Research Article

PUBLIC FINANCE JOURNAL

Using Large Language Models to Forecast Local Government Revenue

Il Hwan Chung[i, c], Berat Kara[ii], Melissa F. McShea[iii], Rahul Pathak[iv], Daniel Williams[v]

We examine the use of a public access large language model (LLM) to make local government revenue forecasts. ChatGPT is an LLM that is not specifically designed to perform quantitative analysis. However, it is capable of completing a wide range of tasks. The goals of this article are to determine the accuracy that can be obtained and to examine its potential bias. This study is based on a government revenue dataset from the Government Finance Officers Association (GFOA). The benefits of determining the accuracy and bias of LLM forecasts include providing a low-cost forecast method for small- and medium-sized governments and enabling external observers to validate forecasts made by official sources. Discovering the limitations of ChatGPT and similar LLMs, as well as the specific conditions required to use them wisely, may help localities avoid adverse outcomes. We find that a combination of LLM and human input provides a viable alternative forecasting method for small- and medium-sized governments, and it enables external observers to validate forecasts made by official sources. Errors in forecasting with the human-in-the-loop can be as low as 9.9 percent at the aggregated annual level. Using ChatGPT results alone can lead to high-error forecasts that may not be reliable.

Keywords: Artificial Intelligence, ChatGPT, Forecasting, Government Revenue, Large Language Model

The objective of this study is to determine the relative effectiveness of using large language models (LLMs) to forecast revenue. Using large language models (LLMs) is a new option that may provide opportunities to enhance revenue forecasting for certain governments. Small local governments are commonly believed to have poor forecasting capacity and to rely principally on judgmental forecasting (Bretschneider et al., 1992). Small-, medium-sized, and other low-resourced governments may have limited resources for such highly skilled technical staff as

[i] Graduate School of Government, Sungkyunkwan University. https://orcid.org/0000-0003-0061-2827.
[ii] Department of Public Finance, Istanbul Medeniyet University. https://orcid.org/0000-0002-6948-2197.
[iii] Department of Public Management, John Jay College of Criminal Justice. https://orcid.org/0000-0002-5263-7033.
[iv] Marxe School of Public and International Affairs, Baruch College. https://orcid.org/0000-0003-2611-6776.
[v] Marxe School of Public and International Affairs, Baruch College. https://orcid.org/0000-0002-3225-5556.
[c] Corresponding Author: ihchung@skku.edu.

https://doi.org/10.59469/pfj.2025.46

forecasters. Instead, forecasting may be one of many duties assigned to relatively untrained staff. This same condition may apply to other poorly resourced governments, such as those found in some developing countries. A survey of 34 developing and transition economies in Sub Saharan Africa and Asia in early 2000s reported that about 85 percent of sample countries used subjective assessments or simple extrapolation techniques as the main method of deriving their budget revenue forecast (De Renzio & Cho, 2020; Pathak et al., 2022). Identification of a new resource that can improve forecast outcomes may improve the budgetary process and fiscal governance.

One benefit of LLMs is that they may provide a low-cost forecasting method that can be implemented by organizations with a wide range of forecasting capabilities (Lee et al., 2024). This may provide alternate or supplemental options for governments of all sizes. A second benefit of LLMs is that they may also provide a method by which informed members of the public can validate official forecasts. State and local governments in the United States tend to conservatively estimate revenue expectations, particularly for income-elastic revenue sources, as a precautionary measure (Williams & Onochie, 2013). While this approach aims to enhance fiscal stability, it may also contribute to the perception that current expenditure practices are unsustainable (Perry et al., 2023; Williams & Onochie, 2013), a narrative that can be leveraged for political purposes (Champeny, 2023).

A third benefit of LLM-based forecasting has to do with the use of naïve baselines. Three common baseline forecasts are (1) Naïve 1 (the next period will be the same as the last observed period); (2) Naïve 21 (the next period will be the last observed period added to or multiplied by some increment(s) based on the recent change) (Chen et al., 2008); and (3) Naïve 22 (the next period is the last period after seasonal adjustment) (Koutsandreas et al., 2022).

These naïve methods serve as essential benchmarks for evaluating predictive accuracy. Baseline forecasts serve as a reference point for evaluating model performance. Any method that consistently underperforms relative to these benchmarks should be reconsidered or refined. The goal of using LLMs is not to replace naïve baselines, but to introduce a potentially more robust Naïve AI alternative that enhances predictive accuracy while maintaining the role of traditional baselines as evaluation tools.

To fulfill these three purposes, low- or no-cost LLMs that require limited skill that can be readily learned in a single brief workshop may increase forecasting accuracy for low-skilled governments, provide more transparency where forecasts are difficult to evaluate, and set a higher baseline standard for forecasting accuracy in general.

Literature Review

A common method for evaluating forecast methods is to compare forecast results across one or more time series. Some of these studies are the cutting edge of forecasting methodology (Makridakis et al., 1982; Makridakis et al., 2022). Others examine the benefits of various methods concerning specific types of data series (Cerqueira et al., 2020; Chung et al., 2022; Hyndman & Koehler, 2006; Makridakis et al., 2020; Noor et al., 2022; Williams & Kavanagh, 2016; Williams & Miller, 1999); these have included several that examine the use of methods in forecasting government revenue data series (Chung et al., 2022; Noor et al., 2022; Williams & Kavanagh, 2016). While the cutting-edge methods studies strictly focus on promising newer techniques, domain-focused approaches can include a variety of processes, including more traditional ways and even those that may be perceived as potentially defective (Chung et al., 2022; Williams & Kavanagh, 2016).

Typical evaluation of state and local government revenue forecast accuracy uses a limited set of standard forecast methods examining a small number of revenue series (Cirincione et al., 1999; Frank & Zhao, 2009; Gianakis & Frank, 1993). However, in recent years, there has been an interest in using machine language and related forecast methods to forecast government revenue (Chung et al., 2022; Kaburuan et al., 2019Noor et al., 2023; Qiumin, 2018; Uddin et al., 2023). As Chung et al. (2022) discuss, these studies, excluding their study, are typically limited to a few data series from a single source. Some studies used machine learning to forecast government expenditures (Yang et al., 2023). In the broader methods literature, there is considerable doubt that the machine language approach is as effective as the best standard methods (Lim & Zohren, 2021; Makridakis et al., 2018). In some studies, machine learning methods were tested against each other (Chung et al., 2022; Febriminanto & Wasesa, 2022; Goulet Coulombe et al., 2022; Li, 2012).

Chung et al. (2022) recommended continuing to study such more complex forecasting methods because there is considerable pressure for government forecasters to adapt cutting-edge “smart” methods (; Vogl et al., 2020; Yoon, 2020).

In recent years, studies utilizing large language models (LLMs) for time-series forecasting have demonstrated promising results across various domains. For instance, Makridis et al. (2023) explored the use of LLMs in forecasting within the food industry, while Wu and Ling (2024) applied LLM-based forecasting methods to predict wind speeds. In the financial sector, Gopali et al. (2024) utilized diverse datasets to highlight the potential of LLMs. Lopez-Lira and Tang (2023) showed that ChatGPT outperformed traditional financial analysis methods in predicting stock returns. Santschi et al. (2024) extended the application of LLMs to budget forecasting, reporting superior performance compared to conventional methods. These studies collectively suggest that LLMs can provide more accurate and versatile forecasting outcomes, warranting further exploration of their potential in public finance forecasting.

Data and Methods

This paper focuses on the use of large language models (LLMs) to forecast government revenue using monthly revenue datasets spanning from July 2009 to June 2019. Although there are some exceptions (Buxton et al., 2019; Cirincione et al., 1999; Frank, 1990; Reddick, 2004; Williams & Kavanagh, 2016; Chung et al., 2022), prior studies on government budget forecasting have generally been limited to a few localities with a limited number of data series. We attempt to analyze the feasibility of large language models (LLMs) as a forecasting tool. Our study builds upon prior research by utilizing a more extensive dataset with seasonality provided by the Government Finance Officers Association (GFOA), under the condition that the specific localities remain anonymous.[1]

The data comprise monthly revenue figures from July 2009 to June 2019 for numerous small- and medium-sized governments. The types of revenue are property tax, fines, and license fees. The source governments are anonymized at the request of GFOA. There are 176 labeled data series with as many as 120 observations. However, many series have only aperiodic data entries and are excluded from this study. After excluding those series that terminate before June 2019, 90 series remain. After removing series with fewer than 24 continuous observations at the end of the series and series with intermittent missing observations, 24 series remained. The experiment was conducted in June 2023.

Our forecasting procedures are as follows. First, we divide our sample into training data and holdout data. We set the last 24 months of data as a holdout set. Subsequently, we identify the pattern in the revenue data using a training set of data from traditional forecasting techniques and machine learning algorithms. Then, estimated parameters from each model in the training set of data are applied to the holdout data. Results are compared across models using the mean of the symmetric absolute percent error (sMAPE) (Chung et al., 2022; Makridakis et al., 2018a). sMAPE is the most recommended and used (Hyndman, 2006; Makridakis & Hibon, 2000; Taieb et al., 2012; Williams & Calabrese, 2019; Williams & Miller, 1999). Other common measures include MSE, which is dependent on the size of the observations, and MAPE, which is biased by the direction of error.

$$sMAPE = \frac{100}{n} * \sum_{\mathrm{t=1}}^{\mathrm{n}} \frac{|\mathrm{Y_t} - \hat{\mathrm{Y}}_\mathrm{t}|}{(|\mathrm{Y_t}| + |\hat{\mathrm{Y}}_\mathrm{t}|)/2}$$

Y_t is the actual revenue and $\hat{Y}_t$ It is the result of the model's forecast at time *t*.

Our primary interest is to explore the feasibility of LLMs in revenue forecasting with different types of forecasters. A growing body of research has documented several challenges associated with large language models (LLMs), such as ChatGPT, developed by OpenAI. These challenges include biases from training data (Dale, 2021; Motoki et al., 2023) and hallucinations (Alkaissi & McFarlane, 2023; Azamfirei et al., 2023; Dwivedi et al., 2023; Ji et al., 2023).[2] Thus, we categorize forecasting with LLMs into two approaches: The first method relies solely on the LLM. At the same time, the second involves conducting revenue forecasting by combining the LLM with consultation or evaluation by human experts.

We also employ different types of forecasters, including traditional time series forecasting and machine learning algorithms. The traditional time series forecasting techniques include Holt's exponential smoothing and the Autoregressive Integrated Moving Average (ARIMA) model. The latter includes the Generalized Regression Neural Network (GRNN) and the K-Nearest Neighbors (KNN) algorithm. Each machine learning method represents distinctive features of machine algorithms: neural net and distance-based regression. Our machine learning algorithms are chosen based on previous studies that focused on time-series model forecasting (Chung et al., 2022; Makridakis et al., 2018).

In our study, we employed two distinct machine learning-based forecasting techniques: the GRNN and the KNN algorithm. These models were chosen due to their frequent application in time-series forecasting within the realm of machine learning, as highlighted in several studies (Ahmed et al., 2010; Makridakis et al., 2018). Each of these algorithms possesses unique features. The GRNN, also known as the Nadaraya-Watson estimator (Specht, 1991), utilizes neural network methodologies, incorporating a rapid, one-pass learning algorithm with a Gaussian function in its hidden layer. On the other hand, the KNN algorithm operates by measuring the distance between data points. The GRNN, a nonparametric method, forecasts by averaging the output of training data points based on their proximity to the new observation, as explained by Makridakis et al. (2018). See Martínez et al. (2019) for further description of GRNN.

While the field of machine learning predominantly emphasizes artificial neural networks for time series forecasting, our approach also includes alternative methods, such as time series

Table 1. Forecasting Approaches

	LLM	**Human-in-the-Loop**
Traditional Forecasting Methods	(1)	(2)
Machine Learning Algorithm Methods	(3)	(4)

KNN (K-Nearest Neighbors). KNN is a method of nonparametric regression that makes predictions based on the Euclidean distance within the feature space. To elaborate, for N given inputs, this technique selects the nearest K training data points. The forecast is then determined by calculating the average of the target values of these selected points, as detailed by Makridakis et al. (2018).

Taken all together, we present the four types of revenue forecasting approaches as described in Table 1. It outlines our primary approaches based on two criteria: 1) whether the forecasting is derived from LLM alone or from a combination of LLM and human input, i.e., human-in-the-loop, and 2) whether the method is a traditional time series forecaster or a machine learning algorithm. We also provided Naïve 1 forecasting model as a benchmark, which projects revenue based on prior year's revenue.

When the forecast is derived solely from LLM, we use ChatGPT 3.5. We provide the training data and directly ask for the forecast for the next 24 months. We generate prompts for four different forecasters (Holt, ARIMA, GRNN, KNN) in 24 localities, which leads to separate 96 prompts (24 time series × 4 forecasters = 96 prompts). The prompt is used as follows:

> "Suppose you are a budget analyst in the municipal government. Using the monthly revenue data (from July 2009 to June 2017) provided, please create a comprehensive revenue forecast for the next 24 months. Just respond "OK" to this prompt, as I will provide you with detailed revenue data in the second prompt. (With giving raw data) Please provide revenue forecast for the next 24 months using [insert forecasting method]"

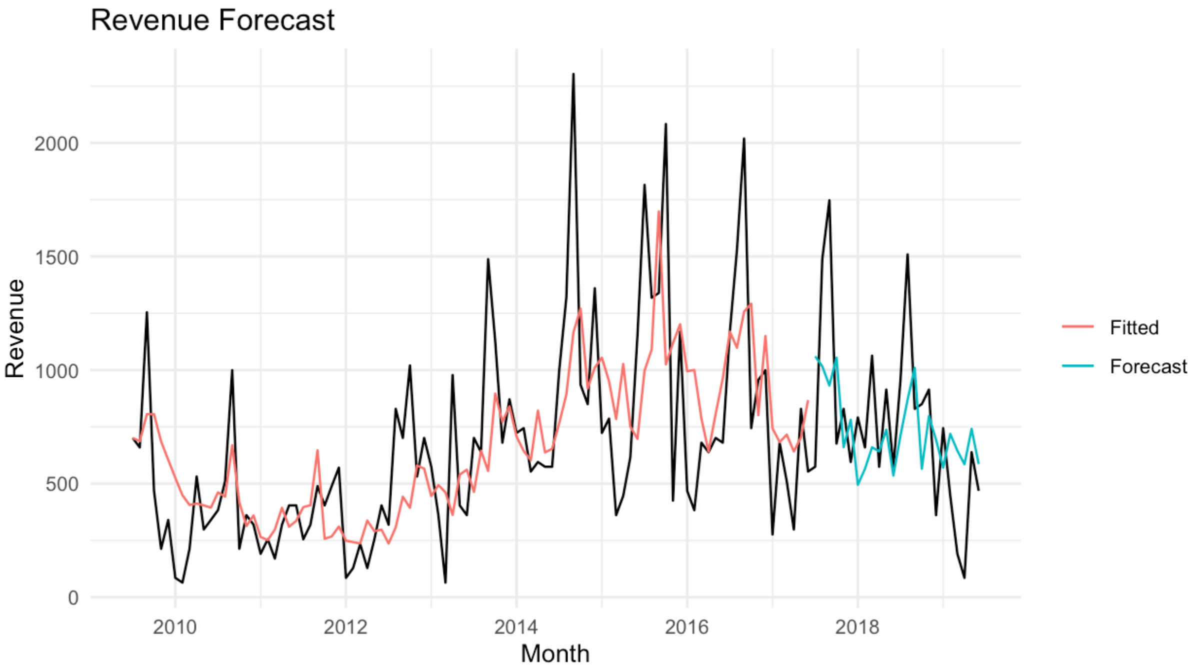
In the human-in-the-loop approach, where forecasting is derived from a combination of LLM and human input, we use the same prompt but do not solely rely on the response from ChatGPT.[3] As the baseline approach, we employ ChatGPT 4.0,[4] which features advanced data analysis and plug-in features. These allow us to run statistical code, such as Python or R, in a virtual environment simultaneously.[5] Additionally, we either provide detrended data for the training dataset or set hyperparameters for the forecaster.[6]

> "Suppose you are a budget analyst in the municipal government. Using the monthly revenue data (from July 2009 to June 2017) provided, please create a comprehensive revenue forecast for the next 24 months. Just respond "OK" to this prompt, as I will provide you with detailed revenue data in the second prompt. (With giving raw data) Please provide revenue forecast for the next 24 months using [insert forecasting method]."

Table 2. Forecasting Results from LLMs

	6 Months Forecasting	**12 Months Forecasting**	**18 Months Forecasting**	**24 Months Forecasting**
Holt Exponential	82.40%	80.24%	83.03%	85.13%
ARIMA	84.06%	79.85%	82.10%	80.34%
KNN	78.19%	75.07%	77.59%	77.29%
GRNN	82.34%	80.80%	82.04%	81.90%
Naïve 1	87.03%	83.90%	85.70%	85.05%

Figure 1. Illustration of Forecasting



Results

Among 120 monthly observations, we split the first 96 months as a training set and the last 24 months as a holdout data set. We conducted forecasts using four different models – Holt exponential model, ARIMA, KNN, and GRNN – to predict the 24 months of the holdout data set. Table 2 shows the comparison of forecasting results from 24 localities by ChatGPT. Since we use monthly revenue data for forecasting, we divide the forecasting into four different forecast horizons: the first 6 months, the first 12 months, the first 18 months, and the first 24 months. This forecast reflects three distinct budget periods in the municipal forecasting practices: the current/midyear, the budget year, and the out years (Williams & Calabrese, 2016). Among four different forecasters, the average forecast accuracy of revenue is relatively better with KNN in each forecast horizon. However, the sMAPE ranges from 78.19% to 77.29%, which indicates

Table 3. Forecasting Results from LLMs

	ChatGPT	**ChatGPT + Human**		
		Plug-In	**Detrend**	**Hyper-Parameter**
Holt	42.73%	41.87%	39.20%	39.05%
ARIMA	43.94%	41.10%	39.60%	39.10%
KNN	43.23%	50.10%	---	---
GRNN	43.18%	---	---	---

Note: sMAPE for naïve 1 is 47.30%.

Table 4. Forecasting Results from ChatGPT

		Chat GPT	**ChatGPT + Human**		
			Plug-In	**Detrended / Deseasonalized**	**Hyper-Parameter**
Holt	1st year	39.10%	37.80%	21.70%	16.30%
	2nd year	2.80%	11.20%	11.80%	10.40%
ARIMA	1st year	43.00%	13.70%	7.30%	9.00%
	2nd year	11.50%	6.10%	19.70%	11.90%
KNN	1st year	41.40%	14.20%	---	---
	2nd year	10.10%	18.30%	---	---
GRNN	1st year	41.30%	---	---	---
	2nd year	10.20%	---	---	---

that forecasting by ChatGPT does not perform well. This pattern is consistent with other forecasters, where its accuracy ranges from 79.34% to 85.13%.

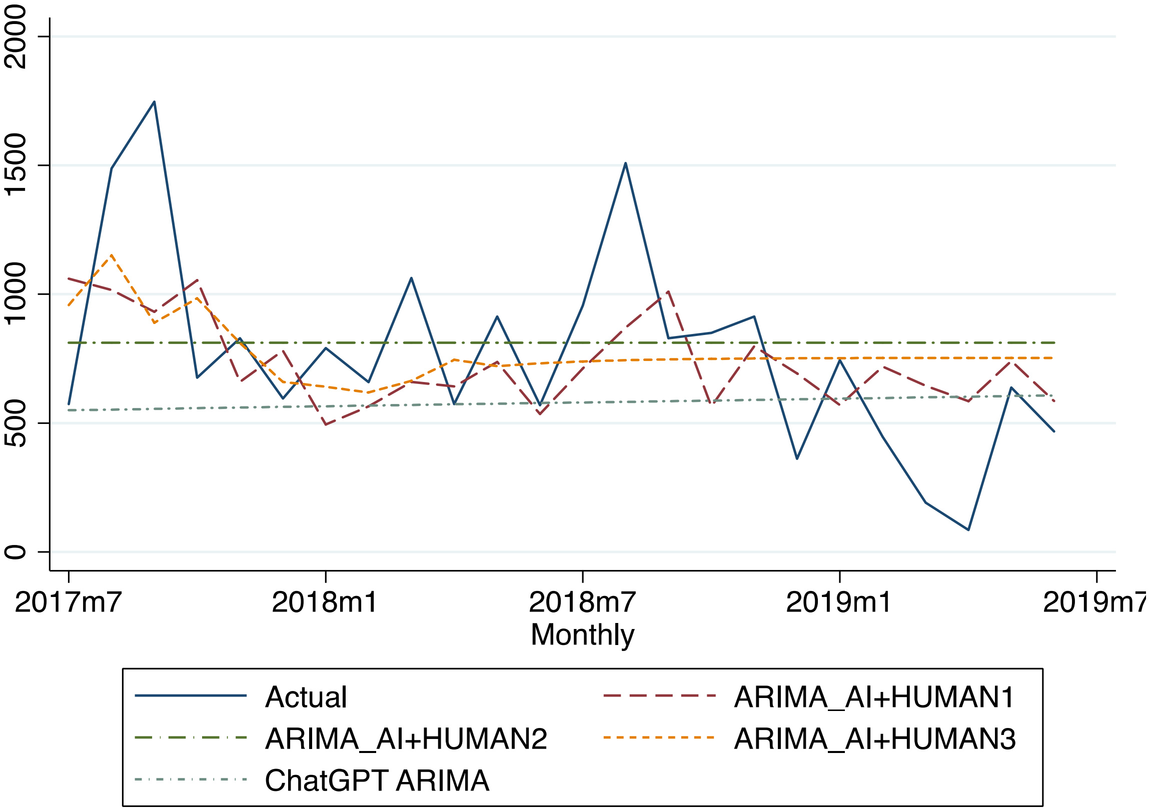
In the next step, we explore whether human engagement in forecasting with LLMs enhances forecast accuracy. Since human-in-the-loop involvement in the forecasting process with LLMs requires extensive time and effort, we illustrate this comparison based on a single time series.

Figure 1 illustrates the case of ARIMA combined with LLM and human input. Table 3 presents the results across different forecasters for each year, measured in terms of sMAPE. Compared to the forecasting accuracy of all 24 localities, the results show a slight improvement. However, they still lack the accuracy of forecasts conducted in other studies (Makridakis et al., 2018). It is essential to note that some results are left blank because ChatGPT was unable to provide relevant forecasting code.

We split the forecasting results for each year to explore the heterogeneous outcomes across different forecast horizons. For instance, the Holt exponential method indicates that the projected error in the monthly revenue forecast from July 2017 to June 2018 is 39.1%, while the sMAPE for the second year is 2.8%. It is worth noting that forecast accuracy tends to be better in the first year than in the second year for all types of forecasting methods: See Table 4. A possible explanation for this is that the revenue for the second year follows a dissimilar pattern from training data, as depicted in Figure 2.

To understand the forecasting mechanism for each month, we present Figure 2, which compares actual revenues with forecasted revenues using the holdout dataset. Interestingly, the

Figure 2. Comparison of Forecasting Results with ChatGPT



Note: y axis refers to monthly revenue in a specific small local government A.

Table 5. Forecasting Results from ChatGPT: Yearly Basis[7]

	ChatGPT	**ChatGPT + Human**		
		Plug-In	**Detrended / Deseasonalized**	**Hyper-Parameter**
Holt	20.90%	24.50%	16.80%	13.40%
ARIMA	27.30%	9.90%	13.50%	10.40%
KNN	25.80%	16.30%	---	---
GRNN	25.80%	---	---	---

Note: sMAPE for naïve 1 is 25.3%.

only LLM model and ARIMA with detrended data projected relatively linear trends in monthly revenue. This could indicate that hallucination problems may occur when generating time series predictions, which calls for human interaction with the LLM.

We have conducted the forecasting based on monthly data. Since the budget data are often aggregated and evaluated at the annual level, we evaluate the sMAPE at this level. Table 5 presents interesting results: In the annualized forecast, the combination of LLM and human input improved substantially, with a reduction in error from 27.3 percent to 9.9 percent. Given that small local governments often employ simple methods, such as judgmental forecasts (Cirincione et al., 1999; Gianakis & Frank, 1993), the combination of LLM and human input offers an alternative forecasting method that may benefit these governments.

Conclusion

ChatGPT is a large language model that is not specifically designed to perform quantitative analysis. However, they have been reported to be capable of completing a wide range of tasks. The purpose of this study is to determine how accurate forecasting with LLMs can be, examine its potential bias, and establish the best prompt to use to obtain a forecast. Given this is a new area of enquiry, there are no peer-reviewed studies that examine the use of LLMs in the context of budget forecasting. In private communication with a forecast consultant, we were advised that a similar AI (BardAI) has provided forecast results within five percent of actual revenues for a limited number of data series. Nevertheless, the results here are not consistent with this optimistic result. Our study finds that a combination of LLM and human input provides a viable alternative forecasting method, enabling small- and medium-sized governments, as well as external observers, to validate forecasts made by official sources. Errors in forecasting with the human-in-the-loop can be as low as 9.9 percent at the aggregated annual level. Using ChatGPT results alone can lead to high-error forecasts that may not be reliable.

This is a rapidly evolving technological environment, and new AI systems are becoming available daily. The focus of this study is not merely to examine the extent of errors but to assess the viability of using these technologies for budget forecasting, particularly for governments with low economic analysis capacity. We have demonstrated that the use of these technologies with some human input is viable and may offer several opportunities for institutional interventions. For instance, representative organizations of budget and finance officials or other international agencies (e.g., GFOA and National Association of State Budget Officers (NASBO) in the US context and International Monetary Fund (IMF) and World Bank in the international context) may help local governments by developing simpler LLM-based applications for specific forecasting contexts (economic and budget forecasting) or by developing toolkits that would enable officials to conduct the analysis. Future research may extend this approach to other datasets and also deploy other AI tools that have recently become available, beyond those used in this study.

We recommend further research examining other easy-to-implement combinatory strategies. In particular, forecast averaging, sometimes labeled “ensemble forecasting” (Kriz, 2019; Makridakis & Winkler, 1983; Yamana et al., 2016), may improve forecasts produced using methods presented here. Thus, we recommend examining the potential benefits of combining judgmental and AI-produced forecasts, simple methods (such as moving averages) and AI-produced forecasts, naïve forecasts, and AI-produced forecasts through averaging. Open AI platforms offer a new tool that is widely accessible and low-cost. However, these tools also have their limitations, and the technology is changing at a rapid pace. Therefore, the use of human judgment while leveraging the analytical capacity of large language models is recommended. Additionally, future studies should acknowledge the limitations of the experiment, which was conducted in June 2023. For instance, limitations include the lack of transparency in closed-source model, the justifiable strategy of prompt design, and model sensitivity.

Endnotes

[1] The application of this approach is similar to the approach seen in the renowned M-competitions spearheaded by Makridakis and his colleagues.

[2] ChatGPT often provides confident responses that appear nonsensical and unfaithful. Such a phenomenon has been referred to as “hallucination” (see Alkaissi & McFarlane, 2023; Ji et al., 2023).

[3] Forecasting accuracy may depend on both the choice of LLM version (e.g., ChatGPT 3.5 versus ChatGPT 4) and the extent of human intervention in the forecasting process. To clarify these influences, we have decomposed the LLM outputs into more granular components, thereby isolating the differences attributable to model version and human-in-the-loop adjustments (see Tables 3 and 4). For instance, Table 3 highlights how comparisons between “ChatGPT” and “ChatGPT with plugin” involve the model difference, since plugin capabilities are exclusively available in ChatGPT 4. Furthermore, we incorporate “detrending” and “hyperparameter” steps to illustrate how domain experts can intervene prior to model execution—either by removing underlying trends from the data or by selecting the model’s hyperparameters in advance. Such human involvement provides a means of refining the forecasting process beyond what the LLM can achieve alone.

[4] ChatGPT 4.0 was released just before we conducted this forecasting experiment and prepared the manuscript.

[5] All the statistical code generated by ChatGPT is unable to provide results due to hallucination in the code. In those cases, a human analyst engages in an analytical process to debug the code.

[6] For instance, Holt's exponential method uses (a=0.5, b=0.01) and ARIMA uses (1, 1, 12) hyperparameters.

[7] We do not focus on how accurate the LLM is as a prompted forecaster. However, we replicated the Holt method with the same hyperparameter in the standard statistical model to check its accuracy. The findings show that LLM’s outputs do not precisely match the forecasts produced by standard statistical software (sMAPE in the statistical model is 11.2% compared to 13.4% by LLM). LLM appears to approximate Holt’s methodology rather than replicate its exact mathematical formulation.

Disclosure Statement

The authors declare no conflicts of interest related to this article’s research, authorship, or publication.

References

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, *29*(5-6), 594-621.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in

Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, *120*, 27. https://doi.org/10.1186/s13054-023-04393-x

Bretschneider, S. I., Bunch, B., & Gorr, W. L. (1992). Revenue forecast errors in Pennsylvania local government budgeting: Sources and remedies. *Public Budgeting & Financial Management*, *4*(3), 721-743.

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, *109*, 1997-2028.

Champeny, A. (2023, August 15, 2023). 5 myths and facts about the NYC FY 2024 budget. *City Blog*. https://cbcny.org/research/5-myths-and-facts-about-nyc-fy-2024-budget#myth5

Chen, R. J. C., Bloomfield, P., & Cubbage, F. W. (2008). Comparing forecasting models in tourism. *Journal of Hospitality & Tourism Research*, *32*(1), 3-21. https://doi.org/10.1177/1096348007309566

Chung, I. H., Williams, D. W., & Do, M. R. (2022). For better or worse? Revenue forecasting with machine learning approaches. *Public Performance & Management Review*, *45*(5), 1133-1154. https://doi.org/10.1080/15309576.2022.2073551

Cirincione, C., Gurrieri, G. A., & Van De Sande, B. (1999). Municipal government revenue forecasting: Issues of method and data. *Public Budgeting & Finance*, *19*(1), 26-46.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113-118.

De Renzio, P., & Cho, C. (2020). Exploring the determinants of budget credibility. *International Budget Partnership*.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., & Ahuja, M. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Gianakis, G. A., & Frank, H. A. (1993). Implementing time series forecasting models: Considerations for local governments. *State and Local Government Review*, *25*(2), 130-144.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920-964.

Gopali, S., Siami-Namini, S., Abri, F., & Namin, A. S. (2024). The performance of the LSTM-based code generated by Large Language Models (LLMs) in forecasting time series data. *Natural Language Processing Journal*, *9*, 100120.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43-46.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Kaburuan, E. R., Lindawati, A. S. L., Putra, M. R., & Utama, D. N. (2019). A model configuration of social media text mining for projecting the online-commerce transaction (case: Twitter tweets scraping). 2019 7th International Conference on Cyber and IT Service Management (CITSM), *7*, 1-4.

Koutsandreas, D., Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2022). On the selection of forecasting accuracy measures. *Journal of the Operational Research Society*, *73*(5), 937-954.

Kriz, K. A. (2019). Ensemble forecasting. In D. W. Williams & T. Calabrese (Eds.), *The Palgrave handbook of government budget forecasting*. Palgrave MacMillan. https://doi.org./10.1007/978-3-030-18195-6

Lee, M., Hayes, D., & Maher, C. (2024). AI as a budgeting tool: Panacea or Pandora's box?. *Public Finance Journal*, *1*(1), 49–65. https://doi.org/10.59469/pfj.2024.6

Lopez-Lira, Alejandro and Tang, Yuehua. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. http://dx.doi.org/10.2139/ssrn.4412788

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111.

Makridis, G., Mavrepis, P. & Kyriazis, D. (2023). A deep learning approach using natural language processing and time-series forecasting towards enhanced food safety. *Mach Learn*, *112*, 1287–1313. https://doi.org/10.1007/s10994-022-06151-6

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE*, *13*(3), e0194889.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). Predicting/hypothesizing the findings of the M4 Competition. *International Journal of Forecasting*, *36*(1), 29-36. https://doi.org/10.1016/j.ijforecast.2019.02.012

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*(4), 1325-1336.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987-996.

Martínez, F., Charte, F., Rivera, A. J., & Frías, M. P. (2019). Automatic time series forecasting with GRNN: A comparison with other models. In I. Rojas, G. Joya, & A. Catala (Eds.), *IWANN 2019. Lecture notes in computer science: Vol. 11506. Advances in computational intelligence* (pp. 198-209). Springer. https://doi.org/10.1007/978-3-030-20521-8_17

Martínez, F., Frías, M. P., Pérez-Godoy, M. D., & Rivera, A. J. (2022). Time series forecasting by generalized regression neural networks trained with multiple series. *IEEE Access*, *10*, 3275-3283.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More human than human: Measuring chatgpt political bias. *Public Choice*. https://doi.org/10.1007/s11127-023-01097-2

Noor, N., Sarlan, A., & Aziz, N. (2022). Revenue prediction for Malaysian federal government using machine learning technique. 2022 11th International Conference on Software and Computer Applications.

Noor, N., Sarlan, A., & Aziz, N. (2023). Government revenue prediction using feed forward neural network. *Journal of Theoretical and Applied Information Technology*, *101*(6).

Pathak, R., Cangiano, M., & Desouve, J. (2022). *Forecast bias and budget credibility in Rwanda, Senegal, and Uganda*. https://www.cabri-sbo.org/uploads/files/Documents/Forecast-Bias-and-Budget-Credibility-in-Rwanda-Senegal-and-Uganda.pdf

Qiumin, L. (2018). A novel design of hybrid polynomial spline estimation and GMDH networks for modeling and prediction. *International Business and Management*, *16*(1), 23-28.

Santschi, D., Grau, M. C., Fehrenbacher, D., & Blohm, I. (2024). Artificial intelligence to improve public budgeting. *ICIS 2024 Proceedings*. 1. https://aisel.aisnet.org/icis2024/iot_smartcity/iot_smartcity/1

Specht, D. F. (1991). A general regression neural network. *IEEE transactions on Neural Networks*, *2*(6), 568-576. https://doi.org/10.1109/72.97934

Uddin, M. M., Begum, H., Hasan, M., & Sultana, Z. (2023). Predicting future government revenues generation using artificial intelligence for ensuring economic development of Bangladesh.

Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart technology and the emergence of algorithmic bureaucracy: Artificial intelligence in UK local authorities. *Public Administration Review*, *80*(6), 946-961. https://doi.org/10.1111/puar.13286

Williams, D. W., & Calabrese, T. (2019). Current midyear municipal budget forecast accuracy. In D. W. Williams & T. Calabrese (Eds.), *The Palgrave handbook of government budget forecasting*. Palgrave MacMillan. https://doi.org/10.1007/978-3-030-18195-6

Williams, D. W., & Calabrese, T. D. (2016). The status of budget forecasting. *Journal of Public and Nonprofit Affairs*, *2*(2), 127-160. https://doi.org/10.20899/jpna.2.2.127-160

Williams, D. W., & Kavanagh, S. C. (2016). Local government revenue forecast competition/comparison. *Journal of Public Budgeting, Accounting, and Financial Management*, *28*(4), 488-526.

Williams, D. W., & Miller, D. M. (1999). Level-adjusted exponential smoothing for modeling planned discontinuities. *International Journal of Forecasting*, *15*(3), 273-289

Williams, D. W., & Onochie, J. (2013). The Rube Goldberg machine of budget implementation, or is there a structural deficit in the New York City budget? *Public Budgeting & Finance*, *33*(4), 1-21. https://doi.org/10.1111/j.1540-5850.2013.12021.x

Wu, T., & Ling, Q. (2024). STELLM: Spatio-temporal enhanced pre-trained large language model for wind speed forecasting. *Applied Energy*, *375*, 124034.

Yamana, T. K., Kandula, S., & Shaman, J. (2016). Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface*, *13*(123), 20160410.

Yang, C.-H., Molefyane, T., & Lin, Y.-D. (2023). The Forecasting of a leading country's government expenditure using a recurrent neural network with a gated recurrent unit. *Mathematics*, *11*(14), 3085.

Yoon, S. (2020). A study on the transformation of accounting based on new technologies: Evidence from Korea. *Sustainability*, *12*(20), 8669.

Zhao, L. (2009). Neural networks in business time series forecasting: benefits and problems. *Review of Business Information Systems (RBIS)*, *13*(3).

Author Biographies

Il Hwan Chung is an associate professor in the Graduate School of Governance at Sungkyunkwan University. He received his MPA from the University of Georgia and his Ph.D. in public administration from Syracuse University. His research focuses on public finance, budgeting, and education finance.

Berat Kara is an assistant professor in the Department of Public Finance at Istanbul Medeniyet University. He received his Ph.D. in public finance from Istanbul University. His research focuses on public budgeting and finance, with a particular emphasis on the accuracy and reliability of budget forecasting, including the development and application of forecasting methods**.**

Melissa F. McShea is an assistant professor in the Department of Public Management at John Jay College of Criminal Justice. She received her master's degrees in economics and public policy along with her Ph.D. in public policy and administration from George Washington University. She also holds a master's degree in applied economics from University of Michigan. Her research focuses on state and local public finance.

Rahul Pathak is an associate professor in the Marxe School of Public and International Affairs at Baruch College and the Director of Howard J. Samuels State and City Policy Center. He received his Ph.D. in public policy from Georgia State University, Atlanta. His primary research interests lie at the intersection of public finance and social policy.

Daniel W. Williams is professor emeritus at the Marxe School of Public and International Affairs at Baruch College. He received his Ph.D. in policy analytics from Virginia Commonwealth University. Williams has 20 academic publications related to forecasting, ranging from methodology to empirical evaluation of forecasts, including two textbooks and a handbook.