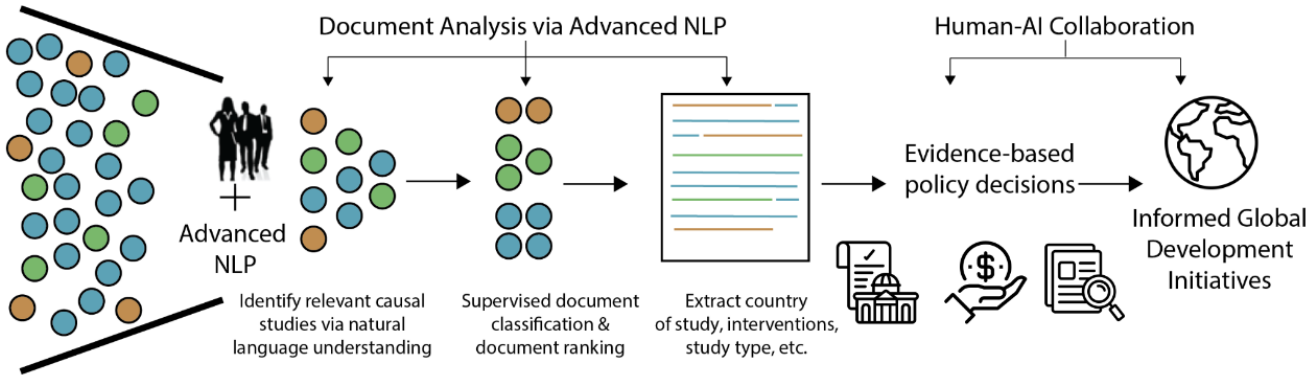

Fig. 3. Proposed utilization of NLP tools in a human-AI team to screen, understand, and classify documents (represented by circles) in order to inform evidence-based policy decisions. Our goal is to accelerate the design process for EGM products in the global development field.

ing irrelevant documents, and screen only the relevant documents. This contrasts the strategy in classical AL to query or screen the documents we are most uncertain of.

3.1 Dataset Description

For the simulated EGM design, our data is provided by 3ie and is derived from manually labeled documents from 3ie's Development Evidence Portal (DEP) [2] [47], an expansive repository of impact evaluations and systematic reviews in global development across a wide range of sectors. We utilize a dataset of 68,539 documents screened for inclusion in 3ie's DEP to develop and evaluate our classification models. Table 1 shows the key attributes of the dataset, such as title, abstract, and inclusion decisions.

Attribute	Description
Title	Title of the paper.
Abstract	Abstract of the paper.
Keywords	Keywords of the paper.
Year	Publication year.
Publication type	Journal, conference proceeding, report, etc.
Source	The source of the paper, e.g., journals or conferences.
Inclusion decision	Whether the paper is included as a relevant study. If not, what is the exclusion criterion.

Table 1. The key attributes in the Development Evidence Portal dataset.

In this study, the title of each paper is integrated into the abstract as a sentence at the beginning. The BERT classification model takes the integrated texts as the input. The label of "included" or "excluded" is derived from the inclusion decision. To train the binary classification model, the "0" class corresponds to the "excluded" papers, and the "1" class comprises the "included" papers. This dataset is highly imbalanced, containing 5,281 included papers and 63,258 excluded papers. The criteria for excluding the papers are also extracted for training the criterion-specific classification models.

For our deployed EGM design, we are actively designing three EGMs. As indicated in Figure 2, and per 3ie's EGM workflow, we gather our initial datasets via a literature search through scholarly databases and grey literature sources. For the three EGMs, their initial dataset sizes are as follows: Agriculture 221k, Nutrition 117k, and Resilience 60k.

3.2 Data Pre-processing

The raw documents are pre-processed to remove noise. Two types of noise are removed in this step. The first is non-English texts. A portion of the papers provide titles and abstracts in multiple languages. Since our models only take texts in English as input, the sentences in languages other than English are noise to the models and should be removed. The second type comprises English text content that is irrelevant to the scope of the document, such as a copyright statement. The pre-processing consists of five steps. (1) Each document is parsed into sentences. (2) A language detection model is used to identify sentences written in non-English languages. (3) We manually label the sentences from 500 documents with the "relevant" and "irrelevant" labels. (4) A BERT classification model is trained on the labeled data to predict the labels of the other sentences. The accuracy of the model is higher than 0.99. (5) Once the irrelevant sentences are removed, the remaining relevant sentences are integrated back into the original documents.

[2] https://developmentevidence.3ieimpact.org/

3.3 Priority Score

In this study, the AI agent is operationalized by a BERT binary classification model, which employs a 12-layer pre-trained uncased BERT embedding module with a hidden size of 768. The BERT embedding module is followed by a dropout layer with a drop rate = 0.1 and a linear layer that outputs a 2-dimensional (2D) vector as the final classification prediction. As described above, the AI agent needs to sample or prioritize the unlabeled papers according to the probabilities of being relevant, as predicted by the classification model. This probability is named the "*priority score*" (PS) in definition 1.

Definition 1. *Priority score* is the probability that a paper is a relevant paper predicted by the AI agent, which is calculated by $PS(p) = softmax(Pred(p))[1]$, where $Pred(p)$ is the prediction output from the classification model for a paper p, which is a 2D vector. The "1" in the equation indicates that PS(p) is the probability of the paper being classified to the "1" class. Following this definition, higher screening priority scores are assigned to the papers with higher predicted probabilities of being relevant.

3.4 Sampling Strategies

According to the predicted PSs, we apply three different query strategies to sample papers from the unscreened subset, which will be labeled and added to the training set in the next iteration.

1. **Least confidence**: The *least confidence (LC)* query strategy is one of the commonly used strategies for AL, which samples papers that the model is least certain how to classify [44], as shown by

$$x^*_{LC} = \arg\max_{x \in X} U(x) \tag{1}$$

The classification uncertainty $U(p)$ of the paper p is derived from the classification model output $Pred(p)$ through

$$U(p) = 1 - max(softmax((Pred(p))) \tag{2}$$

2. **Highest priority**: The *highest priority (HP)* query strategy samples papers with the highest PSs, given by

$$x^*_{HP} = \arg\max_{x \in X} PS(x) \tag{3}$$

This query strategy is adapted from the uncertainty sampling strategies [43]. For evidence synthesis, all relevant papers need to be verified by a human agent, so the papers most likely to be relevant are first sampled.
3. **Random**: The *random* query strategy randomly samples papers from the unlabeled list without using any informativeness measure. In this case, no AL is applied.

3.5 Human-AI Hybrid Team Workflow

We assume the human-AI hybrid team is tasked with screening a set of papers to identify papers satisfying a given scope. The human agents start the TA screening process by specifying the screening criteria. Then, the AI agent randomly samples a subset of papers, which are screened by the human agents as the initial training set. On this basis, our model is trained to learn the screening criteria from the training. With the learned knowledge, the AI agent predicts the PSs of the unscreened papers. According to the predicted PSs, the AI agent needs to check whether the screen-train-predict-sample loop should stop. If not, it employs a certain strategy to sample a set of papers to be screened for the next iteration.

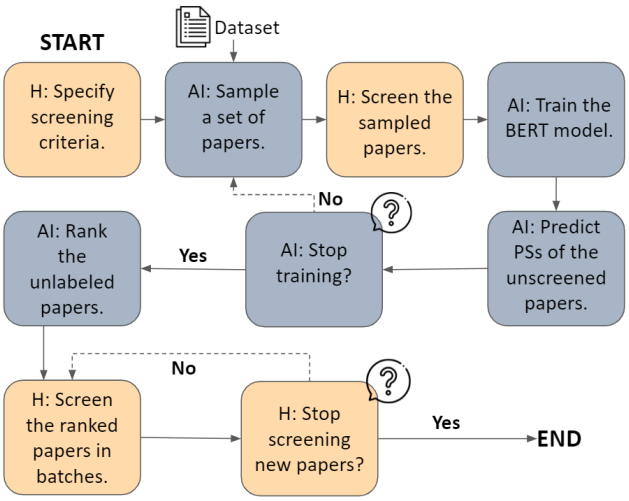

Fig. 4. The workflow of the human-AI hybrid team. H represents human agents.

Once the AI agent decides to stop training the model after a few iterations, it prioritizes all the unscreened papers according to their predicted PSs. Then, the human agents screen the prioritized papers in batches and decide at which batch the screening process should be ended. Since the dataset is imbalanced, a random over-sampling method is applied to the training set to make the numbers of samples from both classes equal. In this study, the AI agent samples a batch of 1,000 papers each time. The batch number is selected because it balances the gain from and the cost of updating the model.

3.6 Evaluation Metrics

In the ML domain, accuracy and F1 score are commonly used to evaluate the performance of classification models. However, these metrics alone are not informative enough to assess the performance and efficiency of the human-AI hybrid teams. In this study, we evaluate the performance of the human-AI hybrid teams in terms of *human effort* in definition 2 needed for achieving an *inclusion rate* in definition 3.

The computational cost reflects team efficiency from another perspective, which is not discussed in this study.

Definition 2. *Human effort* is defined as the ratio (HE) between the number of papers that need to be screened manually ($n_{screened}$) for identifying a given amount of relevant papers and the total number of papers ($n = 68,539$ in the simulated EGM design case study) in the dataset, which is calculated by: $HE = n_{screened}/n$.

Definition 3. *Inclusion rate* is the ratio (IR) between the number of included papers being identified ($n_{identified}$) and the total number of included paper ($n_{included} = 5,281$ in the simulated EGM design case study) in the dataset, calculated by: $IR = n_{identified}/n_{included}$. With limited resources, a higher inclusion rate is preferred.

Given a set of scientific papers and the screening criteria, an efficient human-AI hybrid team should minimize the human effort and computational cost for achieving a satisfying inclusion rate or maximize the inclusion rate with available human effort and computational resources. Additionally, the F1 score of the corresponding classification model in each case is also reported for assessing the performance of the AI agent.

4 Experimental Setup

In this section, we discuss the experimental setups for the two case studies: simulated EGM design and deployed EGM design. The aspects specific to each case study are described in 4.1 and 4.2, while their shared components like baseline techniques and implementation details are described in sections 4.3 and 4.4.

4.1 Simulated EGM Design Experiments

In this case study, we assume the human-AI hybrid team is tasked with screening a set of 68,539 documents to identify documents satisfying a given scope. This dataset is fully labeled, and we can therefore test the efficacy of different training sizes and AL sampling strategies.

4.1.1 Training Size Experiments

For ML model training, a larger training set often benefits model performance but needs more human effort to label the data. In the human-AI hybrid team, the trade-off between the model performance and the required human effort for labeling should be balanced carefully to achieve high team efficiency. We conduct experiments to investigate how the training size affects hybrid team efficiency - both in terms of the model performance (F1 score) and human effort required. Specifically, we start with an initial training set of 1,000 papers; to expand the training set, we randomly sample 1,000 papers, label them, and add them to the training set in each iteration from 1,000 to 6,000. During training, we use 85% of the papers in the training set to train our model and 15% as a validation set. All the other papers compose the testing set.

4.1.2 AL Experiments

When the AI agent samples new papers to be screened, the query strategy used affects the informativeness of the sampled papers, which further influences the ML model performance and hybrid team efficiency. We compare two different sampling strategies, LC and HP, with random sampling through experiments. For each sampling strategy, we start with the same initial training set of 1,000 papers with the random sampling case. After that, we sample 1,000 new papers using the LC or HP strategy to expand the training set in each iteration. We experiment with training sizes ranging from 1,000 to 7,000 for the two sampling strategies.

4.2 Deployed EGM Design Experiments

The human-AI team is tasked with developing three EGMs for deployment: an Agriculture, Nutrition, and Resilience EGM. For each one, the human-AI interaction and, therefore, our experiments are targeted at the title and abstract screening process.

The three EGMs are created independently and following the same process. They start with the following size of datasets: Agriculture 221k, Nutrition 117k, and Resilience 60k. The human-AI workflow used for the title and abstract screening of each EGM is shown in Figure 4 and described above in section 3.5. Once the process is complete for each EGM, the human-AI team has labeled a small subset of each dataset. We utilize these labeled datasets for the following experiments.

4.2.1 Model Performance Experiments

We explore the effectiveness of the industry standard model and our proposed BERT-based model for classifying documents as relevant or irrelevant for each of the three EGMs. Due to the real-world nature of this case study, we only have labels for those documents which we choose to screen. For each of these labeled datasets, we perform an 85% - 15% train-test split and determine the classification accuracy.

4.2.2 Human Effort Experiments

We also compared the trained BERT and industry standard models in terms of human effort to assess hybrid team efficiency. Specifically, we suppose the documents in the test set would be screened in descending order of priority scores predicted by the BERT and industry standard models respectively. Human effort is defined as the percentage of documents that humans need to screen for getting a specific inclusion rate. The hybrid team is more efficient if fewer documents must be screened to obtain the same number of included documents. That is, less human effort is needed.

4.3 Baseline

To answer RQ1 - *how much human effort can be saved when the AI agent is trained on an optimal data size?*, we compare the best case from the experiments with different training sizes with the baseline cases. In the first case, the

human team works alone on the same task without any AI assistance. That is, the human agents randomly screen papers from the dataset. The second case employs an industry standard method: a frequency-based language model using SVM-based classifier, which is developed for retrieving randomized controlled trials and available in the EPPI-Reviewer software [33]. For the second baseline, we also experiment with four different training sizes ranging from 1,000 to 7,000, from which the best model is used as the baseline.

To answer RQ2 - *how much human effort can be further saved by enhancing the hybrid team through AL?*, the best model from the experiments with different training sizes is used as the baseline, where all the sampled papers are randomly selected. We compare the best models from the experiments with the LC sampling strategy and the HP sampling strategy to the baseline model, respectively.

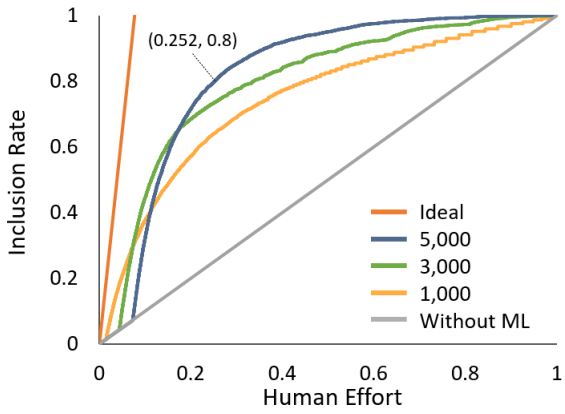

Fig. 5. How inclusion rate varies with human effort when our BERT-based model is trained with different training sizes. A training size of 5,000 performs best, as indicated by reaching an inclusion rate of 0.8 with the lowest human effort.

4.4 Implementation Details

In this study, our models are trained with a learning rate of 1×10^{-5}. There is a warm-up phase at the beginning of the training process, which lasts for one epoch. The experiments were performed on Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz 3.00 GHz, with 18 cores and 256 GB of RAM. Model training and predicting were conducted on Nvidia RTX A5000 GPUs (single GPU per run). Each experiment is repeated five times. When the predicted uncertainties and PSs are needed to sample new papers with the LC and HP strategies, we use the mean values of the predictions from the five runs to improve the repeatability of the results.

5 Results

In the following sections we present the results of our experiments in both case studies: the simulated EGM design, and the deployed EGM design. We compare different ML models, training sizes, and AL sampling strategies and report their effects on model performance and human effort. Further, we go on to discuss the limitations of our work and the future use of human-AI teams in EGM design.

5.1 Simulated EGM Design Results

In this section we display the results comparing the BERT-based model and the industry standard model in the simulated EGM design. We first look at how the two models compare in terms of saving human effort. Then, we incorporate AL strategies and compare both the models and the different AL strategies for both human effort saved and classification performance.

5.1.1 Saved Effort

The performance of the human-AI hybrid team is assessed through inclusion rate (IR) and human effort (HE). Figure 5 shows the variation of IR with HE when our model is trained with different training sizes. The grey "Without ML" line in the figure corresponds to the condition where the human agents work alone without any AI agent. Since the human agents randomly screen papers form the dataset, IR is equal to HE in this case. The orange "Ideal" line close to the y-axis denotes the ideal case, in which each screened document is an included document, and no excluded documents are screened. The slope of this line is 5,281 (the total number of included papers) / 68,539 (the total number of papers in the dataset). The other curves in the figure describe how the IR changes as the human agents invest more screening efforts when the BERT-based AI agent is trained on the datasets with different sizes.

Each curve consists of two parts. The first straight line part indicates the process that the human agents label papers from the dataset to prepare the training set. Since the screened papers are randomly selected, the IR is equal to HE. Once the training set is ready, our model is trained on it to predict the PSs of the unlabeled papers. The curved part following the straight line corresponds to the process during which the human agents screen the unlabeled papers sequentially according to the predicted PSs. Since the unlabeled papers are prioritized for screening, the curves are much steeper in this second portion than in the first, which has the same slope as the "Without ML" line. In the curved portion, the initial slopes are close to the slope of the ideal line, then gradually decrease later on. This trend suggests that the papers with higher PSs are more likely to be identified as included papers than the papers with lower PSs, implying the effectiveness of the AI agent in prioritizing the unlabeled papers for screening.

Since a high-performing human-AI hybrid team can achieve a higher IR with a lower HE, its initial slope should appear closer to the "Ideal" line in 5. As the training size increases, the curve gets steeper, indicating improved model performance. This is in line with the increasing F1 scores shown in Figure 7. However, because a larger training size needs more human labeling effort (i.e., a longer straight line in the first part along the diagonal line), it may also impair the efficiency of the human-AI hybrid team. Given a target IR of 80%, the curves show that the hybrid team gets the

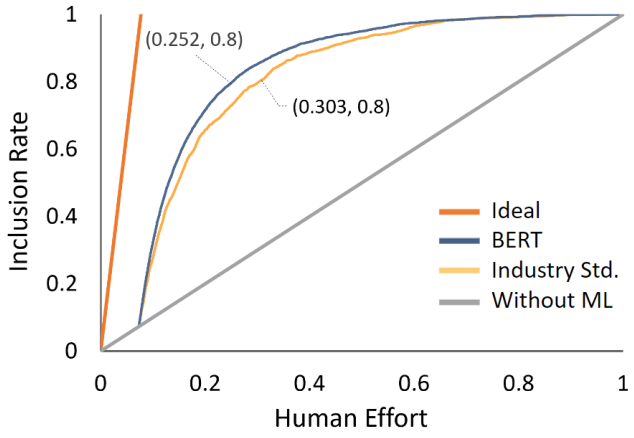

Fig. 6. How inclusion rate varies with human effort for the different ML models: our model (BERT), the industry standard, and "Ideal" and "Without ML" baselines.

highest efficiency when the training size is 5,000. Under this condition, the human agents only need to screen 25.2% of the papers to get an IR of 80%, while they need to screen 80% of the papers to get the same IR in the case without the AI guidance. Therefore, when the BERT-based AI agent is incorporated into the human team, it can save 54.8% human screening effort for getting the IR of 80%.

We also compare the BERT-based model with the model used in the EPPI-Reviewer software in terms of their effectiveness as the AI agent. Similarly, we train EPPI-Reviewer's industry standard model with different training sizes (1,000, 3,000, 5,000, 7,000), among which the training size of 5,000 needs the least human effort for getting the IR of 80%. Figure 6 compares the best BERT-based model (5,000) and the best industry standard model (5,000), suggesting that the BERT-based model enables the human agents to save more screening efforts compared to the industry standard model for getting any IR. Specifically, the human agents can save 5.1% more screening efforts when working with the BERT-based AI agent than working with the industry standard AI agent for getting the IR of 80%. Therefore, our BERT-based model is more effective in acting as the AI agent.

5.1.2 The Effect of AL

In the following section, we discuss how the strategies for sampling new data to expand the training size affect the performance of the AI agent and the efficiency of the human-AI hybrid team for the TA screening task.

AL, Training Size, and Model Performance

Here we report the results of the experiments with different training sizes and different sampling strategies to demonstrate the effect of incorporating the AI agent into the human team, answering RQ1. Following the protocol of the classification problems with the imbalanced dataset, we use the F1 score computed at the default threshold of 0.5 as the classification metric. In these experiments, the sampled papers are randomly selected. The black curve in Figure 7 shows the variation of the F1 score with the training size. As the training size increases, the F1 score improves with diminishing marginal effect, especially when the training size is larger than 5,000.

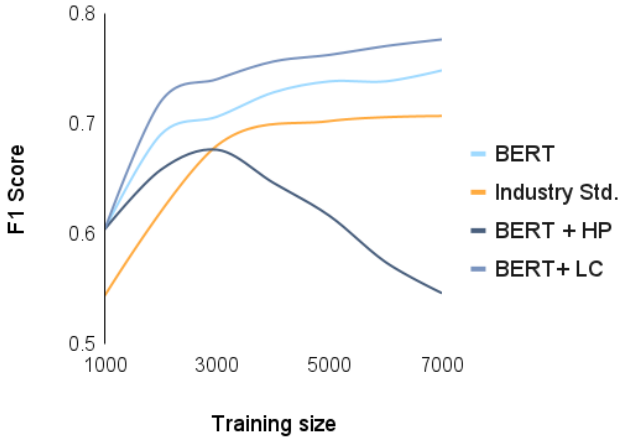

Fig. 7. How model performance, as shown by the F1 score, varies with training size and AL sampling strategy. The bars indicate one standard error. We find that the LC strategy performs the best.

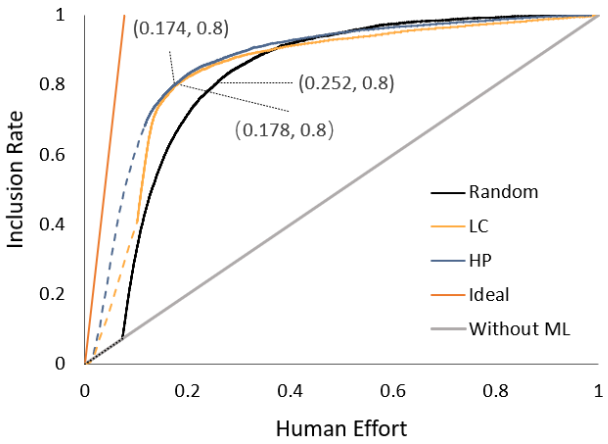

Fig. 8. How different AL sampling strategies affect the human effort and inclusion rate relationship. The dotted line portion of each curve represents the screening-updating-predicting-sampling iterations, while the solid line part corresponds to the process when the human agents screen the prioritized papers.

Similar to the random sampling case, the training size affects the performance of our classification model. As shown in Figure 7, a larger training size improves the F1 score when the LC strategy is applied. If we employ the HP strategy, a moderate training size (e.g., 2,000) benefits the F1 score most, and a larger training set impairs the F1 score when its size surpasses a certain value (e.g., 2,000). Overall, sampling new papers using the LC strategy leads to better classification models than randomly sampling new papers, as

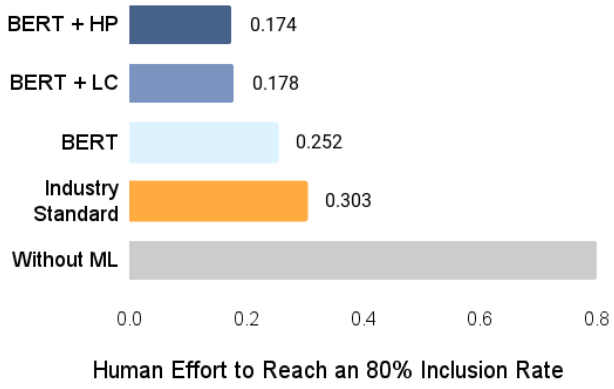

Fig. 9. The human effort required to reach an 80% inclusion rate for various models. Lower human effort is preferred. With no ML, it takes 80% human effort to reach an 80% inclusion rate.

indicated by the higher F1 score; however, the HP sampling strategy results in worse classification models than random sampling, indicated by the lower F1 scores.

AL, Training Size, and Human Effort

The selected AL sampling strategy and the training size also affect the human-AI team efficiency. Under the random sampling condition, a moderate training size can well balance the trade-off between higher model performance and more labeling effort for creating the training data, leading to the highest team efficiency. We observe similar trends for AL. To get an IR of 80%, the human-AI hybrid team achieves the highest team efficiency with a training size of 7,000 when the LC and HP sampling strategies are applied, respectively.

Figure 8 compares the team efficiency among different sampling conditions, including random sampling, LC sampling, and HP sampling. We can see that the efficiency of the human-AI hybrid team is improved substantially with AL. When the LC and HP strategies are applied, the human agents can respectively save 7.4% and 7.8% screening effort for getting the IR of 80%. Specifically, the dotted line portion of each curve represents the screen-update-predict-sample iterations (i.e., the "AI: Stop training?" loop in Figure 4), while the solid line part corresponds to the process when the human agents screen the prioritized papers according to the predictions from the finalized AI model.

We can see that the dotted line portions of the LC and the HP curves are much steeper than the dotted line portion of the random sampling curve. The trends suggest that with both the LC and HP sampling strategies, a larger portion of the sampled papers are included papers compared to the random sampling strategy. That is, the LC and the HP sampling strategies, especially HP, improve team screening efficiency substantially during the screening-updating-predicting-sampling iterations. This can be explained by the sampling strategies themselves. The LC strategy samples the papers with the highest classification uncertainties. Given the highly imbalanced dataset, our model is less confident in classifying the papers from the minor class, i.e., the included papers from the "1" class, leading to more papers being sampled from the minor class. The HP strategy samples the papers with the highest predicted PSs, which are more likely to be included papers by the definition of PS.

Moreover, by sampling the papers with the highest classification uncertainties, the LC sampling strategy also enables our model to learn more efficiently from human labeling compared to the other sampling strategies. This is evidenced by the observation that the solid curve part of the blue curve is steeper at the early phase than the solid curve parts of the red and black curves in Figure 8 and the highest F1 scores for LC in Figure 7.

5.2 Deployed EGM Design Results

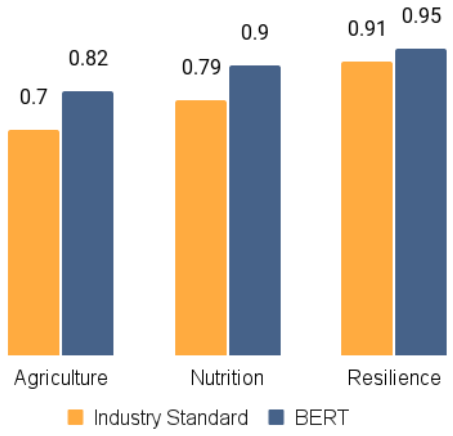

Fig. 10. The classification performance of BERT and industry standard models for the three EGMs created: Agriculture, Nutrition, and Resilience. The metric used is defined in section 4.2.1.

In this section we compare inclusion classification done by our BERT-based model and the industry standard model. We compare model performance across two different metrics: classification performance, and saved effort. The BERT model is our proposed approach, whereas the industry standard model is what tools like EPPI Reviewer [33], Rayyan [31], and RobotReviewer [32] utilize to classify documents, and therefore represents the industry standard.

5.2.1 Model Performance

Figure 10 shows the model performance (as defined in section 4.2.1) of the BERT-based and the industry standard models for each of the three EGMs we created. The results show that for all three EGMs, the BERT model resulted in higher accuracy than the industry standard model. Our results suggest that utilizing BERT for classification has benefits over the industry standard EGM-creation tools.

5.2.2 Saved Effort

Figure 11 shows how inclusion rate varies with human effort. The orange "Ideal" line indicates a perfect inclusion rate, where only relevant documents are screened and

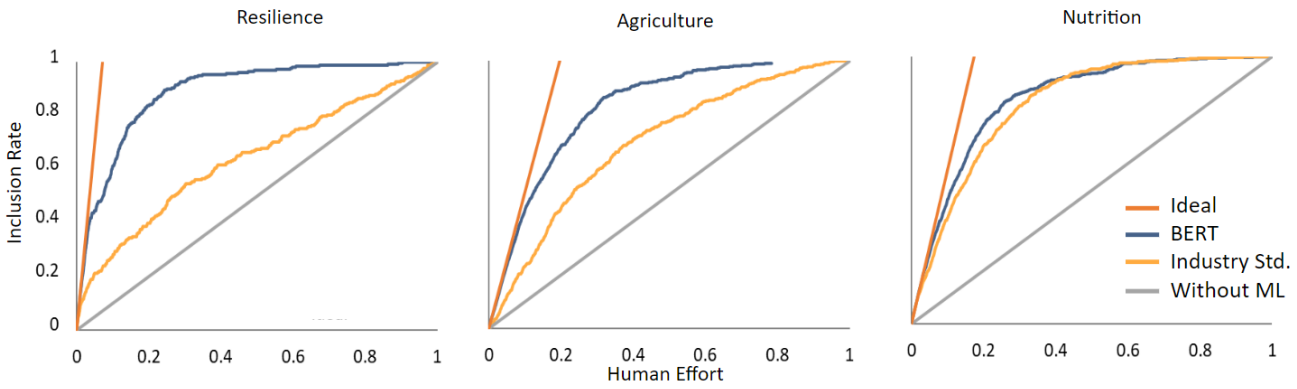

Fig. 11. How inclusion rate varies with human effort for the three deployed EGMs. A higher inclusion rate at a lower human effort is preferred. BERT outperforms the industry standard in all three EGMs.

EGM	Industry Std.	BERT	Percent Effort Saved by BERT
Resilience	68%	17%	**75%**
Agriculture	53%	28%	**47%**
Nutrition	29%	24%	**17%**

Table 2. Human effort required to reach an 80% inclusion rate for each of the three EGMs and the two ML models, the industry standard model and our BERT-based model. The percent effort saved by BERT is calculated as the percent difference between human effort for the BERT-based model and the industry standard model.

therefore all documents seen are included. The grey "Without ML" line indicates a case in which human experts must screen all documents at random in order to find all of the included documents. We compare two ML strategies, BERT and industry standard, and find that BERT outperforms industry standard for all three EGMs.

We carried out a set of experiments with the screened papers of the three EGMs for the comparison. Aiming at an inclusion rate of 80% for the screened papers, we found that the human raters needed 75% less human effort for Resilience, 47% less human effort for Agriculture, and 17% less human effort for Nutrition when working with the BERT-based models rather than with the baseline industry standard model from EPPI-Reviewer. The raw values of human effort for the BERT model and industry standard model for the three EGMs are shown in Table 2. We discuss the variation among the three EGMs in section 6.1.

The effort-saving capabilities of the BERT models are further amplified in the real screening process, in which the models are updated multiple times as new labeled documents come in as training data. In this case, the model improves iteratively over time. As its classification accuracy increases, the model can suggest only the most relevant documents to the human raters. This type of AL is explored in the simulated dataset and described in section 5.1.2.

6 Discussion

Figure 9 provides a comprehensive view of the various ML methods we compared, and their effect on human effort. The figure shows the human effort required to reach an 80% inclusion rate in the simulated EGM design case. We compare results for no ML assistance, the industry standard model, a BERT-based model, and a BERT-based model with the LC or HP AL sampling strategies. We observe that the BERT-based model outperforms the industry standard, with a 16.8% relative reduction in human effort. The results show that the AL strategies reduce human effort even further, by about 30% compared to BERT without AL. Since our motivation is to accelerate the EGM design process and decrease the resource and time intensity of the process, this result is of great significance.

Within the hybrid team, effective interactions and mutual learning between the human agents and the AI agent can improve team performance significantly. When the LC sampling strategy is applied, both the information flow from the human agents to the AI agent (i.e., human knowledge conveyed in the labeled papers) and the information flow from the AI agent to the human agents (i.e., the AI predictions conveyed in the sampled or prioritized papers) play a role in improving the efficiency of the human-AI hybrid teams. In contrast, when the HP sampling strategy is applied, the information flow from the AI agent to the human agents plays a major role in benefiting hybrid team efficiency, especially during the screen-update-predict-sample iterations. However, in HP, the information flow from the human agents to the AI agent is not as beneficial for improving model performance.

In a practical screening process, we only know the labels of a part of the papers in a dataset, which means the actual IR, as well as its overall changing trend, is unknown. In such a scenario, it is difficult to determine when to stop expanding the training set and updating the AI agent and when to stop screening the prioritized papers. The changing scale of the predicted rankings of the unlabeled papers and the growth rate of IR can inform us about the stopping. Small changes in the paper rankings and a low growth rate of IR may suggest we stop updating the AI agent and stop screening the

prioritized papers, respectively.

From the records of twelve human screeners working on an agriculture development EGM, we learn that a human screener can screen 38.6 ($SE = 1.00$) papers per hour on average. On this basis, the AL-enhanced AI agent can save human screeners $(80\% - 17.4\%) \times 68,539/38.6 = 1,111.5$ hours for TA screening compared to the case without the AI agent. Compared to the case using EPPT-Reviewer frequency-based language model and SVM-based classifier as the AI agent, $(30.3\% - 17.4\%) \times 68,539/38.6 = 229.1$ hours can be saved by our model.

6.1 When Does the BERT-based Model Most Improve Results?

We notice an interesting difference in the percent effort saved by our BERT model among the three EGMs, as shown in Table 2. We see, for example, that the BERT model significantly outperforms the industry standard for the Resilience EGM, but has a less pronounced effect on the Nutrition EGM. Interestingly, the BERT model's performance is quite consistent, so this variation is mostly a result of how well the industry standard frequency-based model can perform for different EGMs. We note that, for example, the industry standard model performs quite poorly for the resilience EGM, but quite well for the Nutrition EGM.

This suggests that based on the characteristics of an EGMs specific literature domain (such as resilience or nutrition), the industry standard model may or may not be equipped to properly classify documents. We propose that this is because the industry standard model utilizes a frequency-based language model.

Frequency-based language models, such as n-gram models, have limitations when it comes to semantic understanding due to their simplistic nature and reliance on local statistics. The limitations stem from a number of characteristics. For example, frequency-based models only consider local context (n-grams) to predict the probability of the next word. They do not take into account the broader context or global dependencies between words in a sentence or document. As a result, they may not capture the full meaning and intent of the text. Additionally, words often have multiple meanings (polysemy) or different words may sound the same (homonymy). Frequency-based models treat words in isolation and cannot disambiguate between different senses of a word based solely on local statistics. This leads to ambiguity in semantic understanding. Lastly, these models have limited generalization. Frequency-based models learn representations based on the exact sequences of words they encounter in the training data. They may struggle to generalize to unseen or slightly different contexts, leading to poor performance on tasks that require semantic understanding beyond the training data.

Alternatively, the BERT-based model exhibits strong semantic understanding owing to its bidirectional context, transformer architecture, and pretraining on large text corpora. By considering both left and right context in sentences, BERT captures long-range dependencies and contextual information, essential for understanding word meanings. Its self-attention mechanism allows each word to attend to others, modeling complex semantic relationships. Pretrained on extensive unlabeled data, BERT learns general linguistic patterns, and its contextualized word embeddings resolve word sense ambiguity. Through transfer learning and fine-tuning, BERT adapts its semantic knowledge to various NLP tasks. Its large model size further contributes to capturing intricate semantic nuances, making BERT a highly effective language representation model across diverse applications.

Because of this, for nebulous topics that take quite a bit of semantic understanding, a BERT-based model will likely outperform the industry standard model. This idea is supported in the case of the Resilience EGM. Even the human experts had to spend a long time identifying what truly makes a document relevant to the topic of "resilience." The expert screeners had to use much discernment and semantic understanding to determine if documents should be included. Whereas for other topics, like nutrition, the inclusion decisions were clearer and, often specific keywords were strong indicators that a document should be included. We therefore propose that the BERT-based model will provide more improvement over the industry standard the more nebulous a topic is, or the more semantic understanding is required to properly understand a topic.

6.2 Discussion on AI-assisted Design of EGMs

The nature of our work creating EGMs for deployment and use by USAID meant that our team faced many real-world challenges. In this section, we discuss the unique challenges and limitations that arise when designing EGMs with AIassistance. We also discuss future directions for utilizing natural language processing and human-AI teaming for creating EGMs, including the use of generative large language models.

The Cost of Communication The cost of communicating in a human-AI team can be significant but difficult to quantify as it involves multiple factors. One such factor is the time and effort required to exchange information via email, which can lead to delays and potential miscommunication. Additionally, updating document labels and merging datasets can be a complex and time-consuming task that requires oversight and project management to ensure accuracy. These activities can also be a source of errors that can negatively impact the performance of AI models. Finally, time lags between humans labeling documents, the AI agent receiving the documents and updating the model, and then the AI agent sending back newly ranked documents for human screening means that one team may be operating with incomplete data. Therefore, optimizing communication channels and implementing efficient communication protocols can help reduce the costs associated with human-AI team collaboration.

Trust in AI A major challenge that many AI recommendation systems face is the "cold start" problem. The AI agent must provide some prediction about the documents in the first iteration, but at this point, the model knows noth-

ing about the new domain. In our case, we pretrained our model on documents in the global development space, but this cannot ensure that it would perform well in classifying documents for, say, an Agriculture-specific EGM without any additional training data. This challenge, while common, can lead to distrust in AI from the human team, if they find the initial rankings to be incorrect. Additionally, the training data for our models are labels from people, which can be noisy. The AI model's performance is constrained by the quality of its training data, and therefore to have meaningful and accurate model results, we must begin with consistent high-quality training data.

"When to Stop": A Business Decision Another challenge in the deployed EGM design case study was determining when to stop the screening process, the second question shown in Figure 4. Our human-AI team faced a trade-off here between screening more papers in order to improve model performance, or stopping screening in order to move onto the next step in the EGM process (Figure 2). This is ultimately a business decision in which the team must weigh the resource cost of improving the model, and identifying the most "true positive" documents. We experimented with two techniques for determining "when to stop." The first of these techniques was calculating the similarity of the rankings of the documents when ordered based on the priority score between two consecutive iterations. If the similarity of the two rankings was above a certain value after a screening iteration, we could stop updating the BERT model. The second technique was to terminate the human screening process at a specific real-time inclusion rate, e.g., the number of relevant documents identified from screening 1,000 documents in the current iteration. If the number of relevant documents is lower than a given threshold, the human screening team could stop the screening process. Future work could specifically address the question of when to stop screening, as it is a highly relevant decision for the human-AI team.

Automation of Full-text Screening The EGM design process, as depicted in Figure 2, includes both title and abstract screening, and full-text screening. A natural continuation of our work would be to use NLP to assist in the full-text screening step. This step, however, presents the logistical challenge of obtaining the full-text documents. While many institutions have subscription-based access to scholarly article databases, copyright issues make downloading and using full-text documents a challenge when working among and between different institutions. This ultimately dictated that our project scope remains in the title and abstract screening process alone.

Additionally, to perform full-text screening, one would need to train another BERT model to classify full-text documents for inclusion. This means human screeners would need to generate a training dataset for this task, which would require significant human effort. Large language models (LLMs) may assist in this challenge. LLMs which are trained on billions of documents [48] have a broad understanding of language, and future work can explore whether they can classify full-text documents without domain specific training data.

Counterfactual Analysis Our team faced the challenge of accurately comparing different document screening techniques within the real-world setup- such as using all three methods: a BERT-based model, the industry standard model, and no ML model to make a single EGM. It was infeasible for the human raters to create each EGM three separate times in order to compare the entire process for each technique. Therefore, we standardized our comparisons by performing retrospective experiments after the human-AI team had labeled a subset of the data. We present the results using this labeled subset. We further aimed to address this limitation by including the second case study - the simulated EGM design. In this case study, we utilized a fully labeled dataset of 68,539 documents in order to experiment with the various ML models and AL sampling strategies. However, this challenge means we do not have true counterfactual analyses of how the EGM process would have proceeded without any AI assistance. Future work could further address this limitation by creating each EGM multiple times for each different technique.

6.3 Future Use of Human-AI Teams for EGM Design

The AI sub-field of NLP is experiencing rapid growth. Large language models (LLMs) like OpenAI's ChatGPT [48], and Meta's Galactica [49] are changing the way the world perceives, exploits, and interacts with pretrained language models. These models were released after we had concluded our EGM creation; however, we predict that their capabilities will shift the way that NLP is utilized in EGM design.

We performed a number of exploratory experiments to understand LLMs' capability in EGM design. We explored ChatGPT's understanding of the relationship between certain interventions and outcomes by asking it "How can agriculture transformation change poverty, migration, and food security" The LLM captured the general qualitative relationships between the intervention (agriculture transformation) and the outcomes (poverty, migration, and food security), but did not output any quantitative implications, potential information sources, or indications of how well the relationships have been studied. This suggests to us that ChatGPT can capture the intervention-outcome relationships in a coarse resolution, but cannot provide all the detailed information that a human team or human-AI hybrid team can capture.

ChatGPT Experiment In a small-scale study, we explored whether ChatGPT could correctly output the interventions and outcomes of documents if shown the abstract and a set of intervention options and outcome options. We tested this on 50 documents that are part of 3ie's published EGM represented in Figure 1. The five intervention options we provided in the prompt were as follows:

1. Water access and management
2. Improved seeds
3. Fertilizer access
4. Pesticide/herbicide access
5. Livestock access

The four outcome options were:

1. Income received
2. Assets
3. Output value
4. Profit

Looking at the top 1 accuracy, we gathered the following results. For predicting the intervention, ChatGPT produced the correct output for 34 of the 50 documents. For predicting the outcome, ChatGPT produced the correct output for 22 of the 50 documents. In the cases in which the tool was incorrect, the generative nature of this tool seemed to combat the strict classification guidelines. For example, for some abstracts ChatGPT would output two outcomes. This went against the prompt, and did not align with the setup of each document only appearing under one intervention and one outcome in an EGM. However, identifying multiple outcomes *was* consistent with certain aspects. For example, during the reviewing process, each document may be assigned multiple interventions and outcomes, then expert raters must select one of these options for the final EGM. So while we ultimately want to classify documents into a single intervention and outcome, providing multiple options may still be valuable.

The tool would often misclassify documents that were meant to have the outcome of "Income received," "Output Value," or "Profit" for one another. In fact, 18 of the 28 incorrect outcome results fell into this category. The provided definition of "Income received" is "The total monetary income earned from some activity by an individual, household or firm," of "Output Value" is "Some measure of the value of the output produced as a result of an intervention," and of "Profit"is "Individual and store revenue or profit. Here profit refers to income net of costs." The difference between these three definitions is subtle, which might explain the frequent misclassification among these three options. To add to that, in our experiment we do not provide these definitions, solely the intervention and outcome titles. We hypothesize that effective prompt engineering will be required to obtain the best results possible.

Another case that arose frequently was an abstract that explicitly stated that the intervention was, for example, a "Farm Input Subsidy Programme" (FISP). In these cases, ChatGPT would often stray from the five provided intervention options and return the explicitly stated intervention (e.g. FISP) as its answer. It was unable to determine under which of the five intervention options FISP should fall. Cases like these suggest that providing a generative LLM tool with a set of heuristic rules developed from expert knowledge could improve performance.

7 Conclusion

In this paper, we have studied (1) how incorporating the BERT-based AI agent into the human team affects team efficiency in the EGM design process and (2) how enhancing the hybrid team through active learning (AL) can improve hybrid team efficiency. We propose a human-AI hybrid teaming workflow during TA screening portion of the EGM design process. We a) design and deploy three EGMs for global development in the areas of Agriculture, Nutrition, and Resilience, and b) conduct simulated experiments with a fully labeled dataset to answer the research questions described above. Our results show that the data size for training the AI agent influences hybrid team efficiency. When the training size is optimized, the incorporation of the BERT-based AI agent can reduce human effort by 68.5% compared to the case without AI assistance and by 16.8% compared to the case using the industry standard AI-agent for getting to an inclusion rate of 80%. Moreover, enhancing the hybrid team through AL can further reduce human effort by 30% compared to BERT with no AL. The proposed human-AI hybrid teaming workflow has been validated in the practical construction process of three EGMs. Therefore, the AL-enhanced human-AI hybrid team can accelerate evidence gap map (EGM) design, and decision making in the global development field significantly.

Acknowledgments

This publication was made possible through support provided by the USAID Bureau for Resilience and Food Security, U.S. Agency for International Development. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Agency for International Development.

References

[1] USAID. *Strengthening Evidence-Based Development*. USAID (2016).

[2] 115th Congress, U.S. "Foundations for Evidence-Based Policymaking Act of 2018." (2018).

[3] Kraybill, David and Mercier, Stephanie. "How the United States Benefits from Agricultural and Food Security Investments in Developing Countries." (2019). URL https://www.usaid.gov/sites/default/files/documents/1867/BRIEF_-_US_Benefits_Overview.pdf.

[4] White, Karen. "Publications Output: U.S. Trends and International Comparisons." https://ncses.nsf.gov/pubs/nsb20206/ (2019). Accessed: 14-January-2022.

[5] Donnelly, Christl A., Boyd, Ian, Campbell, Philip, Craig, Claire, Vallance, Patrick, Walport, Mark, Whitty, Christopher J. M., Woods, Emma and Wormald, Chris. "Four principles to make evidence synthesis more useful for policy." *Nature* Vol. 558 (2018): pp. 361–364.

[6] Snilstveit, Birte, Vojtkova, Martina, Bhavsar, Ami, Stevenson, Jennifer and Gaarder, Marie. "Evidence & Gap Maps: A tool for promoting evidence informed policy and strategic research agendas." *Journal of Clinical Epidemiology* Vol. 79 (2016): pp. 120–129. DOI 10.1016/j.jclinepi.2016.05.015. Accessed

2019-07-06, URL https://www.jclinepi.com/article/S0895-4356(16)30190-1/abstract.

[7] 3ie. "Evidence Gap Maps." Available at https://www.3ieimpact.org/evidence-hub/evidence-gap-maps (2021/08/04) (2021).

[8] NLPIE. "BioMedICUS." https://nlpie.github.io/biomedicus/ (2019).

[9] Bommarito, Michael J, Katz, Daniel Martin and Detterman, Eric M. "LexNLP: Natural language processing and information extraction for legal and regulatory texts." (2018). URL 1806.03688.

[10] Porciello, Jaron, Ivanina, Maryia, Islam, Maidul, Einarson, Stefan and Hirsh, Haym. "Accelerating evidence-informed decision-making for the Sustainable Development Goals using machine learning." *Nature Machine Intelligence* Vol. 2 (2020): pp. 559–565.

[11] Engelbert, Mark, Ravat, Zafeer, Quant, Katherine, Respekta, Maciej, Kastel, Fiona, Huang, Carolyn, Frey, Dan, Ahmed, Faez, Song, Binyang, Edwards, Kristen Marie, Porciello, Jaron and Snilstveit, Birte. "Agriculture-led Growth in Low- and Middle-income Countries: An Evidence Gap Map." (2023).

[12] Lane, Charlotte, Storhaug, Ingunn, Tree, Veronika, Cordova-Arauz, Diana, Huang, Carolyn, Frey, Dan, Ahmed, Faez, Song, Binyang, Edwards, Kristen Marie, Porciello, Jaron and Snilstveit, Birte. "Addressing the systemic causes of malnutrition: The nutrition-sensitive agriculture evidence gap map." (2023).

[13] Berretta, Miriam, Lee, Sanghwa, Kupfer, Meital, Huang, Carolyn, Ridlehoover, Will, Frey, Dan, Ahmed, Faez, Song, Binyang, Edwards, Kristen Marie, Porciello, Jaron, Eyers, John and Snilstveit, Birte. "Strengthening resilience against shocks, stressors and recurring crises in low- and middle-income countries: an evidence gap map." (2023).

[14] Briner, Rob B. and Denyer, David. "Systematic review and evidence synthesis as a practice and scholarship tool." *Handbook of evidence-based management: Companies, classrooms and research*. New York University Press (2012): pp. 112–129.

[15] Blaizot, Aymeric, Veettil, Sajesh K., Saidoung, Pantakarn, Moreno-Garcia, Carlos Francisco, Wiratunga, Nirmalie, Aceves-Martins, Magaly, Lai, Nai Ming and Chaiyakunapruk, Nathorn. "Using artificial intelligence methods for systematic review in health sciences: A systematic review." *Research Synthesis Methods* Vol. 13 No. 3 (2022): pp. 353–362. DOI https://doi.org/10.1002/jrsm.1553. URL https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1553, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1553.

[16] Altena, Allard, Spijker, René and Olabarriaga, Silvia. "Usage of Automation Tools in Systematic Reviews." *Research Synthesis Methods* Vol. 10 (2018). DOI 10.1002/jrsm.1335.

[17] 3ie. "Evidence Mapping." Available at https://www.3ieimpact.org/evidence-hub/evidence-gap-maps (2021/08/04) (2021).

[18] O'Mara-Eves, Alison, Thomas, James, McNaught, John, Miwa, Makoto and Ananiadou, Sophia. "Using text mining for study identification in systematic reviews: a systematic review of current approaches." *Systematic reviews* Vol. 4 No. 1 (2015): p. 1. Accessed 2016-03-11, URL http://systematicreviewsjournal.biomedcentral.com/articles/10.1186/2046-4053-4-5. 00000.

[19] Oliver, Kathryn, Innvar, Simon, Lorenc, Theo, Woodman, Jenny and Thomas, James. "A systematic review of barriers to and facilitators of the use of evidence by policymakers." *BMC Health Services Research* Vol. 14 (2014): p. 2. DOI 10/gbfqbx. Accessed 2016-07-30, URL http://dx.doi.org/10.1186/1472-6963-14-2. 00000.

[20] Shannon, C. E. "A mathematical theory of communication." *The Bell System Technical Journal* Vol. 27 No. 4 (1948): pp. 623–656. DOI 10.1002/j.1538-7305.1948.tb00917.x.

[21] Jurafsky, Daniel and Martin, James H. "Speech and Language Processing." 3rd ed. Pearson (2020), Chap. N-gram Models.

[22] Bird, Steven, Klein, Ewan and Loper, Edward. *Natural Language Processing with Python*. O'Reilly Media (2009).

[23] Manning, Christopher D. and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. MIT Press (1999).

[24] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz and Polosukhin, Illia. "Attention Is All You Need." (2017). URL 1706.03762.

[25] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg and Dean, Jeffrey. "Distributed Representations of Words and Phrases and their Compositionality." (2013). URL 1310.4546.

[26] Mikolov, Tomas, Chen, Kai, Corrado, Greg and Dean, Jeffrey. "Efficient Estimation of Word Representations in Vector Space." (2013). URL 1301.3781.

[27] Pennington, Jeffrey, Socher, Richard and Manning, Christopher D. "Glove: Global Vectors for Word Representation." *EMNLP*, Vol. 14: pp. 1532–1543. 2014.

[28] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." (2019). URL 1810.04805.

[29] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke and Stoyanov, Veselin. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." (2019). URL 1907.11692.

[30] Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario and Sutskever, Ilya. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog* (2019).

[31] Ouzzani, Mourad, Hammady, Hossam, Fedorowicz, Zbys and Elmagarmid, Ahmed. "Rayyan—a web and mobile app for systematic reviews." *Systematic Reviews* Vol. 5 No. 210 (2016). DOI https://doi.org/10.1186/s13643-016-0384-4.

[32] Marshall, Iain J, Kuiper, Joel, Banner, Edwards and Wallace, Byron C. "Automating Biomedical Evidence Synthesis: RobotReviewer." 2017. DOI 10.18653/v1/P17-4002.

[33] Thomas, James, McDonald, Steve, Noel-Storr, Anna, Shemilt, Ian, Elliott, Julian, Mavergames, Chris and Marshall, Iain J. "Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews." *Journal of Clinical Epidemiology* Vol. 133 (2021): pp. 140–151. DOI https://doi.org/10.1016/j.jclinepi.2020.11.003. URL https://www.sciencedirect.com/science/article/pii/S0895435620311720.

[34] Aum, Sungmin and Choe, Seon. "srBERT: automatic article classification model for systematic review using BERT." *Systematic Reviews* Vol. 10 No. 285 (2021). DOI https://doi.org/10.1186/s13643-021-01763-w10.1002/jrsm.1335.

[35] Ein-Dor, Liat, Halfon, Alon, Gera, Ariel, Shnarch, Eyal, Dankin, Lena, Choshen, Leshem, Danilevsky, Marina, Aharonov, Ranit, Katz, Yoav and Slonim, Noam. "Active Learning for BERT: An Empirical Study." *EMNLP*. 2020.

[36] Aggarwal, Umang, Popescu, Adrian and Hudelot, Celine. "Minority Class Oriented Active Learning for Imbalanced Datasets." *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021. IEEE. DOI 10.1109/icpr48806.2021.9412182. URL https://doi.org/10.1109%2Ficpr48806.2021.9412182.

[37] Gal, Yarin, Islam, Riashat and Ghahramani, Zoubin. "Deep Bayesian Active Learning with Image Data." (2017). DOI 10.48550/ARXIV.1703.02910. URL https://arxiv.org/abs/1703.02910.

[38] Tur, Gokhan, Hakkani-Tur, Dilek and Schapire, Robert E. "Combining active and semi-supervised learning for spoken language understanding." *Speech Communication* Vol. 45 (2005): pp. 171–186.

[39] Chung, Mu-Huan, Chignell, Mark, Wang, Lu, Jovicic, Alexandra and Raman, Abhay. "Interactive Machine Learning for Data Exfiltration Detection: Active Learning with Human Expertise." *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*: pp. 280–287. 2020. DOI 10.1109/SMC42975.2020.9282831.

[40] Olsson, Fredrik. "A literature survey of active machine learning in the context of natural language processing." Technical report no. Swedish Institute of Computer Science. 2009.

[41] Angluin, Dana. "Queries and Concept Learning." *Machine Learning* Vol. 2 (1998): pp. 319–342.

[42] Atlas, Les, Cohn, David and Ladner, Richard. "Training Connectionist Networks with Queries and Selective Sampling." Touretzky, D. (ed.). *Advances in Neural Information Processing Systems*, Vol. 2. 1989. Morgan-Kaufmann. URL https://proceedings.neurips.cc/paper/1989/file/b1a59b315fc9a3002ce38bbe070ec3f5-Paper.pdf.

[43] Lewis, David D. and Gale, William A. "A Sequential Algorithm for Training Text Classifiers." *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: p. 3–12. 1994. Springer-Verlag, Berlin, Heidelberg.

[44] Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648. University of Wisconsin–Madison. 2009.

[45] Culotta, Aron and Andrew, McCallum. "Reducing Labeling Effort for Structured Prediction Tasks." *AAAI*: p. 746–751. 2005.

[46] Scheffer, Tobias, Decomain, Christian and Wrobel, Stefan. "Active Hidden Markov Models for Information Extraction." *IDA*. 2001.

[47] 3ie. "Development Evidence Portal." Available at https://developmentevidence.3ieimpact.org (2021/08/04) (2021).

[48] OpenAI, TB. "Chatgpt: Optimizing language models for dialogue." *OpenAI* (2022).

[49] Taylor, Ross, Kardas, Marcin, Cucurull, Guillem, Scialom, Thomas, Hartshorn, Anthony, Saravia, Elvis, Poulton, Andrew, Kerkez, Viktor and Stojnic, Robert. "Galactica: A Large Language Model for Science." (2022). DOI 10.48550/ARXIV.2211.09085. URL https://arxiv.org/abs/2211.09085.