Received: 29 January 2024 | Revised: 11 February 2025 | Accepted: 4 April 2025

DOI: 10.1111/psj.70034

RESEARCH NOTE

PSO PSJ POLICY STUDIES JOURNAL

GPT models for text annotation: An empirical exploration in public policy research

Alexander Churchill[1] | Shamitha Pichika[2] | Chengxin Xu[3] | Ying Liu[4]

[1]Albers School of Business and Economics, Seattle University, Seattle, Washington, USA

[2]Department of Computer Science, Seattle University, Seattle, Washington, USA

[3]School of Public Affairs and Nonprofit Leadership, Seattle University, Seattle, Washington, USA

[4]School of Public Affairs and Administration, Rutgers University, Newark, New Jersey, USA

Correspondence
Chengxin Xu, Seattle University, Seattle, WA, 901 12th Avenue, Seattle, WA 98122 USA.
Email: cxu1@seattleu.edu

Abstract

Text annotation, the practice of labeling text following a predetermined scheme, is essential to qualitative public policy research. Despite its importance, annotating large qualitative data faces challenges of high labor and time costs. Recent developments in large language models (LLMs), specifically models with generative pretrained transformers (GPTs), show a potential approach that may alleviate the burden of manual text annotation. In this report, we first introduce a small sample pretest strategy for researchers to decide whether to use Open AI's GPT models for text annotation. In addition, we test if GPT models can substitute human coders by comparing the results of two GPT models with different prompting strategies against human annotation. Using email messages collected from a national corresponding experiment in the US nursing home market as an example, on average, we demonstrate 86.25% percentage agreement between GPT and human annotations. We also show that GPT models possess context-based limitations. Our report ends with reflections and suggestions for readers who are interested in using GPT models for text annotation.

KEYWORDS

content analysis, GPT, qualitative method

INTRODUCTION

Qualitative data, such as government documents or meeting minutes, is valuable for public policy research. It enables researchers to understand mechanisms of policy outcomes and processes for which quantitative data is usually not available. It serves as the starting point of grounded theory building and hypothesis development, informing future quantitative modeling and analysis (Ritchie et al., 2003). Together with quantitative methods, well-developed mixed method approach with rigorous qualitative strands can add considerable value to empirical policy research "by illuminating the context of and complexities inherent in human behavior and improving our ability to explain findings" (Hendren et al., 2023, p. 469).

An essential step shared by many qualitative methods is text annotation, which labels text following a predetermined scheme. Many analytical approaches such as thematic analysis and the Institutional Grammar approach require researchers to summarize the originally massive data to a higher level for further analysis (Braun & Clarke, 2012; Siddiki et al., 2022). High-quality annotated data is also necessary for machine learning models to train classifiers and evaluate the performance of coding results (Gilardi et al., 2023).

However, text annotation bears one common challenge across disciplines: its financial and time cost (Gilardi et al., 2023; Gray et al., 2023; Siddiki et al., 2022). Annotating qualitative data often involves a team effort to ensure accuracy and to manage workloads. Additionally, manual text annotation is highly time-consuming. This is not only limited to the time spent on annotating but also burdens introduced by the university grant and contract administration (Bozeman & Youtie, 2020).

Previous tools for automatic text annotation are either inefficient or difficult to use. For example, automatic text annotation can be achieved by using some existing software such as Nvivo and Altas.ti. However, previous research reports that researchers use these applications primarily as “electronic filing cabinets” because of their unfriendly interface and inefficient algorithm design (Marathe & Toyama, 2018). Meanwhile, the price of software purchase or subscription is also relatively high. Other machine-learning approaches are also available, for example, the Naïve Bayes classifier (Loftis & Mortensen, 2020) and BERTweet (Nguyen et al., 2020). With these algorithms, researchers can train the machine with their own training data and then automate the annotation for the rest of the dataset. However, applying these machine learning methods requires researchers to have substantive programming knowledge, which creates a steep learning curve for researchers who have no such training.

Recently, the growing popularity of large language models (LLMs) has introduced a new potential tool for automatic text annotation, with affordable cost and lower requirement of coding knowledge. Text annotation challenges one model's capacity for natural language processing (NLP), or the extent to which the model can understand human language. Such capacity is closely related to the amount of training data for the machine. Traditional machine learning models usually require researchers to train the model with their own limited data. In comparison, an LLM is pretrained on enormous amounts of data, which largely improves its performance in NLP tasks, as demonstrated by its more accurate prediction of the next word based on context. Most recent LLMs are transformer-based, including Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT). The transformer-based structure enables LLMs to be trained on dramatically larger amounts of data than previous algorithms and thus enhances their performance in NLP tasks. With such capacity, LLMs are becoming increasingly impactful across many fields.

Take GPT models as examples. Early explorations of the efficacy of GPT models show that GPT models perform satisfyingly in academic tests, demonstrating capacities of explanation, reasoning, memory, and accuracy (Geerling et al., 2023; Sumbal et al., 2024). Furthermore, the GPT-based chatbot, ChatGPT, an application developed by OpenAI, offers a powerful tool even for those who have zero programming background. Recent applications of GPT models and ChatGPT for text annotation also show positive results (Gilardi et al., 2023; Gray et al., 2023; Wang et al., 2021).

This research note has two purposes. Our first purpose was to introduce a small-sample pretest strategy to predict the text annotation performance of GPT models, which can help researchers determine when to use GPT models for text annotation. The strategy will offer researchers a relatively accurate estimation of GPT models' performance on text annotation without coding the entire dataset. Given technical similarities between different LLMs, this approach can also be applied by researchers who are considering using other LLMs for text annotation, such as BERT.

The second purpose was to examine to what extent GPT models can substitute humans in text annotation for qualitative data related to public policies.[1] Specifically, we compare the annotation results of two GPT models with three prompting strategies against human annotations. Although previous examination of GPT models' performance shows promising results, it is not clear if these models can produce quality results in contexts such as public policy research, given the diversity and complexity of qualitative data in this field. Our examination relies on 1528 email messages collected by Xu and Lee (2024) through a field corresponding experiment on US nursing homes.

In the following sections, we first introduce GPT models and their potential for text annotation tasks. Second, we demonstrate how to use GPT models through API services to conduct a small-sample pretest strategy to forecast the performance of GPT coders. Then, we compare human and GPT coders' performance on text annotation, which is followed by the discussion on the benefits and costs of using GPT models for text annotation for public policy research.

GPT MODELS FOR TEXT ANNOTATION

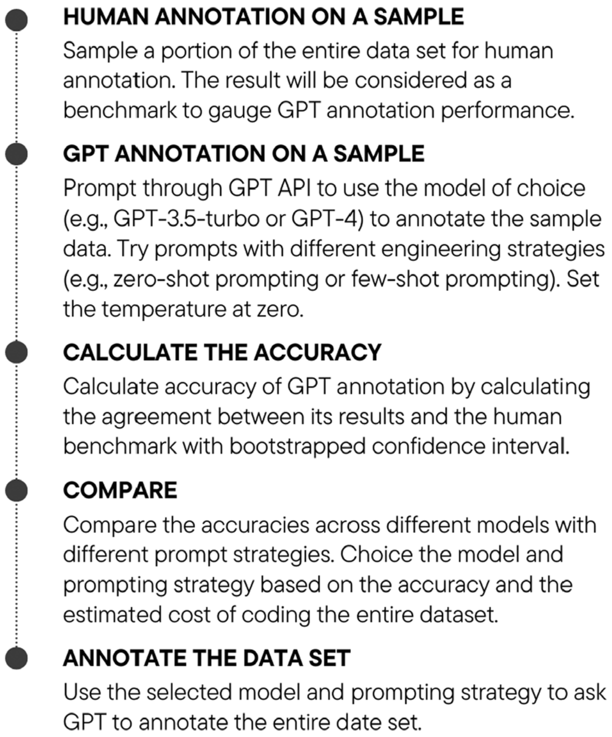
In recent decades, generative AI based LLMs, such as GPT products by OpenAI Inc., Gemini by Google, and LLaMA by Meta, have received attention across diverse fields of academia because of their revolutionary NLP capacity. The key of these LLMs is the transformer architecture, which makes them exceptionally efficient and scalable. Among all these models, GPT models drew the major attention from the market and stimulated competitions among model-developing teams globally.

Since GPT models are pretrained by massive, diverse datasets, their NLP capacity is largely enhanced, making many applications possible. In academia, researchers have found GPT models capable of text summarization, translation, creative writing, developing new measures, generating and predicting data, and so on (Amin et al., 2023; Anastasopoulos & Whitford, 2019; Götz et al., 2024; Lehr et al., 2024; Orwig et al., 2024; Rathje et al., 2024). These days, researchers are still exploring their performance in more specific tasks, such as text annotation. So far, GPT models have been shown to be efficient in annotating various types of texts such as legal documents (Gray et al., 2023), financial news (Azad, 2024), tweets (Gilardi et al., 2023), and academic articles (Toney-Wails et al., 2024). Existing research suggests that OpenAI's GPT-3 model can reduce annotation costs by 50%–90% while maintaining accuracy comparable to human annotation (Wang et al., 2021). With a more advanced model, Gilardi et al. (2023) show that ChatGPT outperforms crowd workers (hired from MTurk) for several annotation tasks. Given its low cost and fast processing time, GPT models offer an efficient alternative to manual coding.

However, the extent to which public policy researchers can use GPT models for text annotation remains uncertain, largely due to the complexity of predicting their performance. In consequence, it is hard for researchers to predict performance for a new task. Although previous literature has suggested some effective strategies to optimize the output (e.g., Hou et al., 2024), whether these strategies can be generalized to policy studies is still an open question. Therefore, a comprehensive examination across models and optimization strategies is necessary. In addition, it is also important to determine when to use GPT models for text annotation at all. However, most recent studies compare GPT and human annotation results, while offering limited guidance about whether to use GPT models for text annotation or not.

USING GPT MODELS FOR TEXT ANNOTATION: A STEP-BY-STEP DEMONSTRATION

In the following sections, we demonstrate the four-step process for annotating text with GPT models using a small sample pretest strategy. We then offer a comprehensive comparison of GPT and human annotation. This demonstration uses the Open AI API (Application Programming Interface) rather than ChatGPT, and thus, the application requires entry-level coding skills in Python or similar programming languages. Figure 1 illustrates the guide.



HUMAN ANNOTATION ON A SAMPLE
Sample a portion of the entire data set for human annotation. The result will be considered as a benchmark to gauge GPT annotation performance.

GPT ANNOTATION ON A SAMPLE
Prompt through GPT API to use the model of choice (e.g., GPT-3.5-turbo or GPT-4) to annotate the sample data. Try prompts with different engineering strategies (e.g., zero-shot prompting or few-shot prompting). Set the temperature at zero.

CALCULATE THE ACCURACY
Calculate accuracy of GPT annotation by calculating the agreement between its results and the human benchmark with bootstrapped confidence interval.

COMPARE
Compare the accuracies across different models with different prompt strategies. Choice the model and prompting strategy based on the accuracy and the estimated cost of coding the entire dataset.

ANNOTATE THE DATA SET
Use the selected model and prompting strategy to ask GPT to annotate the entire date set.

FIGURE 1 Text annotation by GPT API: A step-by-step guide.

Step one: Human annotation

To determine whether to use GPT models for text annotation, we suggest beginning with a small random sample of the whole dataset for a performance pretest. Researchers should first determine the coding scheme for the annotation and then choose a workable size of data to code manually. The sampling strategy should consider the representativeness of the sample to the dataset. This goal may be achieved by simple random sampling; however, for larger datasets such as city council meeting minutes over the years, researchers could use existing covariates to stratify the dataset and then sample randomly. This human annotation needs careful deliberation among the research team, as it will serve as the benchmark to evaluate the performance of GPT models.

Step two: GPT Annotation

After manually coding a random sample of the dataset, researchers can create prompts for GPT models for the annotation task. We recommend using the API service offered by OpenAI for prompting, as it is easier for researchers to choose among a massive number of models with different prices and capacities and to set important parameters for the model. However, if the user has zero programming knowledge, directly prompting ChatGPT can be an alternative (see Gilardi et al., 2023). Although ChatGPT functions similarly, the GPT API offers researchers a wider selection of models. In addition, ChatGPT with GPT-4 only allows 8192 tokens (approximately 5000–6000 English words) for its context window, which may not satisfy researchers' need of processing a large amount of qualitative data in one prompt. In following discussions, we focus on using the API for prompting and annotation.

There are three important factors to consider when prompting GPT models:

Model choice

OpenAI offers access to a variety of GPT models through their API. These models vary in their effectiveness, cost, and context length. At the time of our study, the most widely used models offered were GPT-4 and GPT-3.5-Turbo.[2] Although more recent models are generally more effective, we recommend pretesting various models depending on the use case, as an older model may provide similar performance to a newer one at a lower cost.

The price of each model is determined by the number of *tokens* to be processed, which includes both the *input tokens* (determined by the length of the prompt and the size of the dataset) and *output tokens* (determined by numbers of words in the output). A *token* is a small group of text characters, which is defined as a "non-empty contiguous sequence of graphemes or phonemes in a document" (Mielke et al., 2021, p. 2). Tokens can be a small word such as "is," or part of longer words that are broken into multiple ones. It is reasonable to assume the number of tokens is positively correlated with the number of words or letters to be processed. The cost of annotating a dataset can be calculated beforehand by following the pricing guidelines outlined by OpenAI. At the time of this writing, prices were quoted on OpenAI's website *per 1 million* tokens. For example, the cost for using GPT-4o is 2.50 US dollars per 1 million input tokens (approximately 750,000 words).

GPT models also differ in their *context length*, which is the maximum number of tokens that can be processed at one time. For example, the context length for GPT-4-8k is 8192 tokens, which means it can only process no greater than 8192 tokens at one time, including the prompt and the data. Thus, for a string containing more than 8192 tokens, researchers may want to switch to another model with greater capacity, or to cut the string into pieces and prompt the model multiple times. Thus, researchers should consult with OpenAI or other LLM providers regarding context length before selecting the appropriate model.

Prompt engineering

Prompt engineering is a notion that focuses on strategies for crafting prompts to optimize GPT models' outputs. Accessing LLMs and adjusting parameters through programming languages remains one of the most common ways to interact with LLMs. However, a unique feature of current LLMs like GPT models is their capacity of processing prompts in natural languages. With such capacity, prompt inputs for GPT models do not require researchers to follow a standardized format, wording and contents. Such feature has both pros and cons for researchers. Despite democratizing advanced technologies for researchers with limited programming backgrounds, it also introduces uncertainties, as slight changes in prompt wording can have a big impact on GPT models' outputs (Si et al., 2022).

The prompt for text annotation is recommended to include at least four parts: (1) a short description of the data (e.g., "I am providing emails from a nursing home in response to a potential resident"); (2) the annotation task (i.e., the codebook) (e.g., "If the email indicates that the nursing home has availability, respond with yes"), (3) choices of examples if any, and (4) the data to be annotated.

One technique used to boost the accuracy of GPT annotation is to pass one or more human annotation examples in the prompt. This technique is known as *Few-Shot Learning* (Brown, 2020; Wang et al., 2021). The "Few" refers to the number of examples passed to the model, and a "shot" is a pre-annotated example. For example, for a task asking GPT to code if a statement is positive or negative, researchers can offer some examples of positive and negative statements, and then ask GPT to code other statements. A previous evaluation with GPT-3 shows that one-shot prompts will offer more accurate annotations than a no-shot prompt, whereas a multiple-shot prompt may not outperform one-shot prompts (Wang et al., 2021). As the length of the prompt determines the number of tokens being passed to OpenAI and thus influences the cost of encoding a dataset, researchers may want to optimize cost by balancing prompt length and accuracy. Also, note that the OpenAI API differs from the online ChatGPT interface, which considers previous interactions for future responses. In contrast, OpenAI's GPT API only allows one prompt at a time.

Other model settings

In addition to prompt design, various parameters can be passed to OpenAI models to guide their responses. For most projects, *temperature* and *max tokens* are likely to be useful. *Temperature* indicates the degree of randomness in GPT's responses (Codecademy, n.d.). For reproducibility and results comparison, we recommend setting the temperature to zero so that the model will have the same response to the same prompt each time. *Max token* determines the maximum number of tokens a model can include in its response (Dunn, 2023). For example, if the max token is set at one, then all responses to the prompt will only include one token, such as yes or no. This is useful for projects looking for short codes, such as one-word responses, or for those looking to reduce costs.

Step three: Sample pretest with bootstrap confidence interval

Most previous research evaluates GPT models' performance only after the entire dataset has been labeled by a human coder and a GPT model. For many researchers, this approach may be impossible under their resource constraints. It also completely negates the time-saving potential of GPT models. We suggest overcoming this limitation by evaluating the accuracy of GPT labels against the small sample of human annotated data (see Step two) and then bootstrapping to forecast accuracy for the entire dataset. Bootstrapping is a well-known statistical technique which involves resampling from a sample with replacement and recalculating a statistic of interest from the newly generated pseudo samples (DiCiccio & Efron, 1996). This can be repeated as many times as is computationally feasible to mimic a sampling distribution for that statistic. Bootstrapping a confidence interval using the sample allows researchers to obtain a more accurate estimate of GPT's minimum accuracy for the entire dataset. The accuracy is calculated by percentage agreement between human and GPT coders.

Step four: Multiple pretests comparison

Small-sample pretests with bootstrapping confidence intervals will offer researchers a range of accuracy scores showing the performance of the selected GPT model and prompts being used. This comparison enables researchers to determine: (1) when using a GPT model for text annotation is acceptable; (2) the model for the annotation given its performance, price, and context length; and (3) the prompt strategy (see note 2).

There is no absolute threshold for determining when the GPT models' performance meets acceptable standards. However, we suggest that researchers consider two factors. First, researchers can refer to previous qualitative analyses of similar data types and use their interrater agreements as benchmarks. Second, as the accuracy score is the percentage of agreement between GPT and human annotators, researchers can obtain a rough estimation of the number of inconsistent codes that may require correction. In a large dataset with over 10,000 items, a 90% accuracy means that 1000 items are coded differently by GPT and human annotators. With such estimation, researchers should consider that correcting these discrepancies can be a cost-efficient strategy for their projects.

APPLICATION

Data and research context

The data we use to demonstrate above steps includes 1528 email messages collected from a corresponding experiment on US nursing homes by Xu and Lee (2024), in which they sent out inquiring emails about long-term care facilities and expected responses from targeted nursing homes. The total word count is 108,621 words, with an average of 71 words per email, including greetings and signatures. The purpose of this study was to investigate whether and to what extent Asians and/or noncitizen clients are discriminated against by nursing homes in the United States compared to their white and/or citizen counterparts.

Responses from nursing homes about long-term care services reflect the implementation of important public health policies and programs such as Medicare and Medicaid. These email messages have unique features, as they are less standardized and structured than previously tested data such as legal documents and news. The lack of clarity in emails could pose another challenge for algorithms attempting to systematically apply predetermined codes. In the original paper, Xu and Lee (2024) conducted text annotation with both human coders and GPT models, and the inconsistency was resolved by the authors' judgments. In this examination, we focus on the comparison between first-round annotation results by human coders and GPT models.

Annotation tasks

In the dataset, nursing homes reply to emails regarding inquiries about the availability of long-term care services. Accordingly, we implement the following annotation tasks: (1) *availability*: Whether the nursing home indicated if beds were available for the potential client, categorized as yes, no, waitlist, and not applicable; (2) *citizen flag*: whether or not the nursing home raised concerns about a potential client's citizenship status; (3) *more info*: whether or not the nursing home requested further personal information or documentation from the potential client; (4) *payable*: whether or not the nursing home expressed concerns with the potential client's ability to pay; and (5) *asking call*: whether or not the nursing home asked the potential client to follow up with a phone call.

Small sample human annotation

As suggested in Step One, we start by randomly sampling a small portion from our dataset and ask human coders to annotate the data manually. Our evaluation includes two human annotators: One is a Ph.D. student of public administration (coder 1), and the other is an undergraduate student majoring in computer science (coder 2). Both annotators received the same training for the annotation task. Both coders independently coded the entire dataset, and we randomly selected 200 out of 1528 emails for the pretest.

Model choice

We chose and compared results of two GPT models, GPT-3.5-turbo and GPT-4. We chose these two models since both have sufficiently large context lengths to process all tokens at one time. Technically, GPT-4 is more advanced since it is based on GPT-3.5-turbo, and thus, we expect GPT-4 will outperform GPT-3.5-turbo. However, the price for GPT-3.5-turbo is substantially lower than GPT-4.

Prompt engineering

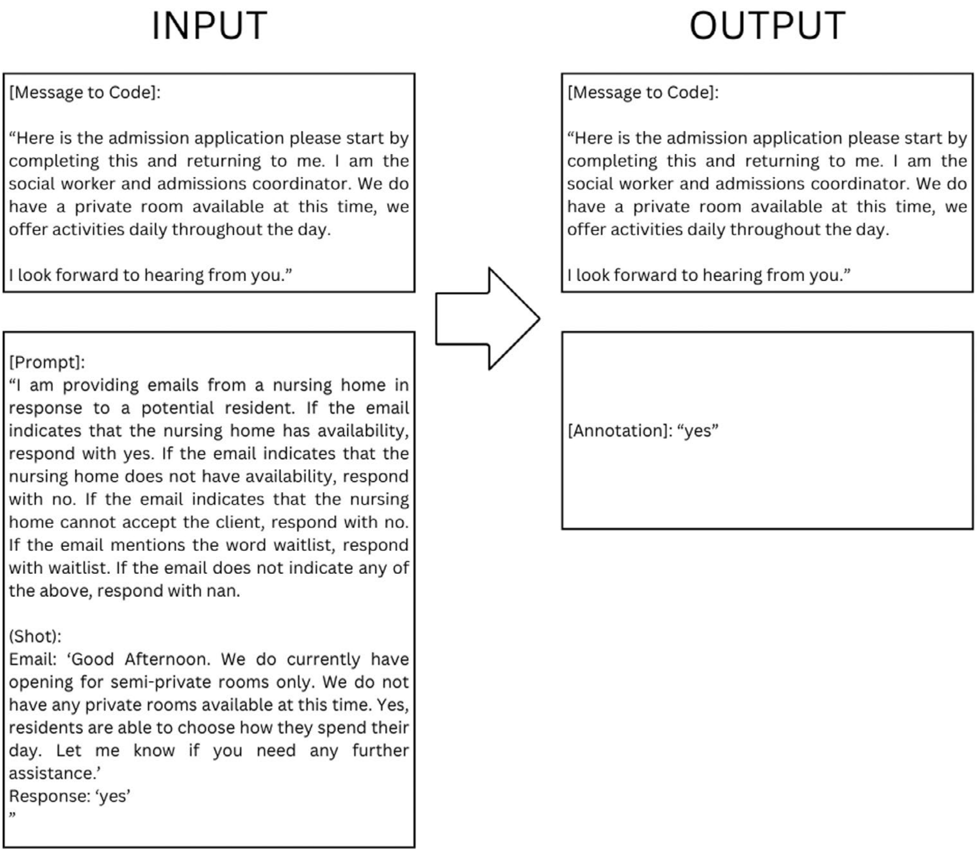
Using both GPT models, we compared results of different prompt strategies, including zero-shot, one-shot, and two-shot prompts. For each annotation task, we used a single prompt with specific settings (model selection is not included in the prompts). Thus, each prompt includes: (1) the task requirement; (2) shots (either zero, one, or two shots are provided); (3) the email to be annotated. The prompt is then sent to the GPT models 1528 times, or once for each email to be labeled. We repeat this process for every model and shot combination, resulting in 15 unique prompts across all models. An example of a one-shot prompt is illustrated in Figure 2.

Other settings

To ensure a fair comparison across all tests, we set the *temperature* to zero. In addition, we limited the *max tokens* of our outputs to one to maintain consistency across annotation results and reduce cost.

Analysis

Before annotating the full sample, we calculated the bootstrapped confidence intervals for each annotation task, model, and prompting strategy using the subsample for the pretest. This step was to determine



[Message to Code]:

"Here is the admission application please start by completing this and returning to me. I am the social worker and admissions coordinator. We do have a private room available at this time, we offer activities daily throughout the day.

I look forward to hearing from you."

[Prompt]:
"I am providing emails from a nursing home in response to a potential resident. If the email indicates that the nursing home has availability, respond with yes. If the email indicates that the nursing home does not have availability, respond with no. If the email indicates that the nursing home cannot accept the client, respond with no. If the email mentions the word waitlist, respond with waitlist. If the email does not indicate any of the above, respond with nan.

(Shot):
Email: 'Good Afternoon. We do currently have opening for semi-private rooms only. We do not have any private rooms available at this time. Yes, residents are able to choose how they spend their day. Let me know if you need any further assistance.'
Response: 'yes'
"

[Message to Code]:

"Here is the admission application please start by completing this and returning to me. I am the social worker and admissions coordinator. We do have a private room available at this time, we offer activities daily throughout the day.

I look forward to hearing from you."

[Annotation]: "yes"

FIGURE 2 Example of a one-shot prompt input and output for annotating availability.

the optimal model choice and prompting strategy for each annotation task. As previously mentioned, after experimenting with different models and prompting strategies, researchers should be able to find the most optimal combination for annotating the full sample. In the following sections, to demonstrate the effectiveness of this approach, we report the full dataset annotation results to demonstrate whether the accuracy scores calculated in the full sample are predicted by the bootstrapped confidence intervals.

To gauge the GPT models' annotation performance, we use coder 1's annotation as a benchmark and calculate GPT accuracy, which is the percentage of annotation agreement between the GPT models and coder 1. To compare it with human coder performance, we also calculate the interrater agreement between two human coders, coders 1 and 2. Since extracting information from emails often requires reading between the lines, we expect a lower interrater agreement between two human coders to indicate highly subjective judgment, in which case GPT models may not perform as effectively.

RESULTS

Small-sample pretest results with bootstrapped confidence interval

Table 1 presents the pretest results and bootstrapped confidence interval based on a sample of 200 emails randomly selected from our dataset. It also includes the annotation accuracies across all models and prompt strategies examined in our full sample test. Across all annotation tasks, we find GPT-4 in general outperforms GPT-3.5-turbo by comparing its annotation accuracy with bootstrapped confidence intervals. For some tasks like labeling *availability*, *citizenship flag*, and *payable*, the performance differences between GPT-3.5-turbo and GPT-4 were not critical. However, for tasks *more info* and *asking call*, GPT-4's annotation accuracy is much higher than GPT-3.5-turbo.

In addition, we find that one- and two-shot prompts do not always outperform zero-shot prompts. For example, when using GPT-3.5-turbo with two-shot prompts for the *more info* task, the two-shot prompt underperformed compared to the one- and zero-shot prompts. Results in Table 1 indicate that the efficacy of the GPT annotator should be evaluated case by case for optimal annotation outputs. However, in general, we may conclude that GPT-4 outperforms GPT-3.5-turbo, and zero-shot prompts may produce results as good as two-shot prompts.

Table 1 also shows that all final accuracy scores fall within the bootstrapped confidence intervals calculated with the small sample ($N=200$). This suggests that the small sample pretest strategy can provide researchers with predictions of GPT models' performance in annotating their own datasets.

Comparing GPT and human annotations

Figure 3 presents GPT annotation performance for all five tasks. In general, we find that GPT models can achieve accuracy levels as high as human coders. The interrater agreement for labeling *availability* is 86.7%. For this task, GPT-4 outperforms GPT-3.5-turbo across all prompts, and zero-shot, one-shot, and two-shot prompts do not produce substantially different results. The highest accuracy score for this task is 88.7%. For tasks of labeling *payable* and *citizenship flag*, both GPT-4 and GPT-3.5-turbo reached levels of accuracy as high as the interrater agreement. Meanwhile, GPT-3.5-turbo with a two-shot prompt in fact underperforms other prompting strategies. For labeling *asking call*, only GPT-4 with a two-shot prompt reached the level of interrater agreement, showing a less ideal performance. The lowest accuracies are found for labeling *more info*. For this task, GPT-4 substantially outperforms GPT-3.5-turbo. Prompting strategies also show a minor difference in this task. Across all tasks, we found that GPT models' accuracy is positively associated with the interrater agreement.

Figure 4 shows the accuracy for each specific label for all tasks. We found that GPT models perform mostly consistently for each label, except GPT-3.5-turbo's performance in labeling, *no availability* prompted by two shots. For other binary labels, we find GPT models' performance is inconsistent for some tasks.

TABLE 1 Sample accuracy and bootstrap confidence interval.

Model	Shots	Small sample accuracy (N=200)	Bootstrapped confidence interval	Full sample accuracy (N=1528)
Task: Availability				
GPT-3.5-Turbo	0	0.86	0.79–0.92	0.81
	1	0.83	0.76–0.9	0.84
	2	0.82	0.74–0.89	0.83
GPT-4	0	0.86	0.78–0.92	0.87
	1	0.87	0.81–0.93	0.89
	2	0.86	0.78–0.92	0.88
Task: Citizen flag				
GPT-3.5-Turbo	0	0.97	0.93–0.99	0.98
	1	0.98	0.95–1.0	0.98
	2	0.96	0.91–0.99	0.95
GPT-4	0	0.99	0.97–0.99	0.99
	1	0.99	0.98–1.0	0.99
	2	0.99	0.98–1.0	0.99
Task: More info				
GPT-3.5-Turbo	0	0.69	0.59–0.79	0.69
	1	0.8	0.71–0.87	0.77
	2	0.56	0.46–0.65	0.56
GPT-4	0	0.79	0.7–0.87	0.82
	1	0.8	0.72–0.87	0.81
	2	0.8	0.72–0.87	0.79
Task: Payable				
GPT-3.5-Turbo	0	0.91	0.85–0.96	0.91
	1	0.9	0.84–0.96	0.90
	2	0.92	0.86–0.97	0.89
GPT-4	0	0.91	0.85–0.96	0.90
	1	0.92	0.85–0.96	0.90
	2	0.91	0.85–0.96	0.90
Task: Asking call				
GPT-3.5-Turbo	0	0.81	0.73–0.88	0.83
	1	0.86	0.79–0.93	0.83
	2	0.78	0.70–0.86	0.77
GPT-4	0	0.88	0.81–0.94	0.87
	1	0.86	0.78–0.92	0.87
	2	0.94	0.89–0.98	0.92

First, we find both GPT-4 and GPT-3.5-turbo failed to identify some *citizenship flags* in the email, whereas interrater agreement between two human coders is higher. Meanwhile, GPT-4 outperforms GPT-3.5-turbo in identifying emails asking *payability* questions. In addition, GPT-3.5-turbo performs well in identifying emails requesting more information, and GPT-4 is more capable of identifying the opposite. Finally, GPT-4 outperforms GPT-3.5-turbo in correctly labeling emails not asking for phone calls from the potential client.

Figure 4 captures a potential pattern of errors made by GPT models regarding Type I Error (false positive), when the model incorrectly identifies a feature or signal (e.g., an email requesting additional information) that is not actually present, and Type II Error (false negative), when the model fails to

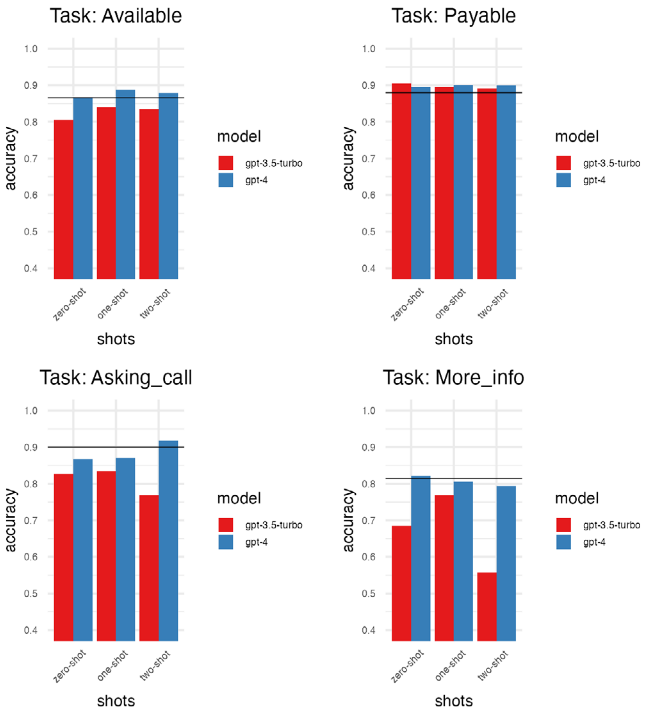


FIGURE 3 Results of text annotation tasks by GPT API. (1) The solid horizontal line indicates the interrater agreement between two human annotators. (2) Accuracy is calculated as the percentage agreement between the GPT and one human annotation.

detect a feature or signal (e.g., an email signaling a concern about citizenship) that is present. By identifying these patterns, researchers can better understand the limitations of GPT models in various contexts and develop strategies—such as targeted human intervention or additional training—to mitigate these errors. In our case, GPT models seem more likely to make Type II (false negative) errors with high-context messages: when an email signals a concern about citizenship, GPT often fails to identify it (as shown in Figure 3, Panel 1). In contrast, GPT models seem more prone to making Type I (false positive) errors with low-context messages: When an email does not request additional information, GPT models incorrectly infer that it does. This pattern aligns with findings from other studies showing that GPT models struggle with understanding high-context messages where subtler cues are present (Binz & Schulz, 2023; Hou et al., 2024; Lehr et al., 2024). On the other hand, in low-context scenarios, GPT tends to overcompensate and detect cues that are not there (Dentella et al., 2023).

DISCUSSION

Qualitative data is an unignorable data source for empirical research in policy research, and text annotation is one essential step to systematically process the data for various analytical approaches. However, text annotation can be costly and time-consuming. In this research note, we propose and show that the small sample pretest with bootstrapped confidence intervals can be an effective tool for predicting GPT models' text annotation performance. Provided with the predicted accuracy scores, researchers would be able to determine whether to use GPT models for certain annotation tasks or not, based on the interrater agreements reported in previous literature with similar data. Given there is no golden standard to determine how high the accuracy score should be acceptable, we encourage researchers to compare the performance of different models and prompts and choose the one that best balances the accuracy and the cost.

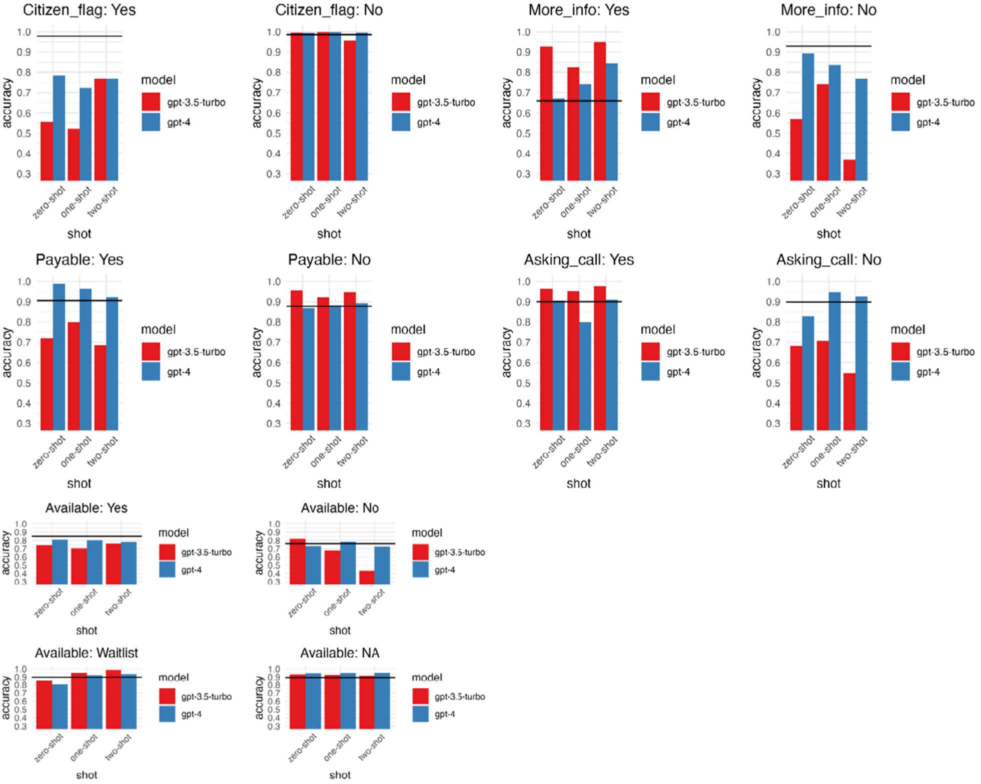


FIGURE 4 Results of text annotation tasks by GPT API by label values. (1) The solid horizontal line indicates the interrater agreement between two human annotators. (2) Accuracy is calculated as the percentage agreement between the GPT and one human annotation.

We then examine whether and how GPT models can offer researchers a cost-efficient approach using the email data collected from a field correspondence experiment. Partly consistent with previous examination, our test shows that GPT models can be as accurate as human text annotators, while their performance might be correlated with the complexity of the task, captured by the interrater agreement between human coders. One important feature of emails is that, compared with other public policy documents, the text is less clear and standardized, and sometimes tends to be implicit. Such features may lead to low interrater agreement on some labels, as the coders' subjective judgment depend on their prior knowledge and characteristics such as age, gender, race and ethnicity, and cultural background might substantially influence the coding results. In that case, it is not clear what assumptions and logic GPT models rely on. Given the source of its training data, GPT models are very likely to be westernized and possess bias (Abid et al., 2021; Bender et al., 2021). In such case, GPT models might serve as a good annotator for text concerning laws, regulations, academic papers, news, government documents, and other data distinguished by its clarity. However, for other data, especially data generated by interpersonal communications, including emails, interviews, focus groups, meeting minutes, and so on, GPT models' annotation performance might be context-based. Researchers need to be cautious about representation and bias issues in GPT annotations, especially in high-context tasks involving high-level cognitive abilities (Binz & Schulz, 2023). In our study, the citizenship flag is not recognized well by GPT models. This could be due to GPT's tendency to represent majority viewpoints, potentially harming marginalized and underrepresented groups (see, e.g., Ghosh & Caliskan, 2023; Lucy & Bamman, 2021; Yang et al., 2024).

For tasks with low interrater agreement, our findings show that GPT-4 outperforms GPT-3.5-turbo by higher annotation accuracy. This result indicates that GPT-4, with its higher cognitive abilities, might be a better model for annotating data that requires logical reasoning and subjective judgment.

However, since GPT-4 is more expansive than GPT-3.5-turbo, researchers might incur higher costs for choosing the more advanced model.

We did not find that the N-shot strategy for prompt engineering led to higher accuracy. Although our results show variations produced by different prompts, the gaps in most cases are moderate. In such cases, using zero-shot prompts might be a more cost-efficient strategy. This finding is contradictory to previous examination of prompt engineering such as Wang et al. (2021). Such discrepancy might be caused by different natures of data, including its shape and context, which creates uncertainties for researchers to choose the most appropriate prompting strategy for their own text annotation task. Thus, we recommend researchers to experiment prompts with zero to multiple shots in small sample pretests to find the one that best serves their purposes and budget.

Limitations of GPT models for text annotation

It is also important to point out several limitations of using GPT models, or other LLMs, for text annotation in general. First, some researchers suggest that computer programs for automated data processing may "interfere with the analysis by creating distance and hindering creativity" (Creswell & Poth, 2017, p. 207). Indeed, heavily relying on GPT models may discourage researchers from digging deeper into qualitative data. In essence, GPT models do not interpret the semantic meaning of data during annotation; instead, they operate as predictive machines that excel at predicting the next word depending on the context. Thus, it is possible that researchers may neglect important insights worth further exploration.

Second, notably, since the training and fine-tuning algorithms of GPT models are not publicly disclosed, it remains unclear how certain pretraining methods might influence their results, underscoring the transparency advantage of self-trained models. In addition, given GPT models and other pretrained LLMs heavily rely on training materials in English, their annotation performance on qualitative data in English might be more acceptable than data in other languages (Cao et al., 2023). Meanwhile, given various ways of tokenization for different languages, the cost of using LLMs for text annotation might also be calculated differently. More exploration and analysis are necessary for evaluating LLMs' capacity and cost of annotating non-English qualitative data.

Third, as GPT models are operated by OpenAI's servers, researchers may have concerns about data privacy and security. OpenAI claims that researchers have the ownership and control over their inputs and outputs from their services, and they will not train the current and future models with researchers' data. But they may run the data through automated content classifiers and safety tools to understand how their services are used. In addition, all data are encrypted at rest (i.e., data in the server, or any data that can be accessed using Internet) and in transit (i.e., data being moved in and out of the server). Despite these privacy measures, we recommend that authors remove any identifiers and confidential information from their data before analysis.

Last, researchers might be particularly interested in understanding any patterns in GPT models' annotations that differ from human annotations. In our study, we have observed several phenomena that merit future investigation. First, GPT models are still limited in inferring high-level context tasks that involve implicit cues and require higher cognitive abilities to understand. For example, when coding for *availability*, emails with sentences like "It sounds like she might be a good fit in our Assisted Living!" "We would be honored to have your mother stay with us long term" were coded as "NA," indicating that the GPT model could not determine if service is available. This limitation may stem from the fact that GPT models we tested have a limited capacity to understand implicit and indirect information; instead, they follow the prompts literally. Second, labeling may require specific experiences and background that GPT models might lack. For instance, messages inquiring whether the prospective patient has a social security number (SSN) were coded by human annotators as "yes" for the *citizen flag* as human coders recognize that asking about an SSN is a common question for foreign-born immigrants (names). In contrast, GPT models marked it as "no." Third, it remains unclear whether the length of the data influences accuracy. For some tasks, the data that

GPT annotated incorrectly tend to be longer than the dataset's average data length, while in the case of the *more info* category, where GPT made the most mistakes, the average message length is shorter than the dataset average. With current data, we are limited to claim any correlations between the length and accuracy.

Indeed, theorizing such patterns of the inconsistency between GPT and human annotation requires more detailed qualitative analyses and additional cases, which exceed the scope of this research note. Meanwhile, any identified pattern may only apply to the specific research context in which it is found. In other contexts, the inconsistencies might not be generalizable, as these errors arise from complex statistical models. However, investigations into these inconsistencies may help researchers refine their prompts and increase accuracy during the small sample pretest stage.

Despite its limitations, the GPT model offers a substantial cost advantage over traditional human text annotation. For example, relying solely on human labor—employing two students at an hourly rate of $18.50 for 70 h each—would cost approximately $3000 for the entire sample. In contrast, our reported workflow, which combines GPT-4 annotation with a pretest method, significantly reduces costs. Specifically, a small sample was annotated by human coders (10 h each, costing $370 in total), and the remaining annotations were completed using GPT-4 with two-shot prompts, incurring a cost of $85.63. Overall, this approach totaled $455.63, representing a cost reduction of 84% compared to the traditional human annotation method.

Limitations and recommendations

Our examination has several limitations that warrant further investigation into the use of GPT models for policy research. First, we relied on nonstandardized email data rather than typical qualitative policy data (e.g., laws, regulations, or policy documents). The ambiguous language in these emails may present unique challenges for current GPT models, suggesting the need to evaluate their performance on more standardized data.

Second, our analysis focused solely on GPT models, while acknowledging the potential of other LLMs (e.g., Claude 2, LLaMA, and Gemini) and alternative text annotation tools such as NVivo and self-trained models like the Naïve Bayes classifier in text annotation tasks. Since our purpose is not to compare different models, and given the rapid development of different models, we cannot claim that GPT models outperform existing automated text processing software. However, we encourage further comparative studies that consider differences in pricing, context length, and specific research needs.

Third, we did not examine the specific patterns by which GPT models' coding decisions diverge from those of human coders. Given that GPT performance is highly context-dependent, identifying generalizable patterns requires a collective effort. One approach is to encourage researchers using GPT for text annotation to pool their inconsistencies and perform a thematic analysis to identify common issues across contexts. Such findings could contribute significantly to the development of a standardized AI-based text annotation tool applicable across various contexts.

Fourth, the annotation task we assigned to GPT models was relatively simple, typically involving binary judgments. In contrast, qualitative data analysis often requires generating topics, summarizing, and categorizing data, which introduces higher complexity and may reduce GPT accuracy. In these cases, researchers should design prompts carefully and conduct frequent comparisons during a small-sample pretest stage. We recommend further testing with more complex text analysis tasks to better understand the capacity and feasibility of GPT models.

CONCLUSION

In this research note, we examined when GPT models (GPT 3.5-turbo and GPT-4) can be considered effective tools for text annotation, potentially providing a more cost-efficient solution for researchers compared to relying solely on human coders for large amounts of qualitative data. Our analysis shows that, on average, GPT-4 may achieve about 86.25% annotation agreement with human coders. To help researchers experiment with GPT models before determining when to use them for annotating an entire dataset, we offer a step-by-step guide for conducting small sample pretests.

FUNDING INFORMATION

The authors receive no funding for this study.

CONFLICT OF INTEREST STATEMENT

The authors report no conflict of interest.

DATA AVAILABILITY STATEMENT

Data and replication files can be found at: https://osf.io/shvrj/?view_only=68fa3f347eaf4dad86039a888572ca55.

ORCID

Chengxin Xu https://orcid.org/0000-0003-4930-9028

Endnotes

[1] Our analysis focuses exclusively on GPT models for three primary reasons. First, at the time of writing, OpenAI's GPT models are among the most widely used large language models (LLMs) on the market and have attracted significant attention from the academic community. Previous studies have demonstrated GPT's capacity for text annotation, suggesting its potential applicability in public policy research. Second, evaluating GPT models benefits researchers without a programming background, as they can access the same models via ChatGPT. Third, although GPT models are not free, their cost is substantially lower than that of hiring student labor for text annotation. By the time this paper was written, GPT-4 has been open to the public for free through ChatGPT.

[2] While prompting strategy is not the focus of this research note, researchers can find more details on improving prompt strategies in Wei et al. (2022), Lee et al. (2024), Mu et al. (2023), and others.

REFERENCES

Abid, A., M. Farooqi, and J. Zou. 2021, July. "Persistent Anti-Muslim Bias in Large Language Models." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 298–306.

Amin, M. M., E. Cambria, and B. W. Schuller. 2023. "Will Affective Computing Emerge from Foundation Models and General Artificial Intelligence? A First Evaluation of ChatGPT." *IEEE Intelligent Systems* 38(2): 15–23.

Anastasopoulos, L. J., and A. B. Whitford. 2019. "Machine Learning for Public Administration Research, with Application to Organizational Reputation." *Journal of Public Administration Research and Theory* 29(3): 491–510.

Azad, S. 2024. "The Effectiveness of GPT-4 as Financial News Annotator Versus Human Annotator in Improving the Accuracy and Performance of Sentiment Analysis." In *Machine Intelligence for Research and Innovations*, Vol 832, edited by O. P. Verma, L. Wang, R. Kumar, and A. Yadav, 105–19. New York: Springer Nature Singapore.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021, March. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623.

Binz, M., and E. Schulz. 2023. "Using Cognitive Psychology to Understand GPT-3." *Proceedings of the National Academy of Sciences* 120(6): e2218523120.

Bozeman, B., and J. Youtie. 2020. "Robotic Bureaucracy: Administrative Burden and Red Tape in University Research." *Public Administration Review* 80(1): 157–62.

Braun, V., and V. Clarke. 2012. "Thematic Analysis." In *APA Handbook of Research Methods in Psychology, Vol. 2. Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, edited by H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher, 57–71. Washington D.C.: American Psychological Association. https://doi.org/10.1037/13620-004.

Brown, T. B. 2020. "Language Models are Few-Shot Learners." *arXiv preprint arXiv:2005.14165*.

Cao, Y., L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. 2023. "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study." *arXiv preprint arXiv:2303.17466*.

Codecademy. n.d. "Intro to OpenAI GPT API." https://www.codecademy.com/learn/intro-to-open-ai-gpt-api/modules/intro-to-open-ai-gpt-api/cheatsheet.

Creswell, J. W., and C. N. Poth. 2017. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Thousand Oaks, CA: Sage.

Dentella, V., F. Günther, and E. Leivada. 2023. “Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias.” *Proceedings of the National Academy of Sciences* 120(51): e2309583120.

DiCiccio, T. J., and B. Efron. 1996. “Bootstrap Confidence Intervals.” *Statistical Science* 11(3): 189–228.

Dunn, Craig. 2023. “OpenAI Tokens and Limits.” https://devblogs.microsoft.com/surface-duo/android-openai-chatgpt-15/.

Geerling, W., G. D. Mateer, J. Wooten, and N. Damodaran. 2023. “ChatGPT Has Aced the Test of Understanding in College Economics: Now What?” *American Economist* 68(2): 233–45.

Ghosh, S., and A. Caliskan. 2023, August. “Chatgpt Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages.” In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal* 901–12.

Gilardi, F., M. Alizadeh, and M. Kubli. 2023. “ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks.” *Proceedings of the National Academy of Sciences* 120(30): e2305016120. https://doi.org/10.1073/pnas.2305016120_.

Götz, F. M., R. Maertens, S. Loomba, and S. van der Linden. 2024. “Let the Algorithm Speak: How to Use Neural Networks for Automatic Item Generation in Psychological Scale Development.” *Psychological Methods* 29(3): 494–518.

Gray, M., J. Savelka, W. Oliver, and K. Ashley. 2023. “Can GPT Alleviate the Burden of Annotation?” In *Legal Knowledge and Information Systems*, edited by G. Sileno, J. Spanakis, and G. van Dijck. Amsterdam, Netherland: IOS Press.

Hendren, K., K. Newcomer, S. K. Pandey, M. Smith, and N. Sumner. 2023. “How Qualitative Research Methods Can be Leveraged to Strengthen Mixed Methods Research in Public Policy and Public Administration?” *Public Administration Review* 83(3): 468–85.

Hou, C., G. Zhu, J. Zheng, L. Zhang, X. Huang, T. Zhong, S. Li, H. Du, and C. L. Ker. 2024. “Prompt-Based and Fine-Tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation.” In *Proceedings of the 14th Learning Analytics and Knowledge Conference, Kyoto, Japan* 518–28.

Lee, K., S. Paci, J. Park, H. Y. You, and S. Zheng. 2024. “Applications of GPT in Political Science Research.” *PS: Political Science and Politics*. https://hyeyoungyou.com/wp-content/uploads/2024/05/gpt_polisci.pdf

Lehr, S. A., A. Caliskan, S. Liyanage, and M. R. Banaji. 2024. “ChatGPT as Research Scientist: Probing GPT's Capabilities as a Research Librarian, Research Ethicist, Data Generator, and Data Predictor.” *Proceedings of the National Academy of Sciences* 121(35): e2404328121.

Loftis, M. W., and P. B. Mortensen. 2020. “Collaborating with the Machines: A Hybrid Method for Classifying Policy Documents.” *Policy Studies Journal* 48(1): 184–206.

Lucy, L., and D. Bamman. 2021, June. “Gender and Representation Bias in GPT-3 Generated Stories.” In *Proceedings of the Third Workshop on Narrative Understanding* 48–55.

Marathe, M., and K. Toyama. 2018. “Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC Canada* 1–12.

Mielke, S. J., Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, et al. 2021. “Between Words and Characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP.” In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2112.10508.

Mu, Y., B. P. Wu, W. Thorne, A. Robinson, N. Aletras, C. Scarton, Kalina Bontcheva, and X. Song. 2023. “Navigating Prompt Complexity for Zero-Shot Classification: A Study of Large Language Models in Computational Social Science.” *arXiv preprint arXiv:2305.14310*.

Nguyen, D. Q., T. Vu, and A. T. Nguyen. 2020. “BERTweet: A Pre-Trained Language Model for English Tweets.” *arXiv preprint arXiv:2005.10200*.

Orwig, W., E. R. Edenbaum, J. D. Greene, and D. L. Schacter. 2024. “The Language of Creativity: Evidence from Humans and Large Language Models.” *Journal of Creative Behavior* 58(1): 128–36.

Rathje, S., D. M. Mirea, I. Sucholutsky, R. Marjieh, C. E. Robertson, and J. J. Van Bavel. 2024. “GPT Is an Effective Tool for Multilingual Psychological Text Analysis.” *Proceedings of the National Academy of Sciences* 121(34): e2308950121.

Ritchie, J., J. Lewis, C. M. Nicholls, and R. Ormston. 2003. *Qualitative Research Practice*, Vol 757. London: Sage.

Si, C., Z. Gan, Z. Yang, S. Wang, J. Wang, J. Boyd-Graber, and L. Wang. 2022. “Prompting GPT-3 To Be Reliable.” In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2210.09150.

Siddiki, S., T. Heikkila, C. M. Weible, R. Pacheco-Vega, D. Carter, C. Curley, Aaron Deslatte, and A. Bennett. 2022. “Institutional Analysis with the Institutional Grammar.” *Policy Studies Journal* 50(2): 315–39.

Sumbal, A., R. Sumbal, and A. Amir. 2024. “Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing.” *Journal of Medical Education and Curricular Development* 11: 23821205241238641.

Toney-Wails, A., C. Schoeberl, and J. Dunham. 2024. “AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications.” In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2403.09097.

Wang, S., Y. Liu, Y. Xu, C. Zhu, and M. Zeng. 2021. “Want To Reduce Labeling Cost? GPT-3 Can Help.” In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2108.13487.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Quoc Le, and D. Zhou. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” In *Advances in Neural Information Processing Systems*, Vol 35 24824–4837. San Diego CA: NeurIPS.

Xu, C., and D. Lee. 2024. “No Country for Model Minorities: Evidence of Discrimination against Asians Noncitizen Immigrants in the U.S. Nursing Home Market.” *Public Administration Review*. Online first.

Yang, Y., X. Liu, Q. Jin, F. Huang, and Z. Lu. 2024. “Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation.” *Communications Medicine* 4(1): 176. https://doi.org/10.1038/s43856-024-00601-z.

AUTHOR BIOGRAPHIES

Alexander Churchill earned the Master of Data Science from the Albers School of Business and Economics at Seattle University and now works as a Casino Marketing Analyst at Holland America Line.

Shamitha Pichika is an undergraduate student of the Department of Computer Science at Seattle University (Class of 2026).

Chengxin Xu is an assistant professor of the Department of Public Affairs and Nonprofit Leadership at Seattle University. His research focuses on decision making that matters to public service equity and effectiveness.

Ying Liu is a PhD candidate at the School of Public Affairs and Administration at Rutgers University-Newark. Her research interests include public and nonprofit management, social equity, local governance, digital government, and accountability.

How to cite this article: Churchill, Alexander, Shamitha Pichika, Chengxin Xu and Ying Liu. 2025. "GPT Models For Text Annotation: An Empirical Exploration in Public Policy Research." *Policy Studies Journal* 00(0): 1–17. https://doi.org/10.1111/psj.70034.