ARTICLE

Using AI to Auto-Tag Graduate Theses

Kyle Morgan

ABSTRACT

This article presents a practical approach to using artificial intelligence (AI) for tagging graduate theses in an institutional repository with the United Nations Sustainable Development Goals. Utilizing strategies requiring no prior programming experience, the article provides a step-by-step guide, cost analysis, and lessons learned from employing two AI-based tagging methods. These methods, attempted with varying degrees of success, highlight the real potential of using AI for the thematic tagging of digital library resources.

INTRODUCTION

Institutional repositories develop over time across eras of varying managerial involvement, staffing, resources, and administrative priorities, leading to the dated and uneven tagging of content. Artificial intelligence (AI) platforms hold the promise of automated workflows to update repository metadata, improve uniformity, and make the content more discoverable to the understandings, vocabularies, and priorities of current users. Building on the success of Portland State University Library to highlight social and environmental justice content in its institutional repository, this article explores whether AI platforms can be utilized to support such endeavors through the automated tagging of content with the United Nations Sustainable Development Goals (SDGs) at Cal Poly Humboldt.[1]

Tagging institutional repositories' content with relevant SDGs would promote social and environmental justice efforts and enhance their discoverability. It would highlight the valuable work of campus researchers, more easily connect student interests with those works, and better communicate the ties between campus research and real-world impact. While the technical complexity of AI platforms and coding can be intimidating, this article demonstrates how AI can be leveraged to navigate the effort with limited or no programming experience. This article shares insights on strategies, costs, and practical considerations for librarians interested in pursuing AI-driven SDG tagging or for those intrigued about its potential for other tagging applications.

DEFINITIONS

The following terms will be useful in understanding this project.

- API (Application Programming Interface): This mechanism allows software applications to communicate and exchange data. For this project, I utilized the Google Sheets API to exchange data with GPT (Generative Pre-trained Transformer) and thereby apply the SDG tags directly to the spreadsheet.
- BERT (Bidirectional Encoder Representations from Transformers): This free natural language processor specializes in understanding the context of words, making it a good candidate for text classification projects.

About the Author

Kyle Morgan (kem8@humboldt.edu) is Scholarly Communications and Digital Scholarship Librarian, Cal Poly Humboldt.

Submitted: 6 April 2025. Accepted for Publication: 9 October 2025. Published: 15 December 2025.

https://doi.org/10.5860/ital.v44i4.17381

- Google Colab: This free browser-based tool allows users to implement code, in this case Python code, to activate an AI project.
- GPT (Generative Pre-trained Transformer): This is a popular AI language model capable of content generation. ChatGPT is a specialized version of GPT for conversational use.
- OpenAI: This is the AI research organization that developed GPT. An OpenAI account is required to use GPT with Google Sheets.
- Python: This is a common programming language.
- Semantic processing: The ability to comprehend the meaning of text is the foundation of natural language processing.
- UN SDGs (United Nations Sustainable Development Goals): The seventeen global social and environmental goals were established by the United Nations to provide a blueprint toward global peace and prosperity.[2]

LITERATURE REVIEW

A systematic study of the articles published from 2019 to 2023 regarding AI-automated subject indexing in libraries found that the AI tagging did not match the quality of human cataloging.[3] While automated AI tagging/classifying/indexing was not accurate enough to work alone, studies of applications in libraries or other repositories of bibliographic content found that AI could serve as a supportive service and might be improved with larger training models, application tweaks, and/or advances in AI.[4]

A 2022 comparison of UN SDG AI classification models demonstrated that an accuracy score of roughly 80% was achievable for the kind of tagging envisioned by this paper's project, showing the promise of AI-based tagging utilizing smaller controlled vocabularies.[5] Although individual studies also found success in developing AI models for the application, much like the generalized tagging projects, authors repeated the views that the projects would benefit from more analysis, larger training sets, application tweaks, and/or advances in AI.[6]

The projects detailed in these articles either used older versions of ChatGPT or BERT or involved technological knowledge and capacity beyond my skill set.[7] I speculated that advances in AI might improve the results enough to be a functional solution for a neophyte practitioner utilizing a limited vocabulary like the UN SDGs. Zavalin's project using Gemini and ChatGPT4 for metadata generation speaks to the promise and limitations of more advanced AI releases.[8]

PROJECT DESCRIPTION

This project aimed to assign UN SDGs to a sample of 788 graduate student theses through an AI analysis of the thesis metadata. Of those, I manually tagged 250 theses to establish a benchmark for AI accuracy.[9] The project explored two primary methods to implement the AI tagging: (1) a classification system using Google Colab and Python code workflows, and (2) an integration of GPT analysis directly through Google Sheets.

I launched this project with a $50 budget and hardly any prior knowledge of AI platforms and Python code. I chose ChatGPT-4o, the most widely used AI tool, as my project guide. So as not to overwhelm the narrative with technical details, specifics about workflows have been relegated to the endnotes for those wishing to replicate this project.

METHOD 1: CHATGPT AND THE BERT MODEL IN GOOGLE COLAB

Initial Attempts with ChatGPT

The first attempt at AI coding involved uploading Excel spreadsheets with each thesis's title, abstract, discipline, and keyword metadata directly to ChatGPT-4o. However, no matter how I formatted the Excel file, I incurred errors through either the input or output exchange processes. When a random .csv file made it through the process, the SDG numbers were so inaccurate as to indicate a misalignment between the article metadata fields and the SDG designations.[10] Eventually, I successfully loaded small batches of thesis metadata and received the SDG tags through bit-sized chunks in the conversation thread. Although there were issues with the tagging, the fifteen theses that were tagged registered an 88% accuracy rate.[11] Even the errors were not so much wrong as merely different interpretations of the nuanced SDG descriptions. Although the theses it tagged were of the less challenging variety and the workflow would not function for any but the smallest of tagging projects, it seeded the promise of larger batching success.

BERT

With the hope of applying this high level of tagging accuracy to a spreadsheet of the metadata from 788 theses, ChatGPT recommended BERT to process the metadata because (1) it was free and (2) it seemed a good match for the kind of natural language processing the project needed.[12] I employed BERT through a Google Colab notebook, another free platform, to operationalize the workflow through Python code.

As intimidating as this might sound, what it meant in practice was simple. ChatGPT would produce code, and I would copy and paste it into the Colab notebook, unfamiliar with the coding language or even generally how it was functioning. When I received errors or inaccurate results, I would paste the information into ChatGPT where it could analyze the issue and tweak the code for me to paste back into the notebook. Except for the monthly subscription for ChatGPT, this approach allowed me to tackle the SDG tagging using otherwise free services and to iteratively improve performance.

Challenges

When I uploaded the spreadsheet with the 250 manually tagged theses to train the BERT model, I received repeated errors related to its formatting. After hours of fixing the spreadsheet so it could finally be processed, the model yielded inconsistent results.[13] I experimented with what ChatGPT called a hybrid approach, which meant supplementing the BERT-model assessment with a semantic similarity method to fill gaps and refine the analysis. The model improved, and after introducing expanded definitions of the SDGs into the code, I achieved a tagging accuracy of 60%. ChatGPT helped me fine-tune the thresholds, which control how confident the model must be to apply a tag, but the accuracy did not improve. The hybrid BERT/semantic similarity model seemed unable to gather the full context of the articles and over-applied SDG tags to articles where they were not relevant. In addition, the fine-tuning required significant computing resources and often resulted in Colab crashes, requiring the insertion of code to prevent exhausting the limitations of Colab's free platform.

Ultimately, the BERT-based model could not capture the diversity of the thesis content or the complexity of the SDG themes from such a limited number of pre-tagged articles. The model was not specialized for this content and would require a much larger dataset for it to learn. ChatGPT recommended replacing BERT with a model called SPECTOR with GPT, hoping that its specialized focus on the semantic understanding of research papers might improve the accuracy of the

tagging. However, even with continual fine-tuning, I continued to experience the same issues as with BERT and never exceeded the 60% accuracy level.

METHOD 2: AI INTEGRATION IN GOOGLE SHEETS USING GPT

Initial Attempts

Although integrating AI into Google Sheets promised more robust AI processing and the ability to deliver SDG tags directly into an active Google spreadsheet, it required hours for the upfront setup, including setting up an OpenAI account.[14] Also, instead of applying the Python code through Google Colab, ChatGPT recommended using a Linux terminal already available on my computer. Once operational, I immediately ran into restrictions with the free OpenAI account that prevented much of the tagging. ChatGPT recommended workarounds such as staggered processing times to circumvent the processing restrictions, but these proved time intensive and impractical. Since my work computer used a Google work environment, I could not upgrade from the free tier and eventually needed to use my personal Chromebook, load my credit card to the OpenAI account, and pay for $10 worth of processing.

Because employing GPT-4o for processing would cost twenty times more than GPT-3.5, I started with GPT-3.5 for the AI analysis. Initial results with GPT-3.5 saw it over-assigning SDGs, especially for metadata covering multiple SDG topics. Introducing SDG definitions did not improve the accuracy as it had in the hybrid BERT-model approach. In fact, I achieved the best results by removing the definitions altogether and relying on simple AI prompts, allowing the model's own internal understanding of the SDGs to inform its selections.[15] I achieved an accuracy rate of roughly 70% with a limited five-thesis metadata sample, a slight improvement but still with enough inexplicable tagging errors to sow doubt about its practical potential. ChatGPT recommended training a custom version of GPT-3.5 with the 250 manually tagged training sample; however, the earlier issues with BERT not successfully learning from this limited sample size made me skeptical that the invested time would achieve the high accuracy the initial Excel upload into ChatGPT-4o promised.

GPT-4o

Switching to GPT-4o significantly improved the comprehension of the content and accuracy of the SDG tags. Through the simplification of the AI prompts, I further improved the accuracy, relying on GPT-4o's own enhanced understanding of the metadata and SDGs. Hopeful about the progress, I expanded the test set to 40 theses and reviewed each tag. The AI applied 121 SDG tags, 22 of which I judged to be erroneous, a roughly 80% accuracy rate. Across that same sample, I had applied 78 SDG tags, 17 of which, upon reevaluation in light of AI's tagging, I also judged to be erroneous.

Although this might appear as if I and GPT-4o made a comparable number of errors, the errors emanated from different causes. Eighteen of the 22 erroneous GPT-4o tags related to four SDGs, two that it had overapplied and two that it underapplied. This demonstrated a misalignment with how I interpreted the SDGs versus GPT-4o, a difference that could be mitigated through customized training. On the other hand, my own 17 errors occurred randomly across 13 different SDG applications, with no SDG tag more likely than any other to be flagged. I attributed these errors to a lack of careful reading of the metadata.

In fact, the term "errors" is not even a fair evaluation. I made errors; GPT-4o made consistent choices that reflected a different interpretation of SDG relevance. Training the model should better align GPT with my own interpretations. With its high level of consistency, the resulting accuracy

should be higher than anything I, or any other human being, could accomplish manually ("should" being the operative word). More on this in the Caveats section.

Perfecting the Model

I introduced four new prompts to better direct GPT-4o regarding the application of the four SDG tags it was consistently misapplying, but the result was an increase in the number of errors to 28. Three of the four SDGs that had been targeted improved, but one increased in errors, and two other SDGs that were not even mentioned in the new prompts now had seven errors as opposed to one previously. This demonstrated how small modifications in the prompts can introduce unintended interactions in how the AI applies tags, underlying the difficulty in honing bulk SDG applications through this method.

A more precise application of the SDGs would necessitate developing a customized version of GPT-4o based on the manually tagged theses. Through that process, GPT-4o would learn how I was applying the SDG tags and hone its approach. Because of the diversity of the theses and the nuances of the SDGs, an expanded training dataset of all 788 theses would be the most useful. This approach would require the theses to be first pre-tagged by the non-customized GPT-4o, discrepancies and all, to help catch the manual errors I made in the training dataset. This would ensure that inconsistencies in my manual tagging were identified and corrected before fine-tuning the customized AI.

DISCUSSION

These two approaches highlight the varied paths libraries can take when using AI for thematic tagging. While BERT models offer value for more focused or single-label tasks, GPT-4o's greater capacity for nuanced comprehension proved more effective for multi-label, complex classification. The customized GPT-4o model holds the promise for the greatest accuracy at the best processing cost, although the meticulous manual labor required would not be insignificant.

My $50 project budget included $40 for two months' access to ChatGPT and $10 for testing through the OpenAI paid tier. The processing of all 788 theses would likely incur $150 of additional costs[16] and 35 hours of additional labor.[17] That put this expansion beyond the budget and time constraints of this current endeavor, but within the scope for future development.

However, considering the labor time and costs, running the non-customized GPT-4o with minimal manual intervention should not be dismissed as an option for limited-sized tagging projects. ChatGPT-4o had an accuracy rate on par with my own applied SDG tags, and its tagging was more consistently applied. The interpretation and application of the SDGs did not exactly meet my own perceptions, but they were not far off and might be entirely functional depending on someone's project goals. The price of implementing ChatGPT-4o instead of a customized model for 788 theses would near $2,000 in processing costs; however, should that be the end-goal of the project, that workflow may prove the most time- and cost-effective method.

CAVEATS

ChatGPT opened a world of possibilities that I could never have navigated on my own, but it did come with caveats. ChatGPT often did not recall earlier discussions, even when working in the same communications thread. This created inefficiencies, particularly when testing iterative changes. For example, when working through the BERT-model process, ChatGPT recommended broader definitions for the SDGs. Yet, when it gave me code later in the day to test new tweaks, it either listed only some of the definitions or described them with abbreviated definitions. This kind

of error was easy for me to catch and direct ChatGPT to fix, but when the change was embedded in Python code, the issue was largely invisible to me and caused significant time delays.[18]

I also relied too heavily on ChatGPT to strategize. The BERT model was not delivering the needed tagging accuracy, yet ChatGPT forged ahead with tweaks to the code and probability matching. Only when attempt after attempt yielded no further gains did I realize that it was up to me to make a change in the strategy. Regardless of whether ChatGPT did not recall the accuracy goal of our initial discussions or could only provide me options within the current context, I realized I needed to be more proactive in the process.

ChatGPT also can have consistency issues, supplying different results when run at different times, even with no obvious changes in the prompts. This proved most frustrating in the unsuccessful attempts to emulate the initial ChatGPT-4o tagging success in the Google Sheets–embedded analyses by GPT-4o. The consistency can also vary within the course of a single batch processing, providing uneven results.[19] Because of such examples, manual checks are necessary after every AI process. I also recommend processing first on a limited sample size before applying any code to a larger dataset, even when seemingly minimal changes occurred in the interim.

I never questioned that ChatGPT tried to provide me with what I wanted, but it is ultimately up to the user to evaluate and validate. When writing this article, I asked ChatGPT to summarize some of the processes that it and I went through using the multiple conversation threads it had saved in the platform. It did so without comment, but when I reviewed the text, I realized many details were missing. When I questioned it, it said it could not access the collection of past saved conversation threads. However, instead of disappointing me with that information, it had fulfilled my prompt by writing what it would have anticipated such a process to entail.

CONCLUSION

This project demonstrates the potential for institutional repositories to use AI-driven solutions for thematic tagging, specifically for aligning content with the UN SDGs. For our library, the goal would be to leverage this project to tag a database of 22,000 legacy articles published by university faculty and students. This would require additional rounds of testing, tweaking, and applying the customized GPT-4o at an OpenAI processing cost of roughly $400–$500 and unknown additional labor costs.[20] Training the model on additional tags, such as to identify traditional ecological knowledge content, would further support and promote these important endeavors on campus and in the community.

Institutional repositories are an untapped treasure trove of environmental and social justice content just waiting to be activated. Tagging with UN SDGs holds the promise of promoting this work and better connecting academia to real-world problems and solutions. I hope this article serves to inspire AI-tagging endeavors on other campuses and spawn a collaboration of librarians aimed at enhancing the discoverability and impact of social and environmental justice content in our institutional repositories.

POSTSCRIPT

I brought on computer science student Courtney Rowe as a paid Spring 2025 intern to advance the project. Instead of engaging either of my two methods, she discovered that a public OSDG Community Dataset had recently been developed that had thousands of volunteer-labeled text snippets mapped to SDGs, utilizing a ModernBERT model. She fine-tuned the model on our particular dataset of thesis metadata and achieved a weighted F1 score of approximately 0.85.[21]

While a weighted FI score is not directly comparable to a percentage accuracy, the result was roughly equivalent to what I had achieved using GPT-4o, with the added benefit that the model was free and could be locally hosted. Her approach laid a foundation that, with an expansion of the training dataset and additional fine-tuning, could become the most accurate approach for applying UN SDG tags.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

I used ChatGPT-4o to provide guidance to the workflow and technical content and to evaluate if my words accurately reflected the process. When incorporating any ChatGPT-generated text regarding the more technical details of the work, I reviewed, edited, and fact-checked the content into my own words. I take full responsibility for the content of the publication. The final question I asked ChatGPT was the correct way to reference it in the article.

ENDNOTES

[1] Sherry Buchanan, "Looking at the Past to Change the Future: Showcasing Featured Collections, Building Communities, and Co-creating," *Humboldt Journal of Social Relations* 1, no. 46 (2024): 17–31, https://doi.org/10.55671/0160-4341.1245.

[2] "The 17 Goals," United Nations Department of Economic and Social Affairs, accessed November 4, 2024, http://sdgs.un.org/goals.

[3] Karen Martínez Concha, Fernanda Palacios Zenteno, and Josefa Tello Alfaro, "Use of Artificial Intelligence in Libraries: A Systematic Review, 2019–2023," *South African Journal Libraries & Information Science* 90, no. 2 (2024): 1–13, https://hdl.handle.net/10520/ejc-liasa_v90_n2_a3.

[4] Marit Asula et al., "Kratt: Developing an Automatic Subject Indexing Tool for the National Library of Estonia," *Cataloging & Classification Quarterly* 59, no. 8 (2021): 775–93, https://doi.org/10.1080/01639374.2021.1998283; Jenny Bodenhamer, "The Reliability and Usability of ChatGPT for Library Metadata" (Oklahoma State University, 2023), https://hdl.handle.net/20.500.14446/339626; Charlene Chou and Tony Chu, "An Analysis of BERT (NLP) for Assisted Subject Indexing for Project Gutenberg," *Cataloging & Classification Quarterly* 60, no. 8 (2022): 807–35, https://doi.org/10.1080/01639374.2022.2138666; Eric H. C. Chow, T. J. Kao, and Xiaoli Li, "An Experiment with the Use of ChatGPT for LCSH Subject Assignment on Electronic Theses and Dissertations," *Cataloging & Classification Quarterly* 62, no. 5 (2024): 574–88, https://doi.org/10.1080/01639374.2024.2394516; Amina El Ganadi et al., "Bridging Islamic Knowledge and AI: Inquiring ChatGPT on Possible Categorizations for an Islamic Digital Library," in *CEUR Workshop Proceedings* 3536 (2023): 21–33, https://ceur-ws.org/Vol-3536/03_paper.pdf; Charlie Harper, Anne Kumer, Shelby Stuart, and Evan Meszaros, "AI-Informed Approaches to Metadata Tagging for Improved Resource Discovery," in *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries,* ed. Sandy Hervieux and Amanda Wheatley (Association of College and Research Libraries, 2022), https://alastore.ala.org/content/rise-ai-implications-and-applications-artificial-intelligence-academic-libraries-pil-78; G. Horton, *Implementing an AI-Generated Subject Indexing Tool for Repositories*, presentation at the Fantastic Futures AI4LAM Conference, Vancouver British Columbia, Canada, November 17, 2023, https://archive.org/details/implementing-ai-generated-subject-indexing-tool-for-repositories; Maja Kragelj and Mojca Borštnar, "Automatic Classification of Older Electronic Texts into the Universal Decimal Classification – UDC," *Journal of Documentation* 77, no. 3 (2021): 755–76, https://doi.org/10.1108/JD-06-2020-

0092; Sugabsen Martins, "Artificial Intelligence-Assisted Classification of Library Resources: The Case of Claude AI," *Library Philosophy and Practice* (2024): 1–22, https://digitalcommons.unl.edu/libphilprac/8159.

[5] Roberto Carlos Morales-Hernández, Joaquín Gutiérrez Jagüey, and David Becerra-Alonso, "A Comparison of Multi-label Text Classification Models in Research Articles Labeled with Sustainable Development Goals," *IEEE Access* 10 (2022): 123534–48, https://doi.org/10.1109/ACCESS.2022.3223094.

[6] Jade Eva Guisiano, Raja Chiky, and Jonathas De Mello, "SDG-Meter: A Deep Learning Based Tool for Automatic Text Classification of the Sustainable Development Goals," in *Asian Conference on Intelligent Information and Database Systems*, 259–71 (Springer, 2022), https://doi.org/10.1007/978-3-031-21743-2_21;Dirk U. Wulff, Dominik S. Meier, and Rui Mata, "Using Novel Data and Ensemble Models to Improve Automated Labeling of Sustainable Development Goals," *Sustainability Science* 19, no. 5 (2024): 1773–87, https://doi.org/10.1007/s11625-024-01516-3; Rui Yao, Meilin Tian, Chi-Un Lei, and Dickson K. W. Chiu, "Assigning Multiple Labels of Sustainable Development Goals to Open Educational Resources for Sustainability Education," *Education and Information Technologies* 29, no. 14 (2024): 18477–99, https://doi.org/10.1007/s10639-024-12566-6; Rui Zhang, Maéva Vignes, Ulrich Steiner, and Arthur Zimek, "Matching Research Publications to the United Nations' Sustainable Development Goals by Multi-label-learning with Hierarchical Categories," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2020), 516–25, https://doi.org/10.1109/DSAA49011.2020.00066.

[7] Bodenhamer, "The Reliability and Usability of ChatGPT for Library Metadata"; Chow, Kao, and Li, "An Experiment with the Use of ChatGPT"; El Ganadi et al., "Bridging Islamic Knowledge and AI"; Guisiano, Chiky, and De Mello, "SDG-Meter"; Asula et al., "Kratt"; Harper, Kumer, Stuart, and Meszaros, "AI-Informed Approaches to Metadata Tagging"; Kragelj and Borštnar, "Automatic Classification of Older Electronic Texts"; Morales-Hernández, Gutiérrez Jagüey, and Becerra-Alonso, "A Comparison of Multi-label Text Classification Models"; Wulff, Meier, Mata, "Using Novel Data and Ensemble Models"; Yao Tian, Lei, and Chiu, "Assigning Multiple Labels of Sustainable Development Goals"; Zhang, Vignes, Steiner, and Zimek, "Matching Research Publications to the United Nations' Sustainable Development Goals."

[8] Vyacheslav Zavalin and Oksana L. Zavalina, "Are We There Yet? Evaluation of AI-generated Metadata for Online Information Resources," *Information Research: An International Electronic Journal* 30, no. iConf (2025): 732–40, https://doi.org/10.47989/ir30iConf47215.

[9] The accuracy rates were determined by comparing the results of GPT's application of the SDGs against my own to create a master list of SDG assignments for each set of thesis metadata. This master list was then compared against GPT's results to evaluate accuracy. If the master list had three SDGs applied (e.g.,. 5.gender equality, 8.decent work and economic growth, and 9.reduced inequalities) and GPT had two SDGs, of which one did not match (e.g., 1.no poverty and 8.decent work and economic growth), then that would count as 25% accurate. That is, it got one right, but missed two, and misapplied a third. I would later apply the same method to measure my own accuracy in applying SDG tags.

[10] I uploaded numerous .csv spreadsheet files. Why all failed the input/output exchange except this one, I have no explanation, nor did ChatGPT. Subsequent improvements to ChatGPT in 2025 have made this process a more functional option for smaller datasets

[11] Instead of processing the rows in order, ChatGPT processed six of the first eight listed articles, then processed the other 12 seemingly at random. Five of those processed thesis titles had no match to what was on my original list. ChatGPT could have invented those titles, but their topical similarity to other titles made me think it may have just reworded the metadata in its process to apply the SDG tags more effectively. Nonetheless, because these five titles could not conclusively be linked to existing articles, they were not included in this accuracy count.

[12] BERT uses numerical representations of text to find patterns between the metadata and SDG descriptions.

[13] I eventually collapsed all four metadata fields (title, abstract, discipline, and keyword) into one cell and deleted empty rows in the spreadsheet. Still, ChatGPT had to iteratively embed code to "clean" the spreadsheet in response to reported errors.

[14] Under ChatGPT's guidance, I created an account with OpenAI to get an API key, launched a Google Cloud Platform (GCP) project, enabled the needed APIs, linked the project script to Google Sheets, then operationalized the process through a Linux terminal. However, when completed, the process of ChatGPT generating the Python code for me to paste into the terminal and then report back for further tweaks was the same as before.

[15] ChatGPT said that since it already had a strong internal understanding of the SDGs, additional definitions only caused confusion in the tagging.

[16] Estimating a processing cost of $0.09 per thesis through this model, running GPT-4o across all 788 theses in the dataset would add roughly $70 in costs. After evaluating and correcting those tags, they could be used to train a customized version of GPT. The per article costs of processing through a customized GPT would be significantly less than through the uncustomized GPT, but it would still add roughly $15 more in processing charges. Budgeting an extra $25 to cover processing costs due to testing and troubleshooting would be reasonable, although that could grow significantly should processing an entire batch of 788 thesis return errors. Two months of ChatGPT to cover this additional work would add $40.

[17] Approximating each manual thesis evaluation and SDG application to take one minute, as well as an additional one-minute comparative analysis against GPT-4o's SDG applications to remove errors, simply creating a training dataset of 788 theses would require more than 25 hours of meticulous work. The processing, troubleshooting, code tweaking, and manual reviews would probably necessitate a minimum of 10 additional hours, although that could expand significantly depending on errors and inconsistencies.

[18] This became apparent when ChatGPT worked iteratively with me to clean the spreadsheet for the BERT-model processing and then again with the Google Sheets API integration. I struggled through the Google Sheets cleaning process for more than an hour before I recognized that the BERT model's rabbit hole of error messages was being replicated, and that I simply needed to direct ChatGPT to the earlier solution to move forward. Similarly, when I abandoned BERT's probability matching method to test an entirely semantic approach to the analysis of the

content, I did not recognize when ChatGPT reintroduced probability matching to refine the semantic SDG applications. That is not to say that ChatGPT did not explain each of the changes it was making, but such was the problem of not understanding the code or fundamentally how it functioned.

[19] When working on another article, ChatGPT formatted the first two-thirds of a bibliography to APA7 specifications, then entirely italicized the listings for the remainder, seeming to have lost its earlier formatting capabilities.

[20] If the customized thesis-trained GPT could be seamlessly applied to academic articles of all kinds, I would estimate $350 in processing charges and 10 hours of labor. If it encountered errors, as I suspect it would, the additional training would probably range from $50 to $150 in additional processing charges, with additional time and labor that could easily exceed the approximately 50 hours spent on the first stage of this project.

[21] An F1 score says how well a model balances precision and recall. Specifically, it is the precision score (how many of the results it pulled up are actually right) multiplied by the recall score (how many of the correct results it pulled up out of all the possible correct results), divided by their total added together, and then multiplied by two.