Data Shop

Data Shop, a department of Cityscape*, presents short articles or notes on the uses of data in housing and urban research. Through this department, the Office of Policy Development and Research introduces readers to new and overlooked data sources and to improved techniques in using well-known data. The emphasis is on sources and methods that analysts can use in their own work. Researchers often run into knotty data problems involving data interpretation or manipulation that must be solved before a project can proceed, but they seldom get to focus in detail on the solutions to such problems. If you have an idea for an applied, data-centric note of no more than 3,000 words, please send a one-paragraph abstract to chalita.d.brandly@hud.gov for consideration.*

Generative AI: Mining Housing Data With a Higher Powered Shovel

Dylan J. Hayden
U.S. Department of Housing and Urban Development

The views expressed in this article are those of the author and do not represent the official views or policies of the Office of Policy Development and Research or the U.S. Department of Housing and Urban Development.

Abstract

This article investigates the potential applications of generative artificial intelligence (AI) models, such as Chat Generative Pre-Trained Transformer (ChatGPT), in housing research by assisting with data analysis. Using the U.S. Department of Housing and Urban Development (HUD) Picture of Subsidized Households dataset, the study employs ChatGPT to generate code and analyze correlations within a housing research context. The methodology includes creating a computer program for calculating correlations and incorporating ChatGPT to analyze the output, leveraging OpenAI's application programming interface. The article addresses concerns related to bias, inaccuracies, and improper citation and examines the benefits and limitations of using ChatGPT in housing research. This study contributes to the ongoing conversation surrounding the responsible and effective use of generative AI models in research across various disciplines.

Introduction

In November 2022, OpenAI released an artificial intelligence (AI) chatbot called Chat Generative Pre-Trained Transformer (ChatGPT) that has since brought AI into the mainstream and, some might argue, ushered in a new era of communication. ChatGPT is a conversational AI system designed to interact naturally and engagingly with users. It is fine-tuned from a series of large language models (LLM), the latest version as of this writing being GPT-4, which uses an extensive set of text data scraped from the internet. ChatGPT is trained using a combination of supervised and reinforcement learning techniques; supervised learning employs human-written inputs as demonstrations, and reinforcement learning leverages human feedback to optimize the model's responses (OpenAI, 2023). The scope of the data on which the model was trained and its meteoric adoption demonstrate the potential of LLM for natural language understanding and generation.

In the first few months following ChatGPT's release, the private sector largely embraced AI, and thousands of businesses have been launched using the LLM system. However, the reaction within academia has been more cautious, with concerns surrounding bias, inaccuracies, and improper citation potentially diluting the quality and validity of research (Van Dis et al., 2023). Although those are valid concerns that warrant research into developing safeguards, it is becoming clear that ChatGPT has the potential to enhance the precision and quality of scientific writing, shorten review times, make scientific writing more accessible to broader audiences, and even give rise to entirely new forms of scientific writing and research (Kappel, 2023). The nascent research into this type of AI's application to the field of education suggests researchers and educators should adopt a programming prompt mindset instead of a search mindset by adopting four categories of programming prompts: Conversational, content analysis, coding, and multimodal (Hwang and Chen, 2023).

For housing and urban studies-related research, ChatGPT can offer several advantages that may enhance the quality and efficiency of data analysis and interpretation. Some of the specific benefits include—

1. Rapid detection of patterns and correlations in extensive datasets, allowing researchers to rapidly generate insights and develop research questions.

2. Creation of custom scripts for data analysis, minimizing the need for extensive programming expertise and promoting equitable access to research opportunities.

3. Generation of natural language summaries of complex statistical analyses, making research findings more comprehensible for nonexperts and improving communication with policymakers and other stakeholders.

4. Support in building predictive models for housing market trends, potentially enabling researchers to forecast changes in housing availability, affordability, and demand.

The primary focus of this article is to explore the potential applications of generative AI models, such as ChatGPT, in assisting researchers with housing data analysis. Because ChatGPT is only a few months old at the time of this writing, there is a dearth of research investigating the application of GPT models to housing policy. This research aims to develop an experimental use case for ChatGPT, enabling it to both construct a tool for processing housing data and summarize the results of the tool's analysis in everyday written language. Through this investigation, this article intends to highlight the advantages and address the concerns associated with using AI models in housing research and beyond.

By examining the capabilities and limitations of ChatGPT in a housing research context, this article contributes to the ongoing conversation surrounding the use of generative AI models. It also seeks to promote a deeper understanding of how these models can be employed responsibly and effectively to enhance the quality and impact of research across various disciplines. Ultimately, the goals are to foster a more nuanced appreciation for the potential of AI in housing research and encourage further exploration of its applications in other areas.

Using ChatGPT as a higher powered shovel for data mining will enable researchers to explore complex housing datasets more efficiently, uncovering novel insights and driving innovation. For this study, developing a functional script took approximately 1 day, with future fine-tuning potentially requiring a few more days. This time investment is considerably shorter than the week or more it could have taken using traditional tools, such as Excel, SPSS, or Power BI. For researchers with minimal coding experience, the ability to generate Python code that automates the process saves months of learning time or the cost of hiring a developer. However, it is crucial to maintain a balance between the advantages of generative AI models and the ethical considerations that arise from their use. By being mindful of those concerns and working toward responsible integration of AI tools, the academic community can continue to advance this field of research and contribute to meaningful progress in addressing pressing housing challenges.

Methodology

This study began by using HUD's Picture of Subsidized Households (POSH) dataset and selecting the most recent public housing data for all census tracts in California. Training the ChatGPT model involved using the comma-separated values (CSV) file containing 56 variables and providing the headings for the data columns. Once the model was primed with the relevant columns, the model was instructed to use the following prompt:

> Generate a Python script that will generate a correlation matrix for the data in these columns. In the data, exclude values of -1, -4, or -5, as these data are either missing or masked to protect privacy.

After several followup prompts to refine the output, ChatGPT produced a Python script that generated a correlation matrix and saved it to a CSV file. This file served as a reference for the correlation coefficients, which were integrated into subsequent steps. To concentrate solely on strong correlations and minimize the number of tokens required for the ChatGPT model to process, a threshold was established to include only correlation coefficients between 0.5 and 1.0 and -0.5 and -1.0, respectively.

The first phase of the process entailed requesting ChatGPT to create a basic code for calculating correlations based on the dataset. The second phase involved incorporating ChatGPT in the analysis of the output that those calculations generated. To accomplish this process, OpenAI's application programming interface (API) was integrated into the Python script. Upon plugging the Python script back into ChatGPT, the model was prompted to employ the OpenAI API to analyze the correlations with the following instruction:

> Update the Python script to print all correlations greater than 0.5 (positive or negative) and import the OpenAI API to provide a summary of the possible reasons for each correlation.

The resulting code included a section with a prompt for the API to summarize the data:

```
prompt = f'Possible reasons for the correlation between {corr[0]} and {corr[1]}:'
```

Subsequent prompts were then issued to ChatGPT to modify the Python script, enabling it to save the correlation summaries to a text file. This methodology demonstrates a novel approach to leveraging ChatGPT's capabilities for generating code and analyzing correlations within a housing research context.

Results

Initial results were mixed. On running the script, the AI generated several pages of summary text for each correlation in a matter of seconds. A review of the output text yielded summaries that could be inconsistent. For example, one summary gave three possible reasons for a 0.99 correlation between "# Reported"—which is the number of households for which reports (Form-50058 and Form-50059) were received—and "Subsidized units available."

1. "The number of reported units available is directly correlated with the number of subsidized units available because the subsidized units are provided by the government in order to help those in need."

2. "The number of reported units available is indirectly correlated with the number of subsidized units available because the government only provides subsidies for a certain number of units, so if there are more units available, it is likely that fewer of them will be subsidized."

3. "There is no clear correlation between the number of reported units available and the number of subsidized units available. It is possible that the subsidized units are provided by the government in order to help those in need, but it is also possible that the government only provides subsidies for a certain number of units, so if there are more units available, it is likely that fewer of them will be subsidized."

Other summaries demonstrated the potential for bias in the ChatGPT language model. For example, the summary text analyzed the 0.51 correlation between "% Hispanic"—which is the percentage of households that the ethnicity of the head of household is Hispanic (in a census tract in this case)—and the number of people per public housing unit.

1. "The Hispanic population is growing rapidly in the United States, so there are simply more Hispanic people to live in close proximity to one another."

2. “Hispanic people tend to have larger families than people of other ethnicities, so they need more space to live.”

3. “Hispanic people may be more likely to live in poverty than people of other ethnicities, so they may be more likely to live in overcrowded housing.”

4. “Hispanic people may be more likely to live in urban areas than people of other ethnicities, so they may be more likely to live in close proximity to one another.”

Statements like “The Hispanic population is growing rapidly in the United States” are not a result of information contained in the POSH dataset; rather, it is an inference drawn from the vast dataset OpenAI uses to build its LLM. Importantly, the ChatGPT model is only basing its summaries by the heading of a column of data. The data dictionary was not incorporated to provide the model with the definitions of each heading for additional context. The summaries more than likely include a mix of information contained in the OpenAI model, which, as the company says, may include biased information, or *hallucinations*, for which the AI uses probabilities to make up information to meet the scenario. Regardless of the bias or hallucinations, statements like this represent potentially testable hypotheses that can be verified with further data analysis. Further refinement of the model would likely separate the wheat from the chaff, narrowing the scope of hypotheses that are truly worth testing.

The initial output also suggested that further refinement of the prompt used to develop the summaries was necessary. The model was given a specific role and more context to develop possible explanations for the correlation coefficients.

```
prompt = f”Act as an Expert Statistician analyzing public housing data. Explain in one
short sentence the possible reasons for the correlation between {corr[0]} and {corr[1]}:
```

As the prompt suggests, responses in the summary text were shortened to a single sentence and did not offer multiple explanations for each correlation as with the initial prompt. The resulting output yielded a slightly less insightful explanation of the correlation between the percentage of Hispanic heads of households and the number of people per housing unit: “The correlation between % Hispanic and Number of people per unit could be due to a higher concentration of Hispanic families living in public housing with more people per unit.”

Conclusion

This study explored the potential applications of generative AI tools, such as ChatGPT, in the domain of housing research, specifically using HUD’s POSH dataset. The adopted methodology involved training ChatGPT to generate a Python script designed to calculate correlation matrices while also excluding specified missing or masked values. Moreover, the OpenAI API was incorporated to analyze and provide summaries of the correlations discovered.

Although the initial results exhibited a mixture of inconsistencies and potential biases in the generated summaries, this study successfully demonstrated the capacity of ChatGPT to assist researchers in the preliminary stages of data analysis. Despite their limitations, the generated summaries presented various testable hypotheses that warrant further investigation and validation through more comprehensive data analysis. These findings underscore the necessity for human supervision and critical evaluation when employing AI tools in research, because these technologies are not devoid of biases and constraints.

Importantly, the demonstrated use case of employing ChatGPT in this study is not limited merely to correlational analysis. The versatility and adaptability of ChatGPT in this study hold significant potential for a wide range of applications within housing research and beyond. Those applications include, but are not limited to, enhancing fair housing analysis, evaluating the effects of housing programs, identifying housing needs for vulnerable populations, assessing the effects of zoning regulations on housing affordability, monitoring and forecasting housing market trends, and facilitating stakeholder engagement and collaboration. The capacity of ChatGPT and similar AI tools to be integrated into robust computer programs for advanced statistical analyses enables researchers to address these research domains more effectively, uncovering deeper insights and novel patterns. As AI tools continue to evolve and improve, their applications in housing research and other social science fields hold the promise of driving innovation and enhancing the quality of research if researchers remain vigilant in addressing biases and limitations inherent in AI-generated output.

Often, when researchers are presented with large datasets, the greatest challenge they face is asking the right questions to yield the greatest insights from the data. A simple conversation with this tool can yield a customized computer program that could produce multiple insights and lines of inquiry instantly. Using AI in this manner has the potential to make research and publications more equitable by lowering the costs of large-scale data analysis for researchers at institutions with fewer resources or those lacking computer programming knowledge to build custom computer programs. By providing access to powerful AI tools such as ChatGPT, researchers can harness the benefits of rapid and automated data analysis without the need for extensive technical expertise, thereby democratizing the research process and encouraging a broader range of perspectives and voices in the realm of housing research and beyond.

To enhance the accuracy and relevance of the generated analyses, future research could consider supplying the AI model with a more comprehensive context, such as data dictionaries or supplementary information. As of this writing, OpenAI is just releasing plugins that connect an AI model to the internet, potentially allowing it to pull the necessary supplemental information on its own. In addition, exploring methods to mitigate biases and *hallucinations*, the term for falsely generated content, in AI output would prove beneficial in establishing the dependability and validity of generative AI tools in housing research and other social science fields.

Although generative AI tools such as ChatGPT exhibit potential in expediting and augmenting data analysis processes within housing research, it is crucial for researchers to remain cognizant of the possible biases and limitations inherent in AI-generated output. By incorporating additional contextual information and continually refining AI models, researchers can effectively leverage the capabilities of generative AI tools to support their data analysis efforts while maintaining the rigor and quality of their research.

Acknowledgments

The author would like to thank the reviewers for their helpful comments. The author would also like to thank the colleagues in the Office of Policy Development and Research who continue to develop and maintain the Picture of Subsidized Households dataset used in this project.

Author

Dylan Hayden is a social science analyst at the U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Office of the Chief Data Officer.

References

Hwang, Gwo-Jen, and Nian-Shing Chen. 2023. “Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions,” *Educational Technology & Society* 26 (2). https://doi.org/10.30191/ETS.202304_26(2).0014.

Kappel, Ellen S. 2023. “How Might Artificial Intelligence Affect Scientific Publishing?” *Oceanography* 36 (1): 5. https://doi.org/10.5670/oceanog.2023.113.

OpenAI. 2023. “GPT-4 Technical Report.” http://arxiv.org/abs/2303.08774.

Van Dis, Eva A. M., Johan Bollen, Robert van Rooij, Willem Zuidema, and Claudi L. Bockting. 2023. “ChatGPT: Five Priorities for Research,” *Nature* 614 (7947): 224–226.