ACS ENVIRONMENTAL Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Open Access

Evaluating GPT Models for Automated Literature Screening in Wastewater-Based Epidemiology

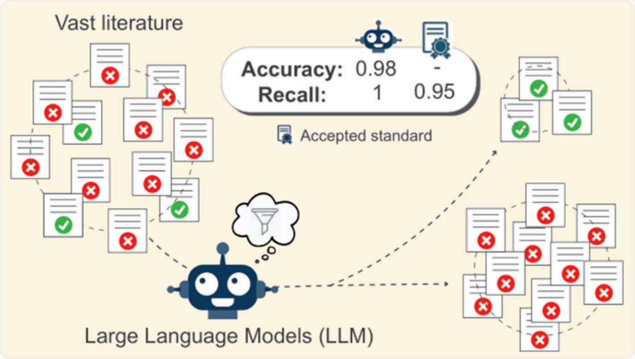
Published as part of ACS Environmental Au special issue "2024 Rising Stars in Environmental Research".

Kaseba Chibwe, David Mantilla-Calderon, and Fangqiong Ling*

Cite This: *ACS Environ. Au* 2025, 5, 61−68

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Methods to quantitatively synthesize findings across multiple studies is an emerging need in wastewater-based epidemiology (WBE), where disease tracking through wastewater analysis is performed at broad geographical locations using various techniques to facilitate public health responses. Meta-analysis provides a rigorous statistical procedure for research synthesis, yet the manual process of screening large volumes of literature remains a hurdle for its application in timely evidence-based public health responses. Here, we evaluated the performance of GPT-3, GPT-3.5, and GPT-4 models in automated screening of publications for meta-analysis in the WBE literature. We show that the chat completion model in GPT-4 accurately differentiates papers that contain original data from those that did not with texts of the Abstract as the input at a Precision of 0.96 and Recall of 1.00, exceeding current quality standards for manual screening (Recall = 0.95) while costing less than $0.01 per paper. GPT models performed less accurately in detecting studies reporting relevant sampling location, highlighting the value of maintaining human intervention in AI-assisted literature screening. Importantly, we show that certain formulation and model choices generated nonsensical answers to the screening tasks, while others did not, urging the attention to robustness when employing AI-assisted literature screening. This study provided novel performance evaluation data on GPT models for document screening as a step in meta-analysis, suggesting AI-assisted literature screening a useful complementary technique to speed up research synthesis in WBE.

KEYWORDS: *GPT-4, wastewater-based epidemiology, fine-tuning, systematic review, meta-analysis*

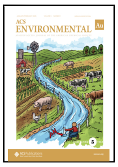
1. INTRODUCTION

Wastewater-based epidemiology (WBE) is a biomonitoring methodology that tracks community-level health statuses through the analysis of untreated wastewater.[1] With the intention of WBE as to facilitate evidence-based public health responses, citable statistics with larger sample size and greater statistical power are required, calling for quantitative synthesis of the vast literature through methods such as meta-analysis.[2,3] In addition, with wastewater-monitoring data generated via multiple techniques, from complex sewage systems, and by numerous research groups at diverse geographical locations,[4−6] meta-analyses are crucial for understanding the sources of variations across studies,[7,8] so that infrastructure built from the COVID-19 pandemics can be effectively leveraged and updated for tracking new or unknown pathogens in a "disease X" framework highlighted by the WHO.[5] Taken together, there is an urgent need of methods to enable timely quantitative research synthesis in WBE.

A necessary step preceding the statistical procedures in meta-analyses is to amass data from multiple studies using a formal and reproducible methodology encompassing literature searching, screening, and critical appraisal, i.e., systematic reviews.[9] This rigorous process can be laborious and time-consuming, taking an average time of 67.3 weeks by a five-person team in medicine[10] and 164 full-time equivalent days in environmental science.[11] Among steps in the systematic review process, literature screening is perhaps the most laborious step. It has been reported that only 2.9% of studies retrieved by literature search are relevant and included in the final synthesis.[10]

To reduce the workload and improve the efficacy of evidence synthesis, researchers are exploring artificial intelligence (AI) and machine learning (ML) methods to automate

Received: June 25, 2024
Revised: November 22, 2024
Accepted: November 25, 2024
Published: December 3, 2024



https://doi.org/10.1021/acsenvironau.4c00042

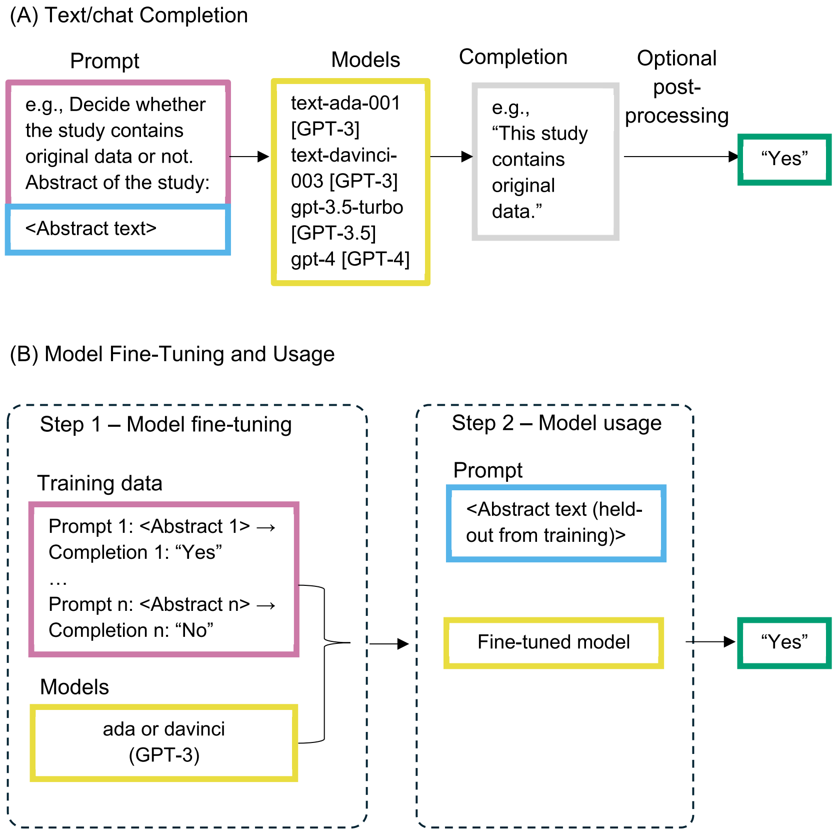


Figure 1. Formulation of Abstract screening as text completion (A) or classification (B) tasks in GPT-3/3.5/4 framework. (A) In the text completion approach, a question and the text of the Abstract of a specific paper is provided in the prompt, and a GPT-3/3.5/4 model is deployed. The model generates a completion based on the question asked and the Abstract provided. Depending on the question asked, a postprocessing step may be required to streamline the answers to a "yes" or "no". (B) In the fine-tuned classifier approach, a data set containing multiple Abstracts and their corresponding labels (yes or no) is provided as the training data, and a GPT-3 model (e.g., ada or davinci) is fine-tuned. The fine-tuned model is used to determine whether a previously unseen Abstract is from a paper that reported original data or not, and a binary outcome (yes or no) is generated as the model output.

steps in systematic reviews. While previous works have applied natural language processing (NLP) models to coarsely categorize research papers,[12,13] less has been reported about automated screening based on custom criteria addressing specific research questions, a much more challenging task, with notable exceptions.[14] Furthermore, with current studies focused primarily on established fields where there are large bodies of literature and universal languages, such as terms widely used in public health (e.g., "population", "exposure", "confounders", and "outcomes" in medicine)[15] or the synthesis conditions in metal−organic frameworks,[16,17] it is unclear whether AI/ML tools can effectively facilitate research synthesis in emerging fields where there are anticipated challenges such as smaller training data (e.g., total quantities of publications) and rapidly iterating semantics.[14,15,18,19]

Generative Pretrained Transformers (GPT) are a suite of neural network ML models trained over texts generated on the Internet.[20] They are attractive for synthesizing literature in WBE as an emerging field because of their ability to perform zero-shot or few-shot learning, where the model generates predictions with seeing zero or few examples. GPT-4 is the most recent development of the GPT models that distinguished itself from previous versions (e.g., GPT-3/3.5) by higher capabilities of handling complex problems.[21] While GPT models are evolving quickly in its complexity, a more complex model is not always better for a specific task. Additionally, users can choose to directly deploy the model as one or several "prompt(s)" (i.e., text/chat completion) or to fine-tune the model, where fine-tuning is known to generally improve performance yet also more costly. Thus, careful performance and cost analyses are required for specific applications. To our best knowledge, fine-tuning and GPT-3 or her more advanced sibling models to perform literature screening to aid systematic reviews in environmental engineering literature has not been reported.

In this study, we explored the use of GPT-3/4 and their siblings to aid research synthesis. Specifically, we asked GPT-3/3.5/4 models to screen published papers and differentiate whether original and relevant data were reported in the paper. Here, we define "original data" as when any new sampling or measurements were performed, in contrast to other types of work, such as research synthesis or computer simulations, and "relevant data" were defined using specific questions about sampling locations. We ask five research questions: 1) How well do pretrained GPT models perform in the task of detecting original data from WBE literature based on Abstracts?; 2) Does fine-tuning improve the performance?; 3) How does the complexity of pretrained models affect the performance and cost?; 4) How does performance or cost change when models are supplied with the texts of the "Methods" section as inputs in comparison to Abstracts?; 5) When a more advanced question is asked, how will the question itself affect the performance and cost of various GPT models?

2. METHODS

2.1. Problem Formulation

In this study, ML models were formulated to take the texts in the "Abstract" or "Methods" section of a paper as inputs, and to output whether that paper meets a specific inclusion criterion relevant for WBE. This task can be formulated as text/chat completion or classification (Figure 1). With various combinations in problem formulations, model choices, input text choices, and screening tasks, a total of 24 test cases were formulated in this study (Figure S1). Detailed descriptions of implementation procedures can be found in the Section 2.3.

2.2. Data Set and Screening Tasks

We constructed a data set based on a set of articles previously curated for meta-analysis on WBE[7] and performed further curation and labeling. Briefly, the data set contained "Abstract" and "Methods" sections of 101 research articles which were retrieved from literature research using the search term combination "TS = (SARS-CoV-2 AND (wastewater OR sewage))" from the Web of Science core collection.

Two screening tasks were tested in this study. First, noting that 70% of papers generated by the above-mentioned search strategy did not contain original data,[7] we set "the study contains original data" as one task (hitherto referred to as "original data detection"). We designed this task because meta-analyses typically build on primary studies where new experiments are performed to estimate an effect size, in contrast to studies that utilize the data generated from other studies, e.g., research synthesis or computer simulations. Second, a screening task specifically of interest to WBE is whether the sampling location was from the wastewater system, thus we set "sampling location is identified as wastewater treatment plant, lift station, manholes or septic tank near a building/hospital" as a second task (hitherto referred to as "relevant sampling location"). We designed this task because sampling locations are important data for WBE, significantly affecting the sampled populations.[22] As a benchmark, all the papers were manually labeled by two authors of this manuscript who also authored the original meta-analysis (DMC and FL).[7]

2.3. Procedures

2.3.1. Data Set Preparation. The Abstracts were retrieved directly from the Web of Science search output. The texts of the Method sections for the papers were retrieved from the full text pdf files using the Python package "scipdf_parse".[23] Papers where the categories were ambiguous even by human screening were excluded from the study, resulting in 92 Abstracts for original data detection and 90 Abstracts for relevant sampling location detection. Since some of the papers in the data set are reviews or perspectives, only 62 papers contained a Method section. The texts from those papers were included in the data set. Papers without a "Methods" section were excluded from the learning tasks using Methods as inputs.

It should be noted that for most models that were used, the maximum number of tokens in a prompt is 2,048. As described by OpenAI, one token is approximately 4 characters or 0.75 words for English text.[24] To apply consistent maximum word counts in the data inputs, we iteratively applied various maximum word counts until no error messages were produced, which led to 6,700 characters as the maximum inputs. None of the Abstracts fell above 6,700 characters, yet some of the Methods exceeded this length. When an input exceeded the maximum length, we included only the first 6,700 characters in the prompt. Because GPT-3.5-turbo and GPT-4 have higher values for the maximum token limits (4,096 tokens and 8,192 tokens, respectively), we did not have to set a limit on the length of the Methods Section for the GPT-3.5-turbo and GPT-4 cases. To remove the influences of special symbols in the Methods section, we also removed non-English words from the Methods. This was achieved by checking whether a certain word can be encoded only with ASCII characters.

2.3.2. Data Set Pretreatment. In order to wrangle the data into a format that can be submitted to OpenAI tools, Python packages "pandas" and "sklearn" were used to clean the data.[25,26]

2.3.3. OpenAI Tool Usage (Text Prompt, Fine-Tuning, and Usage of the Fine-Tuned Models). The tools from OpenAI's developer platform were used. The Python bindings for API interactions developed by OpenAI were used. Among those models, "text-ada-001" and "text-davinci-003" were available through the "/v1/completions" model end point, whereas "gpt-3.5-turbo" and "gpt-4" were available through the "/v1/chat/completions" model end point. GPT-3, GPT-3.5, and GPT-4 models were deployed in this study.

Among the GPT-3/3.5/4 models open to developers at the time of drafting this manuscript, classification is supported by four GPT-3 models. Two of them were explored here: "ada", usually the fastest model, and "davinci", the most capable GPT-3 model. Among the text completion tasks, two API end points were used in this study: the classic text completion and the chat completion (Figure S2). According to the OpenAI documentation, "the difference between these APIs derives mainly from the underlying GPT models that are available in each".[27] In our study, the chat completion and traditional text completion were used to achieve the same goals and were thus referred to as text/chat completions. Four text/chat completion models were explored, "text-ada-001 [GPT-3]", "text-davinci-003 [GPT-3]", "gpt-3.5-turbo", and "gpt-4". More information about these models can be found in the Supplementary Methods.

2.3.4. Model Outputs. For each case, the performance and costs were summarized. In addition, within a test case, the response for each request (a request = a specific paper) was also documented.

2.3.5. Additional Information about the Procedures. A case study is available at WU Box (https://wustl.box.com/s/kjs5kh8e2iq4qm7u02b04euz1han9exa). More information about the fine-tuning parameters is provided in Supporting Information.

2.4. Model Evaluation

Model performances were evaluated by calculating the Precision, Recall, and F1 score of model outputs using the "metrics" function in the "sklearn" package in Python. For classifiers, we fine-tuned the models with 50% of our data and performed evaluations on the other 50% of data (hold-out). The average performance scores and their standard deviations (SD) were computed from the results of four splits using random seeds. As for text/chat completion models, performance scores were computed using all data available because no data requirement from fine-tuning was incurred.

The definitions of the above metrics are shown in eqs 1−4).

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} \tag{1}$$

in which "true positive" refers to the correct classification outcome as positive and "false positive" refers to the incorrect classification outcome as positive. For example, when a study containing original data from wastewater-based monitoring is accurately identified as such, it is considered a true positive. Conversely, when a study without original data (e.g., editorials, research synthesis, or other studies that contributed to science but did not take new measurements) is incorrectly classified by the model as having original data, this outcome would be labeled a false positive.

$$\text{recall} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \tag{2}$$

Recall is also known as sensitivity, with a high recall suggesting that the model is good at capturing positive cases. In the case of classifier-aided literature screening for research synthesis, we emphasize the importance of high recall because we envision the literature screening step to occur at the earlier stage of literature screening when researchers would like to include as many qualified studies as possible.

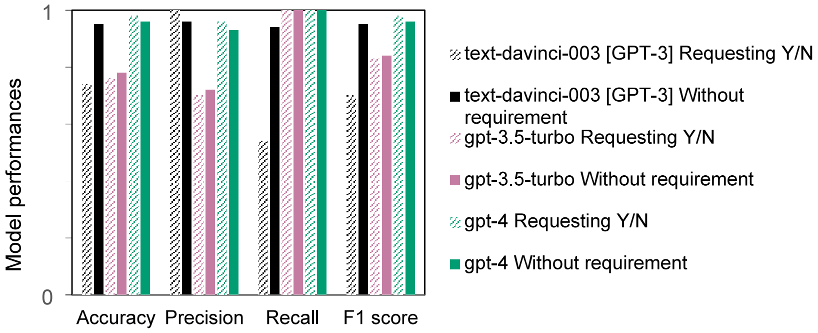


Figure 2. Performances of the text/chat completion models in the original data detection task. Model performances were evaluated on all available data (n_{test} = 92 when Abstract texts were the inputs, and n_{test} = 62 when Methods text were the inputs). There was no fine-tuning step for text-completion tasks; thus, none of the data were reserved for fine-tuning. Gpt-4 and gpt-3.5 achieved ultrahigh Recall, and text-davinci-003 generated excellent recall (>0.95). Query scenarios with and without requesting a "yes" or "no" performed differently.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

in which "precision" is the ratio of true positive predictions to the total predicted positives. The F1 score is a classifier performance measure that balances both precision and recall. In this study, the F1 score was chosen over accuracy because qualified studies usually outnumber unqualified studies in a literature screening task (i.e., imbalanced classification tasks), and the F1 score is more informative in this scenario.

Accuracies were recorded for completeness.

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{number of total predictions}} \tag{4}$$

2.5. Cost Estimation

OpenAI tools reported costs as either USD or as the number of tokens used, where costs were linearly depended on the numbers of tokens used.[28] We converted the costs expressed as tokens used to USDs for comparisons. For comparisons across fine-tuned and the text/chat completion models, normalized costs per 100 papers were calculated based on the testing cost in fine-tuned models and the usage costs of the text/chat completion models.

3. RESULTS AND DISCUSSION

3.1. Abstract Screening Can Be Formulated As a Text Completion or a Classification Task, while Classification Was More Robust

We asked GPT models to "read" a paper's Abstract and determine whether the paper satisfied a screening criterion. We show that this task can be addressed using two approaches, i.e., as a "text/chat completion" or as a "classification" task (Figure 1). In the text/chat completion approach, a GPT-3/3.5/4 model is provided with the text of the Abstract of a paper, and then answers a question "Does this paper contain original data?" The questions were asked either in a way to require the answer to be "yes" or "no", or without the requirement (Figure S2). In the classification approach, the classifiers were fine-tuned using human-labeled examples, and then used on unseen papers to generate binary responses. Essentially, our text completion model performed "zero-shot" learning, and the classifier model had seen examples through the fine-tuning process. Besides, the text/chat completion model needed to understand the question, whereas the classifier model treated the task as producing binary responses following the examples in fine-tuning. While either format generated meaningful responses, nonsensical answers were generated from several text completion models (Table S1). Those models were excluded from performance comparisons. Nonsensical answers is an issue that is documented in LLMs elsewhere and referred to as hallucination.[29] Thus, while literature screening tasks can be formulated either way, the classifier approach lends the advantage of not generating nonsensical answers, thus making a more streamlined process.

3.2. Abstract Screening Performance As Text Completion with GPT-3, GPT-3.5, and GPT-4 models

Effective tools for literature screening require high Recalls, which indicates very few or zero papers were missed from the Abstract screening process, i.e., very low false negative. Notably, the Recalls for gpt-3.5-turbo and gpt-4 were both 100% in Abstract screening for original data (Figure 2), regardless of whether a "yes" or "no" answer was enforced (Figure 2). The model text-davinci-003 generated varied Recalls depending on the way the questions was asked, with not enforcing "yes" or "no" generating higher Recall of 0.94 (Figure 2). With a Recall of 0.95 reported elsewhere as a standard for quality control in literature screening,[30,31] gpt-3.5-turbo and gpt-4 well exceeded that standard, whereas text-davinci-003 came close.

For literature screening, a high Precision indicates a low false positive, and hence a higher efficiency in the process. Between our high Recall models, gpt-4 generated higher Precisions (up to 0.96, Figure 2), whereas gpt-3.5-turbo models performed moderately well (up to 0.72, Figure 2). The balance between Precision and Recall is reflected in the F1 score, which indicates the overall performance (Figure 2). The gpt-4 model had the highest F1 score (0.98) among text/chat completion models, followed by text-davinci-003 and gpt-3.5-turbo (0.94 and 0.89, respectively). Thus, when workload reduction is considered, GPT-4 was more attractive (Figure 2).

The performance of certain models was improved by tweaking the way the questions were asked. For instance, two scenarios were tested when using the text-davinci-003 model. First, we asked "Does this paper contain original data? Answer yes or no" (Figure S2 A), and the Recall was 0.54. Then, we asked the same question but removed the requirement of "yes or no" (Figure S2 B). This adjustment allowed the model to generate a longer reply, resulting in an increase in Recall to 0.94 (Table S2), though the improvement did not occur in all models (Figure 2, Table S2).

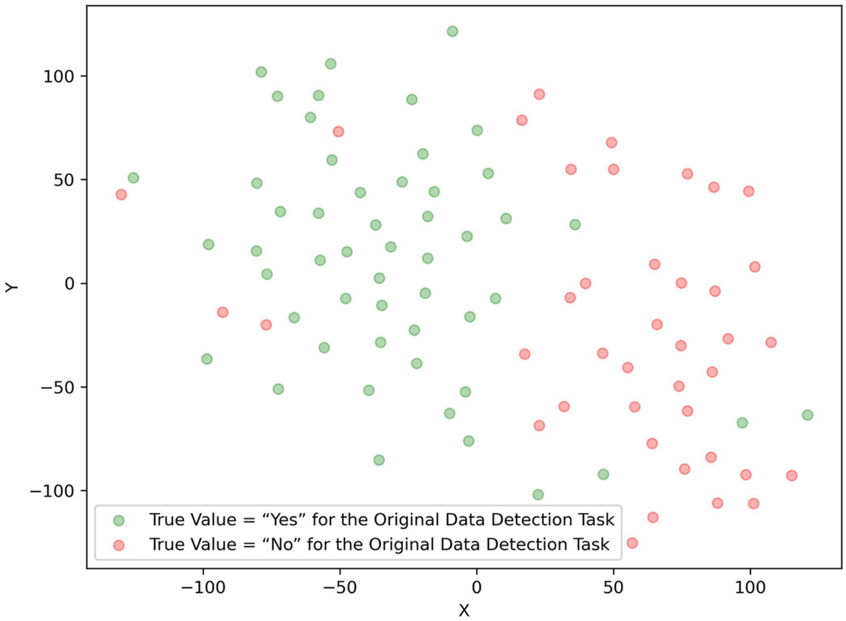


Figure 3. t-SNE plot showing marked differences in the text embeddings of Abstracts from papers containing original data and those that do not. The green dots represent embeddings of Abstracts from papers that contained original data (n = 52), and the red dots represent those from papers that did not contain original data (n = 40).

3.3. Performance of Abstract Screening for Original Data Detection with Various GPT-3 Classifier Models

The classifier models can be fine-tuned in one research group and applied elsewhere, thus are attractive for the reproducibility that they can render. Our fine-tuned classifier built over davinci generated a high Recall (mean = 0.99, SD = 0.02, Table S4), which exceeded the quality standard of 0.95 albeit slightly lower than the gpt-4 (Figure S3). The ada-based classifiers also performed well, generating an average Recall of 0.93 (SD = 0.09), although lower than those of fine-tuned davinci and gpt-4 (Figure S3). As to the overall performance, the fine-tuned davinci classifier yielded an average F1 score of 0.95 (SD = 0.05), which is higher than that of fine-tuned ada yet slightly lower than those of the gpt-4 model.

3.4. Performance of Abstract Screening for to Detect Relevant Sampling Locations

We asked whether the GPT models can provide sensible and accurate answers to a question that required deeper understanding about environmental engineering domain knowledge: "Sampling location is identified as a wastewater treatment plant, lift station, manholes/septic tank near a building/hospital". In terms of robustness, the chat completion models gpt-4 and gpt-3.5, and the fine-tuned classifiers built over davinci [GPT-3] and ada [GPT-3] generated robust responses, whereas the GPT-3 text competition models (text-davinci-003 and text-ada-001) generated nonsensical responses. When Abstracts were taken as inputs, the best performance was detected in the fine-tuned classifiers built over davinci (Figure S4, Table S6), with an average recall of Recall of 0.84 (SD = 0.04) and F1 score of 0.83 (SD = 0.04). This is followed by the fine-tuned ada model, with an average Recall of 0.83 (SD = 0.05) and F1 score of 0.77 (SD = 0.10). The gpt-4 and gpt-3.5-turbo models performed moderately in this task, generating Recalls of 0.60 and 0.43, and F1 scores of 0.66 and 0.60, respectively (Table S5).

3.5. Effects of Input Texts on Model Performance

Using the texts of Methods as the input improved the performance of the relevant location detection task for gpt-4 as well as fine-tuned davinci [GPT-3] and ada [GPT-3] models (Figure S4, Table S5−S6). In particular, the fine-tuned davinci model achieved a Recall of 0.92 (SD = 0.03), bringing it very close to the quality standard (0.95). On the other hand, there was no improvement in gpt-3.5-turbo. The effects also varied across models in the original data detection task (Table S3 and S4, Figure S3). The text-davinci-003 [GPT-3] model was able to take Abstracts and generate meaningful responses, yet not so when the inputs shifted to the texts of Methods. A potential explanation that longer texts added burdens to processing.[29] However, the fact that gpt-4 as the latest GPT model showed improvement with texts in Methods as inputs was promising. Potentially, when LLMs further improves, more sections of a research paper can be utilized to perform screening tasks that require deeper understandings of the domain knowledge.

4. IMPLICATIONS AND DISCUSSION

4.1. Implications

Our study demonstrated that the GPT-3, GPT-3.5, and GPT-4 models can facilitate research synthesis in the WBE literature by accurately detecting paper containing original data from the literature, with gpt-4, gpt-3.5-turbo, and fine-tuned davinci models generating Recall values near 1.00, exceeding the current standards in quality from manual screening (Recall = 0.95),[30,31] while costing less than $0.01 per abstract and taking less than a second. In our previous meta-analysis, we showed that original data detection can reduce near 70% of all publications meeting Web of Science search criteria for WBE literature,[7] thus the automated screening method here can substantially speed up the meta-analysis process. Notable prevention of nonsensical answers was detected in the fine-tuned classifiers based on davinci and ada, making fine-tuning an attractive option even though the completion models also generate excellent performances. When more advanced questions are involved, i.e., the detection of relevant sampling locations from the Abstract, the GPT model performances were moderately high, with the highest Recall rendered by 0.92 among all configurations tested), indicating a necessity for human intervention and continued domain-specific model evaluation as more advanced LLMs become available. While developed using WBE literature, the automated workflow and the model evaluation criteria can also be applied to facilitate research synthesis in other research fields.

4.2. Discussion

4.2.1. Text Embedding Analysis. Embeddings are numeric representations of text strings that a certain AI system uses to represent given input texts. Upon acquiring excellent performance in the Abstract screening task, we asked: are there underlying differences in GPT's embeddings of the Abstracts from papers that contained original data and those that did not? Answering this question is useful for understanding the excellent performance in the Abstract screening task, facilitating the incorporation of GPT models in future research synthesis tasks. We obtained the embeddings of all Abstracts (including fine-tuning and test sets) using OpenAI's "text-embedding-ada-002" model. Then, we performed multivariate analysis using t-SNE (t-distributed stochastic neighbor embedding) analysis. We detected a clear separation between the papers that contained original data and those that did not (Figure 3). This result suggests that there are fundamental differences between the semantics used in Abstracts for papers that contain or do not contain original data, which underlies the success of the screening task.

4.2.2. Time and Cost Considerations. The automated literature screening method saves time. The time required for model training and response for each request was on the order of seconds. Once the pipeline is set up, the only time-consuming part is queuing in model fine-tuning, which occurs after the request is submitted to the OpenAI server. In our experience, the wait time was approximately 10 min. Thus, the total time required for the GPT model-aided screening of 50 papers will be slightly more than 10 min. In comparison, the time required for human screening can be longer and more complex. Because of the domain knowledge involved, screening Abstracts and full texts requires the efforts of trained graduate students or postdoctoral researchers. In our experience, determining whether the original data were reported took 5−10 min per paper. This task can be laborious; thus, break times need to be planned, with the amount varying based on individual attention spans. We anticipate 250−500 min for screening 50 abstracts. With respect to cost, we find that the use of GPT models for paper screening is quite manageable. For example, gpt-4 and gpt-3.5-turbo cost approximately \$1 and \$0.07 per 100 Abstracts screened, respectively. Using fine-tuned davinci and fine-tuned ada costs \$3.63 and \$0.04 per 100 Abstracts, respectively. Owing to longer input texts, using the texts of Methods as inputs increased the cost for all models; however, even when the longer texts (Methods sections) and the costliest method (davinci) were utilized, the normalized cost for screening 100 papers was still below \$15, which is close to the minimal hourly rate at the place where the study was conducted. Additional details on cost and time can be found in Supplementary Results A and Tables S7−S8.

4.2.3. Limitations and Future Studies. OpenAI models are subject to updates and upgrades, which can affect the parameter specifications for a specific task. However, our approach provides a framework for problem formulation and validation in literature screening, paving the way for more applications of large language models in the tasks of automated systematic reviews. We acknowledge that there are potential biases with the paper inclusion procedures, such as biases in the Web of Science results and the focus on English-language literature. Nevertheless, the Web of Science remains a widely used tool in the environmental engineering field; thus, the procedure here has the benefit of reproducibility.

While the paper screening models from this study could provide accuracies as high as 93%, they were not 100% in agreement with screenings performed by professional researchers. On the other hand, it is worth noting that human processing of paper screening tasks may involve individual biases in assessments, thus presenting challenges in reproducibility. In addition, in our experience of manually processing the papers, we noted that clarity levels in published studies varied, with some studies more difficult to assess for their inclusion in systematic reviews even for experienced researchers. Structured data deposits utilizing reporting standards can improve the ease of reuse of published studies, regardless of manual or model-assisted screening.

One of the application areas of GPT-assisted research synthesis is direct data extraction and modeling. While successful applications have been reported for other types of experiments,[16] in our preliminary study (data not shown), the performance of GPT in extracting data from reports of WBE experiments was unsatisfactory despite extensive prompt engineering. This is not an isolated case; automated data extraction using GPT models has been reported as challenging elsewhere.[32] We reason that research fields with more structured and controlled semantics will be more easily integrated into automated literature synthesis, and for emerging fields, human intervention in data extraction is still necessary at the current stage. Nevertheless, with the rapid development of large language models, future iterations of those models may generate more fruitful outcomes.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsenvironau.4c00042.

> Supplementary methods for fine-tuning models; supplementary results on time and cost; examples of nonsensical answers; text/chat completion model performances in original data detection with Abstracts as inputs; text/chat completion model performances in original data detection with Methods as inputs; fine-tuned model performances in original data detection; text/chat model performances in relevant location detection; fine-tuned model performances in relevant location detection; costs and usage of fine-tuned models; costs of text/chat models; schematic of experimental design; schematic of query scenarios for text/chat models; bar plot of fine-tuned model performances in original data detection; bar pot of fine-tuned model performance in relevant location detection (PDF)

AUTHOR INFORMATION

Corresponding Author

Fangqiong Ling − *Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States;* orcid.org/0000-0003-1546-5647; Email: fangqiong@wustl.edu

Authors

Kaseba Chibwe − *Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*

David Mantilla-Calderon − *Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsenvironau.4c00042

Author Contributions

CRediT: **Kaseba Chibwe** writing - original draft, writing - review & editing; **David Mantilla-Calderon** conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing - original draft, writing - review & editing; **Fangqiong Ling** conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing - original draft, writing - review & editing.

Notes

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation CAREER Award (2047470, F. L.) and the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE 2139839, K.C.).

REFERENCES

(1) Lorenzo, M.; Picó, Y. Wastewater-Based Epidemiology: Current Status and Future Prospects. *Current Opinion in Environmental Science & Health* **2019**, *9*, 77−84.

(2) Haidich, A. B. Meta-Analysis in Medical Research. *Hippokratia* **2010**, *14* (Suppl 1), 29−37.

(3) Stroup, D. F.; Berlin, J. A.; Morton, S. C.; Olkin, I.; Williamson, G. D.; Rennie, D.; Moher, D.; Becker, B. J.; Sipe, T. A.; Thacker, S. B. for the Meta-analysis Of Observational Studies in Epidemiology (MOOSE) Group. Meta-Analysis of Observational Studies in EpidemiologyA Proposal for Reporting. *JAMA* **2000**, *283* (15), 2008−2012.

(4) Diamond, M. B.; Keshaviah, A.; Bento, A. I.; Conroy-Ben, O.; Driver, E. M.; Ensor, K. B.; Halden, R. U.; Hopkins, L. P.; Kuhn, K. G.; Moe, C. L.; Rouchka, E. C.; Smith, T.; Stevenson, B. S.; Susswein, Z.; Vogel, J. R.; Wolfe, M. K.; Stadler, L. B.; Scarpino, S. V. Wastewater Surveillance of Pathogens Can Inform Public Health Responses. *Nat. Med.* **2022**, *28* (10), 1992−1995.

(5) Singer, A. C.; Thompson, J. R.; Filho, C. R. M.; Street, R.; Li, X.; Castiglioni, S.; Thomas, K. V. A World of Wastewater-Based Epidemiology. *Nat. Water* **2023**, *1* (5), 408−415.

(6) Safford, H. R.; Shapiro, K.; Bischel, H. N. Wastewater Analysis Can Be a Powerful Public Health Tool—If It's Done Sensibly. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119* (6), No. e2119600119.

(7) Mantilla-Calderon, D.; Huang, K.; Li, A.; Chibwe, K.; Yu, X.; Ye, Y.; Liu, L.; Ling, F. Emerging Investigator Series: Meta-Analyses on SARS-CoV-2 Viral RNA Levels in Wastewater and Their Correlations to Epidemiological Indicators. *Environ. Sci.: Water Res. Technol.* **2022**, *8* (7), 1391−1407.

(8) Li, X.; Zhang, S.; Sherchan, S.; Orive, G.; Lertxundi, U.; Haramoto, E.; Honda, R.; Kumar, M.; Arora, S.; Kitajima, M.; Jiang, G. Correlation between SARS-CoV-2 RNA Concentration in Wastewater and COVID-19 Cases in Community: A Systematic Review and Meta-Analysis. *J. Hazard Mater.* **2023**, *441*, No. 129848.

(9) Gurevitch, J.; Koricheva, J.; Nakagawa, S.; Stewart, G. Meta-Analysis and the Science of Research Synthesis. *Nature* **2018**, *555* (7695), 175−182.

(10) Sampson, M.; Tetzlaff, J.; Urquhart, C. Precision of Healthcare Systematic Review Searches in a Cross-Sectional Sample. *Research Synthesis Methods* **2011**, *2* (2), 119−125.

(11) Haddaway, N. R.; Westgate, M. J. Predicting the Time Needed for Environmental Systematic Reviews and Systematic Maps. *Conservation Biology* **2019**, *33* (2), 434−443.

(12) Yew, A. N. J.; Schraagen, M.; Otte, W. M.; van Diessen, E. Transforming Epilepsy Research: A Systematic Review on Natural Language Processing Applications. *Epilepsia* **2023**, *64* (2), 292−305.

(13) Aziz, S.; Dowling, M.; Hammami, H.; Piepenbrink, A. Machine Learning in Finance: A Topic Modeling Approach. *European Financial Management* **2022**, *28* (3), 744−770.

(14) Aum, S.; Choe, S. srBERT: Automatic Article Classification Model for Systematic Review Using BERT. *Systematic Reviews* **2021**, *10* (1), 285.

(15) Tsafnat, G.; Glasziou, P.; Karystianis, G.; Coiera, E. Automated Screening of Research Studies for Systematic Reviews Using Study Characteristics. *Systematic Reviews* **2018**, *7* (1), 64.

(16) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048−18062.

(17) Dagdelen, J.; Dunn, A.; Lee, S.; Walker, N.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured Information Extraction from Scientific Text with Large Language Models. *Nat. Commun.* **2024**, *15* (1), 1418.

(18) MacDonell, S.; Shepperd, M.; Kitchenham, B.; Mendes, E. How Reliable Are Systematic Reviews in Empirical Software Engineering? *IEEE Transactions on Software Engineering* **2010**, *36* (5), 676−687.

(19) Petersen, K.; Ali, N. B. Identifying Strategies for Study Selection in Systematic Reviews and Maps. *2011 International Symposium on Empirical Software Engineering and Measurement* **2011**, 351−354.

(20) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2020; Vol. 33, pp 1877−1901.

(21) OpenAI. GPT-4 Technical Report. *arXiv* **2023**, DOI: 10.48550/arXiv.2303.08774.

(22) Zhang, L.; Chen, L.; Yu, X.; Duvallet, C.; Isazadeh, S.; Dai, C.; Park, S.; Frois-Moniz, K.; Duarte, F.; Ratti, C.; Alm, E. J.; Ling, F. MicrobiomeCensus Estimates Human Population Sizes from Wastewater Samples Based on Inter-Individual Variability in Gut Microbiomes. *PLOS Computational Biology* **2022**, *18* (9), No. e1010472.

(23) Achakulvisut, T. *SciPDF Parser*, 2023. https://github.com/titipata/scipdf_parser (accessed 2023−06−01).

(24) *OpenAI Platform* https://platform.openai.com (accessed 2023−07−13).

(25) The Pandas Development Team *Pandas-Dev/Pandas: Pandas*, 2020. DOI: 10.5281/zenodo.3509134.

(26) *scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation*. https://scikit-learn.org/stable/ (accessed 2023−07−13).

(27) OpenAI. *OpenAI model overview*. https://platform.openai.com/docs/models/overview (accessed 2023−08−04).

(28) OpenAI. *OpenAI Language Model Pricing*. https://openai.com/pricing#language-models (accessed 2023−08−04).

(29) Tamkin, A.; Brundage, M.; Clark, J.; Ganguli, D. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. *arXiv* **2021**, DOI: 10.48550/arXiv.2102.02503.

(30) Feng, Y.; Liang, S.; Zhang, Y.; Chen, S.; Wang, Q.; Huang, T.; Sun, F.; Liu, X.; Zhu, H.; Pan, H. Automated Medical Literature Screening Using Artificial Intelligence: A Systematic Review and Meta-Analysis. *Journal of the American Medical Informatics Association* **2022**, *29* (8), 1425−1432.

(31) Cohen, A. M.; Hersh, W. R.; Peterson, K.; Yen, P.-Y. Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. *Journal of the American Medical Informatics Association* **2006**, *13* (2), 206−219.

(32) Xiao, Z.; Li, W.; Moon, H.; Roell, G. W.; Chen, Y.; Tang, Y. J. Generative Artificial Intelligence GPT-4 Accelerates Knowledge Mining and Machine Learning for Synthetic Biology. *ACS Synth. Biol.* **2023**, *12* (10), 2973−2982.