Using Generative Pre-Trained Transformers (GPT) for Supervised Content Encoding: An Application in Corresponding Experiments[1]

Alexander Churchill, Seattle University

Shamitha Pichika, Seattle University

Chengxin Xu, Seattle University*

Abstract

Supervised content encoding applies a given codebook to a larger non-numerical dataset and is central to empirical research in public administration. Not only is it a key analytical approach for qualitative studies, but the method also allows researchers to measure constructs using non-numerical data, which can then be applied to quantitative description and causal inference. Despite its utility, supervised content encoding faces challenges including high cost and low reproducibility. In this report, we test if large language models (LLM), specifically generative pre-trained transformers (GPT), can solve these problems. Using email messages collected from a national corresponding experiment in the U.S. nursing home market as an example, we demonstrate that although we found some disparities between GPT and human coding results, the disagreement is acceptable for certain research design, which makes GPT encoding a potential substitute for human encoders. Practical suggestions for encoding with GPT are provided at the end of the letter.

* Address correspondence to the author at cxu1@seattleu.edu.

Introduction

Qualitative data related to public administration and policy research have offered fruitful empirical evidence that supports theory building and orients practices. One common way to use qualitative data for empirical research is supervised content analysis. Compared to unsupervised content analysis, which is more exploratory, supervised content analysis allows researchers to seek specific information, determined by the coding scheme, from qualitative data (Creswell and Poth, 2017). It also benefits quantitative research by enabling scholars to conduct quantitative analyses in fields with limited numerical data. For example, Yackee (2006) conducted supervised content analysis on administrative rules from four federal agencies and explored the extent to which rulemaking in the executive branch was influenced by formal participation of interest groups. Second, it helps scholars to capture more nuanced administrative behavior. For example, scholars have quantified email contents to demonstrate how politicians and street-level bureaucrats discriminate against clients based on race, gender, and ethnicity (e.g., Olsen, Andersen, & Moynihan, 2022). Third, it allows researchers to better understand causal mechanisms. For example, Liu and Lee (n.d.) used content analysis to analyze the missions of collaborative platforms led by the government, for-profit, and nonprofit organizations to understand the causal mechanism in achieving environmental goals. Thus, supervised content encoding is playing an increasingly important role in public administration research.

Supervised content encoding brings two major challenges, both caused by human participation. First, human encoding has a reliability problem, including intercoder and intracoder reliability (Lacy et al., 2015). Intercoder reliability measures coding consistency between encoders. Although a well-designed coding scheme can help to guide the coding process, the application of the coding scheme may vary across different coders. Intracoder reliability involves a coder's consistency across time. Human coders' interpretation of qualitative data is often subjective even with a codebook. Thus, it is expected that the application of the codebook might be inconsistent over time, especially when analyzing a large amount of data over a long period.

The second challenge of human content encoding is cost. Encoding qualitative data often involves a team effort to ensure reliability and manage workloads. In practice, researchers often rely on student labor to encode qualitative data. Given increasing labor costs, scholars with limited resources might not be able to conduct research involving large data sets. Additionally, human encoding is highly time-consuming. If a scholar's career development is time-sensitive, working with a large data set might be a risky choice.

In this research paper, we explore the use of LLMs, such as GPT, to conduct supervised content encoding. The introduction of generative artificial intelligence (AI) is having a large impact on many industries. Given its accurate understanding of natural language and processing efficiency, we view GPT as a potential solution to the aforementioned challenges of supervised content coding. Although computer-assisted coding software existed before GPT, the unique language training model used by GPT makes it outperform other algorithms when processing natural languages (Zhang et al., 2023). However, LLMs are still black boxes and thus their limitations must be explored through experimentation. Without such experimentation, it is difficult to evaluate GPT's usefulness as a supervised content encoder. Thus, in the following sections, we demonstrate how scholars with limited programming background can use OpenAI's GPT Application Programming Interface (API) to encode qualitative data. To validate its effectiveness, we calculate interrater reliability between human coders and GPT coders. Finally, we compare the results of statistical analysis using human encoded data and GPT encoded data. We discuss best practices when using GPT for supervised content encoding at the end.

Using GPT for Supervised Content Encoding: A Step-by-Step Guide

In recent decades, LLMs have received attention across diverse fields of academia. Their potential for processing natural language data makes LLMs desirable in humanities such as social science, as well as in STEM fields. Existing research suggests that OpenAI's GPT-3 model can reduce annotation costs by 50% to 90% while maintaining accuracy comparable to human annotation (Wang et al., 2021). Additionally, the study demonstrated that the best results were achieved using GPT-3 as an initial encoder and humans as editors.

Despite their promise, GPT models are still highly unpredictable and sensitive to the exact wording of prompts (Changlei et al., 2023). Thus, we recommend scholars think carefully about the following decisions and follow the process outlined below to ensure codes generated by GPT models are reasonable and accurate.

Model Choice

OpenAI offers access to a variety of GPT models through their API. These models vary in their effectiveness and cost. At the time of our study, the most recent models offered were GPT-4, and GPT-3.5-Turbo[2]. Although more recent models will generally be more effective, we recommend trying 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. A token is a small group of text characters. A token is defined as a "non-empty contiguous sequence of graphemes or phonemes in a document" (Mielke et al., 2021, p.2). Tokens can be one small word such as "is" but will break up longer words into a series of tokens. It is reasonable to assume the number of tokens is positively correlated with the number of words or letters to be processed. Additionally, the cost of encoding 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 1,000 tokens (OpenAI, n.d.).

Prompt Engineering

Prompts are text strings containing instructions for GPT models, followed by the data for the model to encode. Since GPT models can follow written instructions, scholars can use natural language to design coding schemes, and the model will follow the scheme for each observation in the data. This feature allows scholars with limited coding knowledge to effectively process qualitative data with advanced NLP. However, it also means that there is no standard prompt for each use case, and the accuracy is likely to vary across different research contexts. Additionally, results are likely to vary between slightly different prompts which human coders might view as communicating the same information. It is important to keep this is mind when writing prompts, as slight changes in wording can have a big impact on the codes generated by OpenAI's models (Changlei et al., 2023).

One technique used to boost the accuracy of GPT encoding is to pass a human encoded example in the prompt. This technique is known as N-shot learning (Wang et al., 2021). The "N" refers to the number of examples passed to the model, and a "shot" is a pre-coded example. 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, scholars may want to optimize cost by balancing prompt length and accuracy. Also, note that the OpenAi API is different from the online ChatGPT interface, which considers previous interactions for future responses. Openai's GPT API only allows one prompt at a time.

The cost of labeling a dataset is determined by the number of observations in a dataset that need to be encoded, the length of each observation, the length of the prompt to be passed to OpenAI, and the length of the response returned by OpenAI. For example, our dataset consisted of 1,614 emails which varied in length. As a result, 1,614 prompts had to be passed to OpenAI, one for each unique email. The prompt consists of base instructions which do not vary between observations, plus the email to be labelled, and OpenAI returns a one token response. As a result, the price of passing a prompt is the cost of the tokens in the base prompt plus the cost of the tokens in the email, plus the cost of the response. Formula (1) outlines the cost calculation for a single prompt, which resulted in close estimates to the amount we were billed by OpenAI, where input tokens is the number of tokens in the prompt plus the number of tokens in an observation. Also note that prices quoted by OpenAI are per 1,000 tokens (OpenAI, n.d.).

$$Price = \left(\frac{InputTokens}{1{,}000} * inputTokenPrice\right) + \left(\frac{outputTokens}{1{,}000} * outputTokenPrice\right) \quad (1)$$

In addition to prompt design, various parameters can be passed to OpenAI models to guide their responses. These parameters vary across models, but the parameters for each are available on OpenAI's website (OpenAi, n.d.). For most projects, *temperature* and *max tokens* are likely to be useful. Temperature indicates the degree of randomness in GPT's responses. For reproducibility, temperature should be set to zero so that the model will have the same response to the same prompt each time. Max tokens determines the maximum number of tokens a model can include in its response. This is important for projects which are looking for short codes, such as one-word responses.

Sample Pretest with Bootstrap Confidence Interval

Most previous research offers only retroactive prompt evaluation strategies. That is, codes are evaluated for accuracy only after the entire dataset has been labeled by a human coder and a GPT model. For many scholars, 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 a small sample of human encoded data and then bootstrapping to forecast accuracy for the entire dataset. With this approach, scholars only need to randomly sample a small portion of the full dataset and manually encode this sample. The sample of human encoded data is used as a performance benchmark for the GPT prompts, calculating accuracy by percentage agreement. Bootstrapping a confidence interval using the sample also allows scholars to ensure that their accuracy threshold is above the lowest expected 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. This can be repeated as many times as is computationally feasibly to mimic a sampling distribution for that statistic. In this case, the accuracy is recalculated for many bootstrapped samples of the original sample with human and GPT encoded data. Once the bootstrap distribution is created, a confidence interval can be calculated to see a range of likely accuracy values when the prompt is applied to the entire dataset. It is critical to remember that bootstrapping rests on the assumption that the sample being bootstrapped is in fact representative of the dataset (DiCiccio and Efron, 1996).

Application in a Corresponding Experiment

To demonstrate how GPT can facilitate supervised content coding for public administration research, we apply this strategy to a previous experiment (Xu and Lee, 2023). We will introduce our model choice, prompt design, and the results of small sample accuracy test with bootstrapping confidence interval. After that, we will validate the pre-test results with a retroactive evaluation, including estimating the interrater reliability and comparing analytical results between human and GPT coded data.

Background of the Corresponding Experiment

Xu and Lee (2023) investigate whether and the extent to which Asians and/or noncitizen clients are discriminated against by nursing homes in the U.S. compared to their white and/or citizen counterparts. This experiment is a specific type of field experiment in which scholars send emails to target organizations with artificial requests and observe how these organizations respond (Bertrand and Duflo, 2017). It is considered an effective way to identify and evaluate race and gender-based discrimination among front-line workers. This method has unearthed important evidence of discrimination in labor markets (Bertrand and Mullainathan, 2004), housing markets (Hanson and Hawley, 2011), healthcare markets (Jilke, Van Dooren, and Rys, 2018), political arenas (Grose, 2014), public education (Pfaff et al., 2021), and so on.

In order to analyze the data collected in this experiment, it was integral that the emails be encoded in order to compare response patterns between groups. Our demonstration included 1,614 response emails from a corresponding experiment conducted in 2022 in the U.S. nursing home market. In the following sections, we evaluate the quality of GPT encoded data using human encoded data as a benchmark. We accomplish this by calculating the interrater reliability score and comparing the analytical results using human and GPT encoded datasets.

Setting

Given the cost and the nature of the data, we chose GPT-3.5-Turbo for GPT encoding. The price estimation and the pretest showed that GPT-3.5-Turbo was more cost effective than other models such as GPT-4 and Davinci-003. The results of the GPT encoding are displayed in Table 1. The prompts used for each category are available in the Supplementary Document.

The categories tested in this experiment were *availability*, *citizen flag*, *more info*, and *payable*. *Availability* refers to whether or not the nursing home indicated if there were beds available for the potential client; *citizen flag* refers to whether the nursing home raised concerns about a potential client's citizen status; *more info* refers to whether or not the nursing home requested further personal information or documentation from the potential client; and *payable* refers to whether or not the nursing home was concerned with the potential client's ability to pay. *Citizen flag*, *more info*, and *payable* were binary categories, while *availability* was encoded with yes, no, NA (if no indication was given either way), or waitlist (if the potential client was waitlisted).

In this demonstration, we use one unique prompt for each category to reduce the complexity of the prompts. As mentioned above, the temperature was set to zero, and the max tokens were set to one. This setup was validated by the sample pretest. As shown in Table 1, the bootstrap confidence interval for each category suggests that our choice of GPT model and prompts can produce acceptable annotation results.

Table 1: Sample Accuracy and Bootstrap Confidence Interval.

Category	Accuracy	Bootstrap Confidence Interval
Availability	0.88	0.81 – 0.94
Citizen Flag	0.95	0.95 – 1.00
More Info	0.85	0.78 – 0.92
Payable	0.85	0.78 – 0.92

Evaluation

Interrater Reliability

To evaluate the accuracy of GPT encoding, we first calculated the simple agreement score in terms of the percentage of agreement between GPT and human coding results. The results reported in Table 2 show that the true accuracies achieved on the entire dataset were correctly forecasted by the bootstrap confidence intervals.

Table 2: Accuracy and Kappa Score on entire dataset

Category	Accuracy	Kappa
Availability	0.87	0.79
Citizen Flag	0.98	0.63
More Info	0.78	0.63
Payable	0.81	0.50

We then calculated the Cohen's Kappa statistic (McHugh, 2012). The Kappa statistic, which ranges from 0 to 1, measures the correlation of encodings created by two independent coders on one dataset. It considers the percent agreement between the coders, but also penalizes the agreement for the likelihood that agreement was achieved through random guessing. An acceptable Cohen's Kappa is subjective and depends on the application. For medical testing, percent agreement and Kappa above 0.9 may be the only acceptable threshold. However, for communication research, Lacy et al. (2015) suggests using 0.8 as the rule of thumb unless the area is truly exploratory. Table 2 includes the Cohen's Kappa statistic between results from a human coder and from the GPT coder. It shows that three out of four of the agreement between GPT-3.5-Turbo and the human coder may be due to chance, whereas the GPT encoding of *availability* is more acceptable.

Comparison of Analytical Results

Recall that the purpose of the corresponding experiment is to gauge the extent to which Asian and/or noncitizen clients of nursing homes are treated differently compared to their White and/or citizen counterparts. Thus, in the following analyses, we focus on the difference-in-proportions of four major outcome variables we encoded depending on race and citizenship status.

Table 3 shows the nonparametric comparison across experimental groups using both GPT-3.5-Turbo and human encoded data. The results are mostly similar to each other and support our hypotheses:

nursing homes are more likely to offer access to white and citizen clients than to Asian and noncitizen counterparts, and more likely to ask for more information from white and citizen clients. Meanwhile, nursing homes are more likely to raise concerns about clients' citizenship when the email discloses the immigration status of the prospective client. The only difference found between the human and GPT encoded data is whether the nursing home raised questions about payment. Specifically, only results from human encoded data show that noncitizens are more likely than citizens to receive concerns about service payment, whereas only those from GPT encoded data show that white clients are more likely to receive such questions than Asian clients.

Table 4 shows the regression results using linear probability models based on both GPT and human encoded data. The comparison shows limited qualitative differences between findings from the two datasets. As predicted, nursing homes are more likely to indicate bed availability in the email to white and citizen clients than Asian and noncitizen ones. Meanwhile, disclosing a noncitizen status will raise concerns from the nursing home. In addition, nursing homes are more likely to ask white and citizen clients for more information than Asian and noncitizen clients. Similar to the nonparametric results, the major difference between results from GPT and human encoded data concerns payment. Specifically, only the result from the GPT encoded data shows that Asians are less likely to receive questions related to payment, whereas only the result from the human encoded data shows that noncitizens are more likely to receive questions about payment. Although the directions of coefficients in these models are the same, their levels of statistical significance differ substantially. Meanwhile, one can observe from Table 4 that although coefficients in models 1 to 3 indicate similar results to those in models 5 to 7, the estimation of coefficients using GPT encoded data are greater in absolute values than those using human encoded data. This result suggests users should be cautious about using GPT for supervised encoding when estimating models for causal identification.

Discussion, Suggestions, and Conclusion

Qualitative data is an unignorable data source for empirical research in public and nonprofit administration. However, using such data can be extremely costly and time-consuming. In addition, qualitative data analysis may involve human-related biases and errors, making results difficult to reproduce. In this study, we examine whether LLMs such as those developed by OpenAI can assist scholars in combatting these challenges in a supervised content encoding context.

Our examination shows acceptable results using GPT models for supervised content encoding. Good results were seen across all tests including small sample pretests using bootstrapped confidence intervals, tests for intercoder reliability, and comparisons between analytical results using experimental data encoded by GPT models and humans. Results from GPT encoding have a high to acceptable level of agreement with human encoding results. In addition, such agreement leads to similar analytical results using GPT and human encoded data. Given that the data comes from a randomized experiment, encoding disagreements and errors are less likely to influence the estimation of the average treatment effect for hypothesis testing. However, it is important to notice that inaccurate encoding will lead to fuzzier estimation of coefficients.

Table 3 Nonparametric Comparison of Descriptive Statistics

	GPT Code					Human Code				
Variable	N	Percent	N	Percent	Test	N	Percent	N	Percent	Test
	Citizen		Noncitizen			Citizen		Noncitizen		
available	3187		3184		X^2=114.952***	3187		3184		X^2=139.247***
... no info	2669	84%	2890	91%		2225	70%	2532	80%	
... no	103	3%	108	3%		562	18%	489	15%	
... waitlist	135	4%	98	3%		102	3%	71	2%	
... yes	280	9%	88	3%		298	9%	92	3%	
moreinfo	3187		3184		X^2=38.639***	3187		3184		X^2=38.639***
... no	2684	84%	2850	90%		2684	84%	2850	90%	
... yes	503	16%	334	10%		503	16%	334	10%	
citizenflag	962		652		X^2=49.401***	959		649		X^2=53.042***
... no	949	99%	595	91%		952	99%	599	92%	
... yes	13	1%	57	9%		7	1%	50	8%	
paymentGPT	962		652		X^2=1.972	958		648		X^2=14.719***
... no	692	72%	447	69%		876	91%	552	85%	
... yes	270	28%	205	31%		82	9%	96	15%	
	White		Asian			White		Asian		
available	3195		3176		X^2=13.25***	3195		3176		X^2=33.703***
... no info	2751	86%	2808	88%		2295	72%	2462	78%	
... no	108	3%	103	3%		593	19%	458	14%	
... waitlist	118	4%	115	4%		81	3%	92	3%	
... yes	218	7%	150	5%		226	7%	164	5%	
moreinfo	3195		3176		X^2=19.643***	3195		3176		X^2=19.643***
... no	2715	85%	2819	89%		2715	85%	2819	89%	
... yes	480	15%	357	11%		480	15%	357	11%	
citizenflag	900		714		X^2=0.727	897		711		X^2=6.943***
... no	857	95%	687	96%		855	95%	696	98%	
... yes	43	5%	27	4%		42	5%	15	2%	

payment	900		714		X^2=5.147**	898		708		X^2=1.628
... no	614	68%	525	74%		790	88%	638	90%	
... yes	286	32%	189	26%		108	12%	70	10%	

Table 4 Results of Linear Probability Regression Models using GPT- and Human-coded Data

	Results based on GPT-coded Data			
	(1) Availability	(2) Citizen Flag	(3) More Info	(4) Payment
Race: Asian	-0.024**	-0.003	-0.044***	-0.049*
	(0.008)	(0.011)	(0.009)	(0.024)
Citizenship: Noncitizen	-0.072***	0.075***	-0.056***	0.014
	(0.008)	(0.012)	(0.009)	(0.025)
(Intercept)	1.195***	1.003***	1.213***	1.265***
	(0.029)	(0.033)	(0.033)	(0.084)
Covariates	Yes	Yes	Yes	Yes
Num.Obs.	5782	1462	5782	1462
R^2	0.027	0.047	0.018	0.009
	Results based on human-coded Data			
	(5) Availability	(6) Citizen Flag	(7) More Info	(8) Payment
Race: Asian	-0.015*	-0.023*	-0.026***	-0.023
	(0.007)	(0.009)	(0.008)	(0.017)
Citizenship: Noncitizen	-0.075***	0.070***	-0.041***	0.051**
	(0.007)	(0.012)	(0.008)	(0.017)
(Intercept)	1.193***	1.033***	1.155***	1.142***
	(0.028)	(0.031)	(0.029)	(0.062)
Covariates	Yes	Yes	Yes	Yes
Num.Obs.	5782	1456	5782	1455
R^2	0.026	0.045	0.011	0.019

Standard errors are reported in parentheses.
+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

It is also important to point out the limitation of using computer programs for supervised content encoding in general. Some scholars suggest that computer programs may "interfere with the analysis by creating distance and hindering creativity" (Creswell and Poth, 2021, p. 207). Indeed, heavily relying on GPT models may discourage scholars from digging deeper into qualitative data. Thus, it is possible that scholars may neglect insights which are not widely observed in the data but still offer important information for further exploration.

Thus, although our study demonstrates positive results, we recommend scholars take the following steps when employing GPT models for supervised content encoding. First, research team members should work closely to establish a coding scheme by manually encoding samples of the data. If the dataset is large, researchers may consider using related covariates of the unit of analysis to purposefully sample the whole dataset to ensure the sample is representative. Second, it is necessary to pretest the model's encoding performance on selected samples by calculating agreement and intercoder reliability. Ideally, researchers should identify disparities between human and GPT coders and try to reconcile disagreements. Third, after addressing disagreements, researchers may need to run multiple rounds of pretests to optimize the prompts,

especially to test the number and types of shots needed to produce reliable results. Fourth, during the pretests and coding, set the temperature parameter to zero so that the coding results can be reproduced.

In conclusion, our examination shows that encoding qualitative data using Openai's GPT API can be an effective way to quicky encode qualitative data for an affordable cost. Additionally, the reproducibility of GPT results compared to human coders makes the coding process more transparent. Therefore, although GPT may not be a perfect substitute for human coders, supervised content encoding with GPT can enhance the rigorousness of qualitative data analysis and substantially reduce the cost. Scholars with limited resources should consider GPT models as a valid alternative to human encoding.

Notes:

1. Data and code for this paper can be found at: https://osf.io/shvrj/?view_only=68fa3f347eaf4dad86039a888572ca55
2. GPT-3.5-Turbo is an updated version of the GPT-3 model, and is cheaper than some GPT-3 models such as Davinci-003 (OpenAI, n.d.). GPT-4 is the latest model released by OpenAI and is more effective than GPT-3.5-Turbo and GPT-3 but is more expensive.

References

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. The American Economic Review, 94(4), 991–1013.

Creswell, J. W., & Poth, C. N. (2017). Qualitative inquiry and research design: Choosing among five approaches. Sage publications.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. Statistical Science, 11(3), 189–228.

Grose, C. R. (2014). Field experimental work on political institutions. Annual Review of Political Science. https://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-072012-174350

Hanson, A., & Hawley, Z. (2011). Do landlords discriminate in the rental housing market? Evidence from an internet field experiment in US cities. Journal of Urban Economics, 70(2), 99–114.

Jilke, S., Van Dooren, W., & Rys, S. (2018). Discrimination and Administrative Burden in Public Service Markets: Does a Public–Private Difference Exist? Journal of Public Administration Research and Theory, 28(3), 423–439.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and Best Practices in Content Analysis. Journalism & Mass Communication Quarterly, 92(4), 791–811.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara / HDMB, 22(3), 276–282.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (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

Olsen, A. L., Kyhse-Andersen, J. H., & Moynihan, D. P. (2022). The unequal distribution of opportunity: A national audit study of bureaucratic discrimination in primary school access. American Journal of Political Science, 66(3), 587–603.

Open AI. (n.d.) Pricing. Retrieved from: https://openai.com/pricing

Pfaff, S., Crabtree, C., Kern, H. L., & Holbein, J. B. (2021). Do street-level bureaucrats discriminate based on religion? A large-scale correspondence experiment among American public school principals. Public Administration Review, 81(2), 244–259.

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

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

Yackee, S. W. (2005). Sweet-Talking the Fourth Branch: The Influence of Interest Group Comments on Federal Agency Rulemaking. Journal of Public Administration Research and Theory, 16(1), 103–124.

Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). QualiGPT: GPT as an easy-to-use tool for qualitative coding. arXiv preprint arXiv:2310.07061.