Optimizing Real Estate Portfolios: The Role of Generative AI in Geographic Diversification*

Timothy Dombrowski
Assistant Professor of Finance
Department of Finance and Legal Studies
University of Missouri–St. Louis
tdombrowski@umsl.edu

Cayman Seagraves
Assistant Professor of Finance
Department of Finance
University of Tulsa
cayman-seagraves@utulsa.edu

April 18, 2025

*We are grateful for comments from participants at the Generative AI in Finance conference at Concordia University. In particular, we thank Paolo Giudici for his insightful feedback and discussion, as well as comments from Nasser Arshadi, Yiuman Tse, and Philip Seagraves.

Optimizing Real Estate Portfolios: The Role of Generative AI in Geographic Diversification

Abstract

This study investigates the data analysis capabilities of GPT-4o in real estate portfolio selection by integrating predictive modeling, model evaluation, and investment decision-making into a fully autonomous AI-driven framework. Unlike the earliest large language models (LLMs) that primarily process textual data or recent LLMs such as OpenAI's o1 and DeepSeek's R1, which are designed for complex reasoning, GPT-4o actively executes code and conducts quantitative analysis using the Code Interpreter tool. Leveraging a dataset of Zillow home price data and several predictive factors, the AI-generated portfolios achieved three of the four top-performing Sharpe ratios in our out-of-sample backtest. Further, we find that data obfuscation – removing city names, states, and dates – reduces geographic diversification and produces lower Sharpe ratios than the unobfuscated portfolios. Overall, our findings highlight the potential of generative AI in advancing data-driven portfolio management.

Keywords: Generative AI, Large Language Models, ChatGPT, Real Estate Investment, Portfolio Optimization, Geographic Diversification, Google Trends

JEL: C45, C53, G11, G17, R31, R32

"*I expect over the next five years, so much of what human investors and analysts can do can be better done with machine intelligence.*"

– *Greg Jenson, Co-Chief Investment Officer at Bridgewater, July 8, 2024*[1]

1 Introduction

The emergence of generative artificial intelligence (AI) is fundamentally reshaping the landscape of financial decision-making and portfolio management. This technological breakthrough arrives at a critical juncture in financial markets, where traditional approaches to portfolio selection face modern challenges. The market dominance of the "Magnificent Seven" tech stocks demonstrates a new iteration of a historically persistent pattern of asset concentration.[2] From 1926 to 2018, merely 4% of listed companies generated the entire net gain for the U.S. stock market (Bessembinder, 2018). Yet, poorly diversified active strategies tend to underperform market averages, highlighting the enduring tension between identifying outperforming assets and maintaining adequate diversification (Bessembinder, 2018).

Amid these challenges, generative AI offers promising solutions for investment strategy and corporate financial management. The transformative potential of this technology is exemplified by the Magnificent Seven themselves, which accounted for roughly 57% of the S&P 500's nearly $10 trillion market value increase in 2024 while investing heavily into AI development and deployment (Dulaney, 2025). Academic research demonstrates AI's effectiveness in financial applications, from predicting stock market movements (Lopez-Lira and Tang, 2023; Pelster and Val, 2024) to corporate policy and financial statement analysis (Jha et al., 2024; Kim et al., 2024). However, businesses face mounting pressure to move beyond experimentation and demonstrate concrete returns on their substantial AI investments (Broughton and Maurer, 2024; Lin, 2024). Despite these recent breakthroughs in equity markets, AI's potential for advancing real estate investment strategies remains largely unexplored.

[1] See Parmar and Burton (2024) for the full article and video from *Bloomberg*.

[2] The "Magnificent Seven" consists of Alphabet, Amazon, Apple, Meta, Microsoft, Nvidia, and Tesla. See these recent articles in the *New York Times* (Russell and Rennison, 2024) and *Wall Street Journal* (Singh, 2023) for more details.

Real estate markets present a compelling testing ground for examining AI's potential in investment strategy and portfolio optimization, given their unique physical characteristics and today's complex market dynamics. Geographic diversification in real estate investment, like diversification in stock portfolios, traditionally serves to spread risk and capture varying growth opportunities across regions and property types. This geographic spread is particularly important for managing exposure to environmental risks and natural disasters, as demonstrated by recent events like the California wildfires and Hurricanes Helene/Milton.

Recent shifts in U.S. population trends, driven by factors such as remote work and changing economic opportunities, have underscored the importance of strategic market selection, with states like Idaho and Florida experiencing rapid growth while others like New York and Illinois have seen population declines (Fitzpatrick and Beheraj, 2023). These demographic shifts highlight an increasingly uneven economic landscape where identifying high-growth real estate markets has become crucial for maximizing returns while maintaining appropriate diversification.

Given these unique market dynamics and the promising potential of AI in portfolio optimization, this study investigates whether generative AI models, specifically GPT-4o, can effectively identify and select high-growth real estate markets proposing a refined approach to geographic diversification. Building on recent advances in AI applications for financial markets (Jha et al., 2024; Kim, 2023; Ko and Lee, 2024; Lopez-Lira and Tang, 2023; Pelster and Val, 2024), we develop a novel methodology that integrates GPT-4o's code execution and predictive capabilities with real estate market selection. Our approach leverages a training dataset that combines variables from multiple authoritative sources, including Zillow, the U.S. Census, the Bureau of Labor Statistics (BLS), Freddie Mac, and Google Trends. That data is provided to GPT-4o, which analyzes the data and generates city selections for a real estate portfolio with a goal of achieving the maximum risk-adjusted return, or Sharpe ratio. To evaluate the performance of the AI-generated portfolios, we backtested them against several benchmarks from Zillow and the Case-Shiller Indices and find promising results for this approach.

This paper's contribution is as follows: (1) We conduct an experiment with GPT-4o to evaluate its ability to analyze real estate data, develop predictive models, and select cities for a residential real estate portfolio. By doing so, we propose a novel approach to geographic diversification that prioritizes strategic market selection over broad-based risk spreading. (2) We provide empirical evidence on the viability and effectiveness of using generative AI models to enhance real estate portfolio performance, offering academically rigorous and practically relevant insights.[3] (3) We evaluate the performance of these AI-generated real estate portfolios across time, and (4) we evaluate whether the obfuscation of city names, states, and dates has an impact on the city selections and performance of the portfolios, contributing to the literature documenting look-ahead bias in AI models (Levy, 2024).

To provide a preview of our main results, we find that the AI-generated portfolios achieved three of the four largest mean Sharpe ratios across our 12 out-of-sample backtests. In addition to demonstrating GPT-4o's capabilities in conducting quantitative analysis and generating investment decisions, this paper also explores a novel intersection between data obfuscation and the portfolios generated by the AI models. By comparing the city selections between the obfuscated and unobfuscated versions of the training data, we examine the impact of *data quality* on the analysis and decisions generated by AI. With higher-quality data (unobfuscated city names, states, and dates), we expect the AI to leverage this geographic/temporal information and to generate more diversified portfolios that achieve higher risk-adjusted returns. Our findings in Section 4.2 validate this expectation by demonstrating that the two versions of the training data produce very different sets of city selections and that the unobfuscated selections are more geographically dispersed. Then, in Section 5, we backtest these AI-generated portfolios and find that the unobfuscated portfolios do indeed perform better on a risk-adjusted basis.

As the landscape of financial markets becomes increasingly complex, leveraging AI for strategic market selection represents a significant advancement in portfolio management. This paper contributes to the growing body of literature on the application of AI in finance

[3]In addition to being the subject of our experiment, GPT-4o was also used for the purpose of generating sample code to conduct the analysis and create the tables/figures for the paper. In addition to this, we also used GitHub Copilot for coding assistance, as well as several other generative AI tools for various project-related tasks. All code generated by these AIs has been reviewed by the authors and verified to be correct for the analysis. See our declaration on generative AI usage at the end of the paper for more details.

and real estate, demonstrating how these technologies can be harnessed to allocate capital in real estate markets more efficiently. By conducting this generative AI experiment, our research advances the academic understanding of AI's ability to manage a real estate portfolio and provides practical guidance for investors aiming to outperform the market. In summary, the overall scope of our project can be summarized by the following research question:

> **Research Question**: *Can generative AI tools (specifically, GPT-4o) construct a data-driven strategy for selecting cities for a residential real estate portfolio, such that the AI-generated portfolios outperform various benchmarks?*

The remainder of this paper is as follows: Section 2 positions our study within the academic literature focused on applying AI to finance and real estate topics. Section 3 provides details on the datasets that we constructed for our generative AI experiment. Section 4 specifies the prompting structure for GPT-4o and backtesting strategy. Section 5 documents our results and comparisons. Section 6 concludes with a review of our findings and a discussion around potential future extensions of this project.

2 Background on AI, LLMs, and ChatGPT

With origins in science fiction and philosophy, the concept of AI has found its way into the modern technical lexicon through the development of complex statistical models that can be tailored to particular tasks. The explosive adoption of ChatGPT in late 2022 has sparked an AI boom evident across several of the largest tech giants. This has resulted in several competing large language models (LLMs) that have varying capabilities and focuses.

Mollick (2025) compares the capabilities of seven different LLMs, including OpenAI's o1 and DeepSeek's R1. He found that GPT-4o with Code Interpreter was the best for data analysis, executing code, and implementing advanced statistical techniques. Some other models, such as Claude and Gemini, have some data analysis capabilities. However, they fall short of GPT-4o. Since those capabilities are central to our research question, we use the GPT-4o model as the subject of our experiment.[4]

Our research question helps further develop a young branch of academic literature on the application of generative AI in financial contexts. With a focus on real estate finance and geographic diversification, our study demonstrates the current capabilities of modern AI in data analysis and business decision-making. To further position our paper within this literature, the remainder of this section discusses a variety of related studies in this newly-developing field.

2.1 AI in Finance

Prior to the recent surge in LLM-based AI tools, there are various other AI-related topics that have seen some interest in finance and real estate academia. One example is the application of machine learning models for statistical analysis and prediction.[5] Some other AI-related topics that have been explored include deep learning (Fang et al., 2024; Fischer and Krauss, 2018; Nazemi and Fabozzi, 2024), computer vision/image recognition (Glaeser et al., 2018; Hamilton and Johnson, 2023; Jiang et al., 2023), and explainable AI (XAI) (Erel et al., 2021; Krämer et al., 2023; Nazemi and Fabozzi, 2024).

Another sub-field of AI research investigates natural language processing (NLP) and textual analysis. In the finance literature, the work of Loughran and McDonald (2011, 2014, 2016) popularized their libraries for financial terms and sentiment analysis. These libraries have been integrated into the edgar R package (Lonare et al., 2021), which enables programmatic access to company filings in the SEC's EDGAR database and word count statistics that incorporate the Loughran and McDonald libraries.[6] Beyond these sentiment-based libraries, researchers have also explored tonal analysis of earnings calls (Blau et al., 2015; Davis et al., 2015; Price et al., 2012), part-of-speech tagging (Bhagwat et al., 2024),

[4]Specifically, we created an AI Assistant through the OpenAI API that uses the GPT-4o model with the Code Interpreter tool enabled. After uploading the training data, the files were passed to the assistant along with the prompt. See Section 4 for more details on the experimental design and backtesting procedure.

[5]There are far too many to list all of them here. So a non-exhaustive list of some recent examples is: Ban et al. (2018); Breuer and Steininger (2020); Gu et al. (2020); Kelly et al. (2024); Khandani et al. (2010); Liu and Pun (2022); Nazemi and Fabozzi (2024); Viriato (2019); Wan and Lindenthal (2023).

[6]See `https://cran.r-project.org/package=edgar` for more details.

and even the development of NLP techniques for identifying greenwashing (Bingler et al., 2024; Gorovaia and Makrominas, 2024).

In recent years, the development of LLMs has popularized another AI-related sub-field called generative AI. This concept of generative AI encompasses more than just text-based content from LLMs; it also involves image generation, video generation, and audio/music generation. The Generative Pre-trained Transformer (GPT) models created by OpenAI (Radford et al., 2018) are built on top of the Transformer model architecture introduced by Vaswani et al. (2017). Researchers have even developed some finance-tailored LLMs (Araci, 2019; Huang et al., 2023), which are built on Google's Bidirectional Encoder Representations from Transformers (BERT) algorithm (Devlin et al., 2018).

In regard to academic research on LLMs in finance, Chen et al. (2022) provides one of the earliest studies, which focuses on using ChatGPT and LLaMA to extract news sentiment for predicting stock returns. Dowling and Lucey (2023) explores the potential for using these LLM tools for conducting academic financial research. Fieberg et al. (2023) and Oehler and Horn (2024) explore the potential for ChatGPT to give sound financial advice in an advisory role. Similarly, Ko and Lee (2024) evaluates ChatGPT's ability to improve investment decisions and concludes that "ChatGPT can serve as an assistant or co-pilot to portfolio managers."

2.2 AI in Real Estate

On the real estate side, the most relevant studies before the LLM boom were focused on applying machine learning models for data analysis (Calainho et al., 2024; Lorenz et al., 2023; Viriato, 2019), applying textual analysis techniques to multiple listing service (MLS) data (Nowak and Smith, 2017), and building real-estate-specific dictionaries for textual analysis (Nowak et al., 2021). Similarly, Goodwin et al. (2018) apply textual analysis techniques and examine the connotation of keywords in real estate listings.

On the LLM side of real estate research, this is still a relatively unexplored area, and this paper aims to begin filling that gap. On the construction side of real estate, Prieto et al. (2023) investigates the use of ChatGPT for scheduling construction projects. Cheung (2024) focuses on the potential for using ChatGPT in property valuation processes. Haurum et al. (2024) develops an LLM-based real estate agent, and Gloria et al. (2024) create a real estate domain-specific LLM that outperforms more general LLMs, such as GPT 3.5 and Mistral 7B, when taking a 1,000+ question mutliple-choice exam. Seagraves and Seagraves (2024) looks into the potential impacts that LLMs and other AI tools can have in real estate education.

This paper further extends the literature on the application of generative AI in finance by exploring its capabilities in the context of real estate finance. Our experiment with GPT-4o tests not just the data analysis capabilities of a modern LLM, but also its ability to generate forecasts and select cities for a real estate portfolio. Additionally, we compare the city selections and portfolio performance across time and across two versions of the final dataset: (1) a version that obfuscates the city name, states, and dates and (2) a version that includes the unobfuscated city names, states, and dates. These comparisons provide insights as to whether any improved performance persists over time and whether data obfuscation impacts the city selections and portfolio performance. This latter aspect furthers the literature on evaluating the impact of data obfuscation in AI systems (Búadóttir et al., 2023; Decarolis et al., 2023; Levy, 2024). Lastly, when building our training dataset for the AI, we propose a novel approach to programmatically extracting city-month level real estate search interest from Google Trends. This is described in Section 3.5.

3 Data

This section reviews the data sources, cleaning process, and final datasets for our experiment. Our training dataset includes house price data from Zillow (Section 3.1), residential population data from the U.S. Census Bureau (Section 3.2), unemployment rates from the U.S. Bureau of Labor Statistics (Section 3.3), mortgage rates from Freddie Mac (Section 3.4), and Google Trends data for city-specific search queries within some of their real estate categories (Section 3.5). These are merged together to form a city-month panel of data that we slice into various time windows and provide to GPT-4o to conduct the experiment.

In addition to that training dataset, we also developed a backtesting dataset for evaluating the performance of the AI-generated portfolios against some benchmarks. Those benchmarks include several portfolios constructed from the Zillow data, as well as the S&P CoreLogic Case-Shiller Home Price Indices. These are discussed more in Section 3.6.

3.1 Zillow Housing Value Index (ZHVI)

The starting point for both of our datasets is the Zillow Home Value Index or ZHVI, which is available online from Zillow Research (2024).[7] This study uses the default option on the webpage – *ZHVI All Homes (Single-Family Residential, Condo, Co-op) Time Series, Smoothed, Seasonally Adjusted* – at the city-month level. We focus on 433 of the 500 largest cities in the U.S., as per Zillow's SizeRank variable, over the 235-month period from January 2004 through July 2023.[8]

In other words, the Zillow data provides us with more than 100 thousand city-month observations. These are numerically summarized in Table 1. Figure 1 depicts the distribution of this ZHVI panel as a series of annual boxplots that summarize the city-level cross-sections of the average monthly ZHVIs in each city-year. For some specific time series curves, Figure 2 shows several monthly ZHVI series for several benchmark portfolios that are constructed from the city-month level ZHVI data. These include the national ZHVI (ZHVI_USA) and several averages across the top N cities, where $N \in \{5, 10, 20, 40\}$.

For our final training dataset that is provided to GPT-4o, we keep these ZHVI values in their raw dollar-denominated units and allow the AI to determine how it further cleans or transforms the data. For our backtesting of the AI-generated portfolios and benchmarks, we transform the ZHVI values into continuously-compounded growth rates. This process is

[7] Specifically, the ZHVI reflects the typical value for homes in the 35th to 65th percentile range, which provides insights into market dynamics across various regions and property types. These monthly estimates of "the typical home value" are available at several different geography levels, ranging from the single national U.S. ZHVI time series down to panel data at the ZIP-month level and neighborhood-month level. These indices are also available as smoothed, seasonally adjusted values and as raw values.

[8] These filters ensure matching temporal coverage with population data from the Census and reliable city matches with the Google Trends data. Any missing ZHVI values in the middle of a city's time series were imputed using linear interpolation. However, there were still 44 cities with some missing values in the ZHVI data. Initially, we had planned to keep those cities in the final set; however, the missing values would sometimes result in problems with GPT-4o's analysis and city selection. Thus, we dropped those cities from our final set to ensure no missing values for any variables.

described in more detail in Section 4.3.

3.2 Census Population Estimates

In addition to the Zillow data, we also collected several predictors for the AI to use for its analysis and forecasting. The first among these is city population estimates from the U.S. Census Bureau (2024). Their intercensal estimates are available at an annual frequency, representing residential populations on July 1 of each year. As of this writing, these estimates are available at the city level from 2000–2023.

The full annual series was aggregated across three separate tables: 2000–2009, 2010–2019, and 2020–2023.[9] The annual observations for each city were then mapped to July of each year and linearly interpolated down to a monthly frequency. These city-month level populations were then merged to the training dataset. By incorporating this population data as a predictor, the AI will be able to use this data to model the relationship between population trends and real estate values. Table 1 presents summary statistics for this predictor, and Figure 3 depicts annual boxplots of the city-level population estimates with a log-scale y-axis to better depict the distribution.

3.3 Unemployment Rates

Next, we incorporated monthly state-level unemployment rates from the U.S. Bureau of Labor Statistics (2024) into the training dataset. These seasonally-adjusted series were downloaded via FRED using the fredr R package (Boysel and Vaughan, 2021), and then merged to the training dataset at the state-month level. With this data, the AI can control for general economic conditions across states over time when generating its city selections. Table 1 presents summary statistics for this predictor, and Figure 4 shows annual boxplots for the unemployment rates at the state level.

[9]See the reference item (U.S. Census Bureau, 2024) for the specific Census variable codes of each table used.

3.4 Mortgage Rates

Another relevant predictor for residential real estate values is mortgage rates. The MORTGAGE30US variable in FRED represents the monthly 30-year fixed mortgage rate from Freddie Mac (2024). This was also downloaded via the fredr package (Boysel and Vaughan, 2021) and then merged to the training dataset at the month level since there is no geographic variation for this variable. Table 1 presents summary statistics for this predictor, and Figure 5 shows the dynamics for this time series. By including data on mortgage rates, the AI's predictions can incorporate modeling of the relationship between mortgage rates and property values, which is well-established in the academic literature (Carrillo et al., 2023; Chambers et al., 2016; Forster and Sun, 2024; Harris, 1989).

3.5 Google Search Volume Index (GSVI)

Next, we collected data from Google Trends (2024), which has been widely used in finance and real estate research as a measure of public interest in specific search queries and topics (Chen et al., 2021; Dombrowski et al., 2020; Irresberger et al., 2015; Liu and Pun, 2022; Meshcheryakov, 2018; Vlastakis and Markellos, 2012; Wu and Brynjolfsson, 2015). To ensure that our dataset accurately reflects local real estate interest across different cities, we systematically selected the appropriate search topic codes using the process described below and gathered the corresponding Google Trends data using the pytrends Python package (Hogue and DeWilde, 2023).

The process began by identifying Google Trends topic codes for each city. These topic codes represent the options that appear in the web interface as suggestions once a user starts entering text. The actual topic codes are text strings that look like this: */m/0d35y*, which is the topic code for Phoenix, Arizona. Accompanying each topic code is a topic code type, which indicates a category for the topic code. In the case of Phoenix, that topic code type is "City in Arizona" indicating the geographic region and not a fictional immortal bird.

After identifying the "City of" topic codes for each of the cities in our sample,[10] we

[10]Most of these topic codes are of the form: "City in State." However, there are some other variations

validated the process by ensuring that the topic code type contained the text of the city's state name. Our algorithm produced validated matches for 497 of our initial 500 cities. The remaining cases included Washington D.C. (no state name), which was correctly matched to "Capital of the United States of America," and two cases where only a "Topic" match was available.[11] In other words, we effectively achieved a 100% match rate for identifying the correct topic codes.

Using those city topic codes, we pulled the Google Search Volume Index (GSVI) for each of those queries over the 2004–2023 period. These monthly GSVI series are scaled separately for each city and range from 0–100 with 100 representing the month with peak search interest.[12] The geographic scope of the query was the entire U.S., and we considered two real estate categories available from Google Trends: *Real Estate* and *Real Estate Listings*. Thus, each city's monthly GSVI series represents search interest from around the U.S. about the city within a real estate context.

After the data collection, the GSVI data was processed into a city-month panel format, and then merged to the training dataset. This results in a unified dataset that captures both housing market dynamics, population dynamics, economic controls, and public interest in real estate across 433 U.S. cities and 235 months. To summarize the this data, Table 1 shows the numeric summary statistics, Figure 6 presents annual boxplots showing the distribution across cities in the "Real Estate" category, and similarly, Figure 7 depicts the boxplot trends for the "Real Estate Listings" category.

---

that are needed to improve the match rate. So in addition to "City in" topic code types, we also consider "Municipality in," "Town in," "Township in," "Borough in," "Suburb in," "Village in," "Census-designated place in," "Capital of," (for D.C.) "Unincorporated community in," and "County in."

[11] Those two cases are Chicago, Illinois, which only had a topic code type of "Topic" for Chicago, and Clinton Township, Michigan, which similarly only had a "Topic" match. The former case is included in our final datasets, and the latter case is dropped due to not having a match in the Census population data.

[12] Although it is possible to pull multiple queries at once to gain some degree of relative comparability between cities, the GSVI values are limited to whole numbers between 0–100. Thus, we would lose much of the variation in smaller cities with this approach.

3.6 Benchmarks

In addition to the training dataset described above, we create another dataset for backtesting the portfolios generated by GPT-4o against several benchmarks. Several of those benchmarks were briefly discussed in Section 3.1 and are constructed from the Zillow data using the SizeRank variable to identify the largest cities.[13] Those are:

- Equal-weighted portfolio across Top 5 Cities (ZHVI_5)
- Equal-weighted portfolio across Top 10 Cities (ZHVI_10)
- Equal-weighted portfolio across Top 20 Cities (ZHVI_20)
- Equal-weighted portfolio across Top 40 Cities (ZHVI_40)
- The U.S. national ZHVI series (ZHVI_USA)[14]

The full time series for each of those is presented in Figure 2, which shows that the portfolios concentrated in the largest cities tend to have more expensive residential properties.

To go along with those benchmarks, we include several additional real estate benchmarks to compare with the AI-generated portfolios. The S&P CoreLogic Case-Shiller Composite Home Price Indices provide an additional set of benchmarks for comparison. Those are:

- S&P CoreLogic Case-Shiller 10-City Composite Home Price Index (SPCS_10) (S&P Dow Jones Indices LLC, 2024a)
- S&P CoreLogic Case-Shiller 20-City Composite Home Price Index (SPCS_20) (S&P Dow Jones Indices LLC, 2024b)
- S&P CoreLogic Case-Shiller U.S. National Home Price Index (SPCS_USA) (S&P Dow Jones Indices LLC, 2024c)

[13]This Zillow SizeRank variable lists the 5 largest cities as: (1) New York City, NY, (2) Los Angeles, CA, (3) Houston, TX, (4) Chicago, IL, and (5) San Antonio, TX. However, the 2023 intercensal estimates suggest that the top 5 are: (1) NYC, (2) LA, (3) Chicago, (4) Houston, and (5) Phoenix, AZ. Since we used the Zillow SizeRank variable for our initial set of the 500 cities, we use it again here for forming our benchmarks.

[14]This national ZHVI series is bundled with the metro-level ZHVI data on the Zillow Research webpage.

Those variables are also downloaded via the fredr package (Boysel and Vaughan, 2021) and are merged to the ZHVI benchmarks to form the backtesting dataset. Since the scaling of the Case-Shiller indices is set to an arbitrary value of 100 in January 2000, these benchmarks are omitted from the dollar-denominated plot in Figure 2. Instead, Figure 8 presents the scaled indices for all of the benchmarks together, beginning at 100% in January 2004 for all indices. This depicts the cumulative growth of each benchmark over the full sample period.

4 Experimental Design

For our generative AI experiment, we started with the cleaned training dataset and split it into 12 training windows, each with five years (60 months) of data. The first of these training windows ranged from January 2004 through December 2008, and the windows roll in annual steps until the 2015–2019 window. Each of these training subsets was uploaded to GPT-4o, which was prompted to analyze the data and make city selections for a portfolio. Then for our backtesting process, we evaluated the performance of each window's selections over the following three years (36 months). Thus, for the 2004–2008 training window, we compared the portfolios against the benchmarks over the 2009–2011 period. Then we trained on the 2005–2009 data and backtested on the 2010–2012 period, and so on until the 2015–2019 window, which was backtested over the 2020–2022 period.

As described in Section 3, the training dataset consisted of the following variables:

- **ZHVI**: Zillow Home Value Index for the region
- **POP**: Population of the city
- **UnemployRate**: Unemployment rate (%)
- **MortgageRate**: Average 30-year fixed mortgage rate
- **GSVI:RealEstate**: Google search intensity for general real estate terms
- **GSVI:RealEstateListings**: Google search intensity for real estate listings

In addition to those numeric variables, we included identifiers for the city and month dimensions of the panel data. When doing this, we created two versions of the training subsets: (1) an unobfuscated version that includes the original city names, states, and dates, and (2) an obfuscated version where the city names and states are replaced with a randomly-assigned numeric identifier (CityID $= 1, 2, 3, \ldots, 433$), and the actual dates are replaced with YearMonthID $= 1, 2, 3, \ldots, 60$. These are defined as follows:

- **CityID**: A unique randomly-assigned identifier for each city in the dataset
- **YearMonthID**: A numerical identifier for the time period
- **RegionName**: Name of the city **excluded from obfuscated version*
- **State**: State of the city **excluded from obfuscated version*
- **Date**: Calendar date of the observation **excluded from obfuscated version*

Since city population data is provided, a reasoning AI could potentially infer the city names and states from that variable. That would suggest that we would see similar predictions and city selections between the obfuscated and unobfuscated datasets. However, if we assume that the AI is not capable enough to deducing the city identities, then we might see different city selections depending on the version of the dataset provided. The source of such differences could be due to some probabilistic aspects of GPT-4o's responses, or perhaps even from the AI incorporating additional information about the cities from its corpus into the models.

After preparing the training subsets for a given window, they were uploaded to OpenAI as comma-separated value (CSV) files. We then prompted GPT-4o to analyze the data and generate investment recommendations for the next three years. Specifically, we requested that the AI select cities by their CityID to create four portfolios that each have different sizes: 5 cities, 10, cities, 20 cities, and 40 cities. Those respective portfolios are denoted as GPT_5, GPT_10, GPT_20, and GPT_40 for selections from the unobfuscated data and as OGPT_5, OGPT_10, OGPT_20, and OGPT_40 for the obfuscated data portfolios.

After receiving the city selections from GPT-4o in its response, we first compared the selections from the obfuscated vs. unobfuscated training datasets. Then, we loaded those selections into our backtesting codes and constructed an equal-weighted portfolio for each set of city selections. Those portfolios were then compared against the various benchmarks. The specifics for these comparisons are discussed in Section 4.3 after we provide an overview of our prompt in Section 4.1 and evaluate the impact of data obfuscation on the city selections in Section 4.2. Appendix A includes two diagrams that provide a visual overview of the experimental design and the architecture of our real estate portfolio management AI agent.

4.1 Prompt Construction and Delivery

The full text for our prompt to GPT-4o can be found in B. This section will provide an overview of the prompt structure. To summarize the process, we first created an AI Assistant on the OpenAI Platform.[15] The assistant used the GPT-4o model with the Code Interpreter tool enabled, which allows for the assistant to execute code and conduct data analysis. After uploading the training data via the OpenAI API, we provided the text-based prompt to GPT-4o, which is tailored to our specific dataset and context.

The prompt begins by instructing the AI to assume the role of a PhD-level research assistant. It then provides a brief description of each variable in the data file. After describing some of the nuances around the different variables, we task it with developing 4–5 different predictive models and to use cross-validation for evaluating their performance.[16] We instruct it to use the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and mean squared error (MSE) as metrics to compare between the candidate models.

After comparing those criteria, our prompt instructs the AI to select the best model and to use it to create forecasts for each city over the next three years. Using those forecasts, it is tasked with selecting cities to form the portfolios. To provide the AI with more guidance and domain knowledge for this financial context, this component of the prompt instructs the AI to select cities with a goal of attaining the largest risk-adjusted return (Sharpe ratio).

[15] https://platform.openai.com/docs/assistants/overview

[16] We instruct the AI to use cross-validation, but do not specify how many folds that it should use.

The final aspects of the prompt include some reminders about our desired structure for the response, which help ensure that we get meaningful responses and city selections. This includes an example of the specific Python syntax that allows us to directly parse the city selections from the API response. Those vectors of CityID selections were then compared across the obfuscated/unobfuscated dimension and processed into our backtesting strategy.

4.2 Evaluating the Effects of Data Obfuscation

After parsing the city selections from GPT-4o, we first evaluate whether or not the obfuscation of city names, states, and dates leads to different selections. To do this, we first count the number of shared cities between the obfuscated and unobfuscated versions. Those results are presented in Table 2, which shows that the shared city selections between the two versions are typically quite low. Figure 9 transforms those counts into city match rates and shows that no pair of obfuscated/unobfuscated portfolios has more than 20% overlap. The single 20% case is for 1 common city between GPT_5 and OGPT_5 for the 2015:2019 training window. Since we find that the two versions of the training data produce mostly different sets of city selections, we retain both sets of portfolios for our backtesting analysis.

In addition to simply counting the number of cities in common, we also examined the most frequently selected cities across all portfolios and windows. Those counts are done separately for the obfuscated and unobfuscated versions, and we create top 10 lists for each. Those are presented in Table 3. Interestingly, the unobfuscated data city selections appear to be more diversified, with the most commonly selected city (Edmond, Oklahoma) only being in 9 of the 48 potential portfolios.[17] For the obfuscated data, San Francisco, California, was selected in 15 times, and 5 cities have more than 10 selections.

Beyond the counts in Table 3, another observation is that the unobfuscated top 10 appears to be more geographically diversified as well. The unobfuscated top 10 spans cities across 8 different states, whereas the obfuscated top 10 only spans 4 states. Although both top 10's include New York City and Los Angeles, the unobfuscated version contains several smaller cities, such as Edmond, OK, and Overland Park, Kansas. On the other hand, the obfuscated

[17] Each count in Table 3 is across all 4 portfolios in each of the 12 training windows. Thus, $4 \times 12 = 48$.

version has 6 California cities in the top 10, as well as NYC, Detroit, and Chicago.

4.3 Backtesting Strategy

After extracting the city selections from the GPT-4o responses, we imported them into our backtesting program and constructed equal-weighted portfolios across the cities that were selected. Those portfolios were then compared with the various benchmark portfolios described in Section 3.6 over the three-year period following the respective training window.

To do this, we calculated the continuously-compounded growth rates for each city-month in the Zillow data, as well as for the benchmarks. Equation 1 transforms city/benchmark i from a monthly (t) series of index levels ($x_{i,t}$) into a series of housing returns.

$$MonthlyReturns_{i,t} = \log\left(\frac{x_{i,t}}{x_{i,t-1}}\right) \tag{1}$$

After calculating the monthly housing returns, we transformed them into annualized excess returns. This was done as in Equation 2 where $RiskFree_t$ is the risk-free rate, which we proxy with the 3-Year Treasury yield to match the time horizon of our testing window.[18]

$$ExcessReturns_{i,t} = 12 \cdot MonthlyReturns_{i,t} - RiskFree_t \tag{2}$$

These excess returns were then used to compute Sharpe ratios for each of the portfolios in each testing window. This was done as in Equation 3 where w indicates the testing window, T_w is the final month of that window, and $\sigma_{i,w}$ is the standard deviation of the excess returns for that window. Thus, $T_w = 36$ for our 3-year testing windows.

$$SharpeRatio_{i,w} = \frac{\frac{1}{T_w}\sum_{t=1}^{T_w} ExcessReturns_{i,t}}{\sigma_{i,w}} \tag{3}$$

After calculating the backtesting Sharpe ratios for each of the AI-generated portfolios and the benchmarks, we then ranked the portfolios based on their Sharpe ratios (1=best, 16=worst).

[18]Specifically, we merge the DGS3 variable from FRED as the monthly averages of the daily values (Board of Governors of the Federal Reserve System (US), 2024).

After doing this for each of the backtesting windows, we then averaged the rankings across the backtesting windows to determine a final ranking based on those means.

In addition to comparing the Sharpe ratios between the GPT portfolios and the benchmarks, we also compared the cumulative returns of each portfolio over each testing window. Those are calculated in Equation 4 using the cumulative sum of the monthly returns.

$$CumulativeRets_{i,t} = 100 * \exp\left(\sum_{\tau=1}^{t} MonthlyReturns_{i,\tau}\right) \tag{4}$$

For each testing window, the cumulative returns of the portfolios are compared in a horse race for the respective testing period. These results are simplified by extracting the total return of each portfolio at the end of the testing windows. To formalize this in our nomenclature, let $TotalReturns_{i,w} = CumulativeRets_{i,T_w}$. Those total returns were compared and ranked just as the Sharpe ratios were. All those results are presented and discussed next.

5 Backtesting Results

The full set of results from our experiment and backtesting procedure are summarized across Tables 4–7 and Figures 10–14. This section will discuss these results and highlight our most interesting findings.

5.1 Total Returns

Our first comparison between the AI-generated portfolios and benchmarks is with the total (cumulative) returns of each portfolio over the various backtesting windows. Figure 10 traces the cumulative returns for each portfolio over each of the testing windows. To simplify those results, the ending values (total returns) from each window are presented in Table 4, and Table 5 transforms those into ranks from largest (1) to smallest (16) for each window. These tables also include the averages across all testing windows, which is how the rows of the tables have been sorted. The results from Table 4 are presented visually in Figure 11 where each of portfolio types is depicted using a different line type. The AI-generated portfolios are presented as solid lines, the ZHVI benchmarks are dashed lines, and the SPCS benchmarks are dotted lines. Figure 10 also uses these same line types.

Those figures begin with the testing window of 2009–2011, which evaluates the performance of the portfolios selected from the training window of 2004–2008. Thus, most portfolios begin with net depreciation during the first two windows before appreciating for the remainder of the testing windows. The average three-year returns across all 12 testing windows ranged from 15–21%, which represents annual growth of 5–7%. The top performing portfolio by this metric was the 20-city OGPT portfolio, which was generated by the GPT-4o selections using the obfuscated training data. It outperformed the second-place ZHVI_20 portfolio by 70 basis points, on average.

One of the key elements of this paper is examining the impact of data obfuscation on AI's city selections, and on the performance of those portfolios. In regard to performance between those two versions of the AI-generated portfolios, Figure 12 presents the difference in the total returns between each of the GPT-OGPT portfolio pairs. In general, there does not appear to be a clear winner between the two for this metric. The 2011–2013 testing window saw all four unobfuscated portfolios outperform their obfuscated counterparts. However, for the 2012–2014 testing window, the opposite was true with the obfuscated portfolios outperforming the unobfuscated portfolios.

In regard to the total return ranks in Table 5, the OGPT portfolios tended to perform better than the GPT portfolios. However, these differences are not consistent across time and do not reflect a risk-adjusted return. This motivates our next analysis, which examines the Sharpe ratios of the portfolios to provide a comparison that incorporates both the risk and return aspects of portfolio performance.

5.2 Sharpe Ratios

In addition to comparing the total returns of the AI-generated portfolios and benchmarks, we also compare the Sharpe ratios over the backtesting periods. Table 6 reports the Sharpe ratios numerically, Figure 13 depicts these ratios visually, and Table 7 presents the Sharpe ratio rankings for each window. The Sharpe ratio comparisons provide insights into the risk-adjusted performance of AI-generated portfolios.

The GPT_40 portfolio consistently ranks among the highest in terms of risk-adjusted returns, with the largest mean Sharpe ratio of 1.89, outperforming even the more diversified benchmarks, ZHVI_USA index (1.87) and SPCS_USA (1.28). This suggests that while absolute returns may not be consistently higher for AI-generated portfolios, their ability to balance risk and return can be comparable to, if not better than, traditional diversification strategies.

In regard to the Sharpe ratio rankings in Table 7, we find that the unobfuscated GPT portfolios are frequently among the top ranks. In addition to the top performing GPT_40 portfolio, two others (GPT_20 and GPT_10) also ended up in the top 5 in terms of risk-adjusted returns. These findings suggest that the unobfuscated AI-generated portfolios show promise in generating diversified city selections that perform well on a risk-adjusted basis.

As for the comparison between the obfuscated and unobfuscated AI-generated portfolios, it appears that the more diversified GPT portfolios perform better in terms of Sharpe ratios, whereas the obfuscated (OGPT) portfolios tended to perform better on a total return basis. Figure 14 shows the differences between the Sharpe ratios for each pair of GPT-OGPT portfolios. Although the differences between these is close to zero for the three most recent testing windows, there are several windows in the middle where the GPT portfolios achieve substantially larger Sharpe ratios than their OGPT counterparts.

This finding confirms that removing explicit geographic and time identifiers severely hinders AI's ability to construct optimal real estate portfolios, likely because it disrupts its capacity to detect high-growth areas, economic cycles, and regional risk factors. Another takeaway from the Sharpe ratio analysis is that larger portfolios (in regard to the number of cities) typically demonstrated stronger risk-adjusted returns than smaller portfolios. This suggests that AI benefits from a broader selection pool, which may help mitigate idiosyncratic risks tied to individual real estate markets.

6 Conclusion

This study demonstrates the capabilities of GPT-4o in conducting data analysis, model evaluation, and investment decision-making. To summarize the results from our generative AI experiment on real estate portfolio diversification, we find that GPT-4o is capable of analyzing real estate data and generating city selections that outperform various benchmarks on a risk-adjusted basis. Additionally, we find that the obfuscation of city names, states, and dates in the training dataset results in substantially different city selections from the AI. Those differences suggest that the unobfuscated training data generated more geographically diversified portfolios and better performing portfolios. Specifically, the unobfuscated AI-generated portfolios achieved three of the four largest mean Sharpe ratios across the out-of-sample backtest, outperforming our benchmarks. Our top-performing portfolio was GPT_40 (the 40-city AI-generated portfolio from unobfuscated data), which achieved the largest average Sharpe ratio (1.89) and outperformed all eight benchmarks.

Beyond these findings from our experiment, we proposed a novel approach for portfolio optimization that leverages generative AI to analyze data and inform business decisions. Additionally, we developed an algorithm for identifying Google Trends city topic codes and building a city-month panel of city-specific real estate search interest. This was combined with several other predictors to form the training data that GPT-4o analyzed. Lastly, by examining the impact of obfuscating the city names, states, and dates, this research is among the first to explore the impact of data quality on AI's ability to conduct quantitative analysis and provide investment recommendations.

It is possible that future generative AI models may produce even more effective analysis. For example, including additional predictors or a larger set of cities to select from could generate better-performing portfolios. Some other ideas for expanding on this experimental design are to compare selections across multiple data-analysis-capable LLMs or to experiment more with the prompt design and structure.[19]

[19] For example, one idea that we didn't get a chance to implement for this draft is to expand the prompt to include a request for the AI to create a table that summarizes the fit and performance metrics for the different predictive models that it considered. One could even instruct the AI to generate the LaTeX code to

In conclusion, the fact that we were able to even conduct this experiment is a testament to the progress that has been made in recent decades with computer processing, and particularly over the past few years since generative AI and LLMs have become mainstream technology in daily life. Our findings suggest that modern generative AI is capable of much more than just simple language-related tasks. In addition to conducting quantitative analysis on the training data, it was able to evaluate several statistical models and generate city selections that performed impressively when compared against the various benchmarks.

typeset that table in a formal manuscript.

References

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.

Ban, G.-Y., N. El Karoui, and A. E. B. Lim (2018). Machine Learning and Portfolio Optimization. *Management Science 64*(3), 1136–1154.

Bessembinder, H. (2018). Do stocks outperform Treasury bills? *Journal of Financial Economics 129*(3), 440–457.

Bhagwat, V., S. E. Shirley, and J. R. Stark (2024). Task-oriented speech and information processing. *Journal of Banking & Finance 161*, 107095.

Bingler, J. A., M. Kraus, M. Leippold, and N. Webersinke (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance 164*, 107191.

Blau, B. M., J. R. DeLisle, and S. M. Price (2015). Do sophisticated investors interpret earnings conference call tone differently than investors at large? evidence from short sales. *Journal of Corporate Finance 31*, 203–219.

Board of Governors of the Federal Reserve System (US) (2024). Market Yield on U.S. Treasury Securities at 3-Year Constant Maturity, Quoted on an Investment Basis [DGS3], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 23, 2025.

Boysel, S. and D. Vaughan (2021). fredr: An R Client for the 'FRED' API.

Breuer, W. and B. I. Steininger (2020). Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms. *Journal of Business Economics 90*(7), 963–974.

Broughton, K. and M. Maurer (2024, March). CFOs Tackle Thorny Calculus on Gen AI: What's the Return on Investment? *Wall Street Journal*.

Búadóttir, T., O. Mascio, and J. Eckroth (2023). Kira: A Financial Chatbot Using ChatGPT and Data Obfuscation. *Journal of Computing Sciences in Colleges 39*(3), 277–294.

Calainho, F. D., A. M. van de Minne, and M. K. Francke (2024). A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate. *The Journal of Real Estate Finance and Economics 68*(4), 624–653.

Carrillo, P. E., W. M. Doerner, and W. D. Larson (2023). House Price Markups and Mortgage Defaults. *Journal of Money, Credit and Banking 55*(4), 747–782.

Chambers, D. R., Q. Lu, Y. Lu, Y. Quan, and G. Xia (2016). The Effects of Housing Price Volatility on Mortgage Rates. *Journal of Housing Research 25*(1), 17–38.

Chen, H.-Y., H.-C. Chen, and C. W. Lai (2021). Internet search, fund flows, and fund performance. *Journal of Banking & Finance 129*, 106166.

Chen, Y., B. T. Kelly, and D. Xiu (2022). Expected Returns and Large Language Models. Available at SSRN 4416687.

Cheung, K. S. (2024, January). Real Estate Insights Unleashing the potential of ChatGPT in property valuation reports: the "Red Book" compliance Chain-of-thought (CoT) prompt engineering. *Journal of Property Investment & Finance 42*(2), 200–206.

Davis, A. K., W. Ge, D. Matsumoto, and J. L. Zhang (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies 20*(2), 639–673.

Decarolis, F., G. Rovigatti, M. Rovigatti, and K. Shakhgildyan (2023). Artificial Intelligence & Data Obfuscation: Algorithmic Competition in Digital Ad Auctions. Available at SSRN 4660391.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Dombrowski, T., D. Ratnadiwakara, and V. C. Slawson Jr. (2020). The FIMA NFIP's Redacted Policies and Redacted Claims Datasets. *Journal of Real Estate Literature 28*(2), 190–212.

Dowling, M. and B. Lucey (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters 53*, 103662.

Dulaney, C. (2025, January). How Reliant Are Markets on the AI Trade? *Wall Street Journal*.

Erel, I., L. H. Stern, C. Tan, and M. S. Weisbach (2021, 04). Selecting Directors Using Machine Learning. *The Review of Financial Studies 34*(7), 3226–3264.

Fang, Y., D. Luo, and Z. Yao (2024). Belief dispersion in the Chinese stock market and fund flows. *Journal of Banking & Finance 166*, 107252.

Fieberg, C., L. Hornuf, and D. J. Streich (2023). Using GPT-4 for Financial Advice. *CESifo Working Paper No. 10529*. Available at SSRN 4488891.

Fischer, T. and C. Krauss (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research 270*(2), 654–669.

Fitzpatrick, A. and K. Beheraj (2023, April). See the fastest growing (and shrinking) U.S. states. *Axios*.

Forster, R. and X. Sun (2024). Heterogeneous Effects of Mortgage Rates on Housing Returns: Evidence from an Interacted Panel VAR. *The Journal of Real Estate Finance and Economics 69*(3), 477–504.

Freddie Mac (2024). 30-Year Fixed Rate Mortgage Average in the United States [MORTGAGE30US], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 11, 2025.

Glaeser, E. L., M. S. Kincaid, and N. Naik (2018, October). Computer Vision and Real Estate: Do Looks Matter and Do Incentives Determine Looks. Working Paper 25174, National Bureau of Economic Research.

Gloria, B., J. Melsbach, S. Bienert, and D. Schoder (2024). Real-GPT: Efficiently Tailoring LLMs for Informed Decision-Making in the Real Estate Industry. *Journal of Real Estate Portfolio Management 0*(0), 1–17.

Goodwin, K. R., B. D. Waller, and H. S. Weeks (2018). Connotation and Textual Analysis in Real Estate Listings. *Journal of Housing Research 27*(2), 93–106.

Google Trends (2024). Google Search Volume Index (GSVI), Various Queries and Parameters. Retrieved via the pytrends Python package on January 11, 2025.

Gorovaia, N. and M. Makrominas (2024). Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*. forthcoming.

Gu, S., B. Kelly, and D. Xiu (2020, 02). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies 33*(5), 2223–2273.

Hamilton, T. L. and E. B. Johnson (2023). The amenity value of natural views. *Real Estate Economics 51*(5), 1079–1107.

Harris, J. C. (1989). The effect of real rates of interest on housing prices. *The Journal of Real Estate Finance and Economics 2*(1), 47–60.

Haurum, K. R., R. Ma, and W. Long (2024). Real Estate with AI: An agent based on LangChain. *Procedia Computer Science 242*, 1082–1088. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024).

Hlavac, M. (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

Hogue, J. and B. DeWilde (2023). pytrends: Pseudo API for Google Trends.

Huang, A. H., H. Wang, and Y. Yang (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research 40*(2), 806–841.

Irresberger, F., J. Mühlnickel, and G. N. F. Weiß (2015). Explaining bank stock performance with crisis sentiment. *Journal of Banking & Finance 59*, 311–329.

Jha, M., J. Qian, M. Weber, and B. Yang (2024, February). ChatGPT and Corporate Policies. *Chicago Booth Research Paper No. 23-15*.

Jiang, J., B. Kelly, and D. Xiu (2023). (Re-)Imag(in)ing Price Trends. *The Journal of Finance 78*(6), 3193–3249.

Kelly, B., S. Malamud, and K. Zhou (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance 79*(1), 459–503.

Khandani, A. E., A. J. Kim, and A. W. Lo (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance 34*(11), 2767–2787.

Kim, A., M. Muhn, and V. Nikolaev (2024). Financial Statement Analysis with Large Language Models.

Kim, J. H. (2023). What if ChatGPT were a quant asset manager. *Finance Research Letters 58*, 104580.

Ko, H. and J. Lee (2024). Can ChatGPT improve investment decisions? From a portfolio management perspective. *Finance Research Letters 64*, 105433.

Krämer, B., C. Nagl, M. Stang, and W. Schäfers (2023). Explainable AI in a Real Estate Context – Exploring the Determinants of Residential Real Estate Values. *Journal of Housing Research 32*(2), 204–245.

Levy, B. (2024). Caution Ahead: Numerical Reasoning and Look-ahead Bias in AI Models. *Available at SSRN* (5082861).

Lin, B. (2024, October). Companies Had Fun Experimenting With AI. Now They Have to Show the Returns. *Wall Street Journal*.

Liu, R. and C. S. Pun (2022). Machine-Learning-enhanced systemic risk measure: A Two-Step supervised learning approach. *Journal of Banking & Finance 136*, 106416.

Lonare, G., B. Patil, and N. Raut (2021). edgar: An R package for the U.S. SEC EDGAR retrieval and parsing of corporate filings. *SoftwareX 16*, 100865.

Lopez-Lira, A. and Y. Tang (2023). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. Available at SSRN 4412788.

Lorenz, F., J. Willwersch, M. Cajias, and F. Fuerst (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics 51*(5), 1178–1208.

Loughran, T. and B. McDonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. and B. McDonald (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance 69*(4), 1643–1671.

Loughran, T. and B. McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research 54*(4), 1187–1230.

Meshcheryakov, A. (2018). Using Online Search Queries in Real Estate Research with an Empirical Example of Arson Forecast. *Journal of Real Estate Literature 26*(2), 331–361.

Mollick, E. (2025). Which AI to Use Now: An Updated Opinionated Guide. *One Useful Thing*.

Nazemi, A. and F. J. Fabozzi (2024). Interpretable machine learning for creditor recovery rates. *Journal of Banking & Finance 164*, 107187.

Nowak, A. and P. Smith (2017). Textual Analysis in Real Estate. *Journal of Applied Econometrics 32*(4), 896–918.

Nowak, A. D., B. S. Price, and P. S. Smith (2021). Real Estate Dictionaries Across Space and Time. *The Journal of Real Estate Finance and Economics 62*(1), 139–163.

Oehler, A. and M. Horn (2024). Does ChatGPT provide better advice than robo-advisors? *Finance Research Letters 60*, 104898.

Parmar, H. and K. Burton (2024, July). Bridgewater Co-CIO Sees AI Adding 'Incredible Strength' for Investors. *Bloomberg*.

Pelster, M. and J. Val (2024). Can ChatGPT assist in picking stocks? *Finance Research Letters 59*, 104786.

Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance 36*(4), 992–1011.

Prieto, S. A., E. T. Mengiste, and B. García de Soto (2023). Investigating the Use of ChatGPT for the Scheduling of Construction Projects. *Buildings 13*(4).

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving Language Understandingby Generative Pre-Training. OpenAI Blog.

Russell, K. and J. Rennison (2024, January). These Seven Tech Stocks Are Driving the Market. *The New York Times*.

Russell, S. and P. Norvig (2020). *Artificial Intelligence: A Modern Approach* (Fourth ed.). Pearson.

Seagraves, C. and P. Seagraves (2024). Revolutionizing Real Estate Education: Integrating AI and LLMs into Undergraduate Curriculum. Working Paper.

Singh, H. (2023, December). It's the Magnificent Seven's Market. The Other Stocks Are Just Living in It. *Wall Street Journal*.

S&P Dow Jones Indices LLC (2024a). S&P CoreLogic Case-Shiller 10-City Composite Home Price Index [SPCS10RSA], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 19, 2025.

S&P Dow Jones Indices LLC (2024b). S&P CoreLogic Case-Shiller 20-City Composite Home Price Index [SPCS20RSA], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 19, 2025.

S&P Dow Jones Indices LLC (2024c). S&P CoreLogic Case-Shiller U.S. National Home Price Index [CSUSHPISA], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 19, 2025.

U.S. Bureau of Labor Statistics (2024). Unemployment Rate by State, Seasonally Adjusted [XXUR], retrieved from FRED, Federal Reserve Bank of St. Louis. Retrieved via the fredr R package on January 11, 2025.

U.S. Census Bureau (2024). Intercensal Subcounty Resident Population Estimates. Cohorts [IDs]: 2000–2009 [SUB-EST00INT-TOT], 2010–2019 [SUB-EST2020INT], and 2020–2023 [SUB-EST2023]. Retrieved via the Census website on January 11, 2025.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30*.

Viriato, J. C. (2019). AI and Machine Learning in Real Estate Investment. *The Journal of Portfolio Management 45*(7), 43–54.

Vlastakis, N. and R. N. Markellos (2012). Information demand and stock market volatility. *Journal of Banking & Finance 36*(6), 1808–1821.

Wan, W. X. and T. Lindenthal (2023). Testing machine learning systems in real estate. *Real Estate Economics 51*(3), 754–778.

Wu, L. and E. Brynjolfsson (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In A. Goldfarb, S. M. Greenstein, and C. E. Tucker (Eds.), *Economic Analysis of the Digital Economy*, pp. 89–118. University of Chicago Press.

Zillow Research (2024). ZHVI All Homes (Single-Family Residential, Condo, Co-op) Time Series, Smoothed, Seasonally Adjusted, City Geography Level and U.S. ZHVI. Retrieved via the Zillow website on January 11, 2025.

Figure 1: Zillow Home Value Index (ZHVI) – Annual City-Level Boxplots

The city-month-level ZHVIs were first averaged to the city-year level. Then each annual cross-section is summarized with a boxplot.

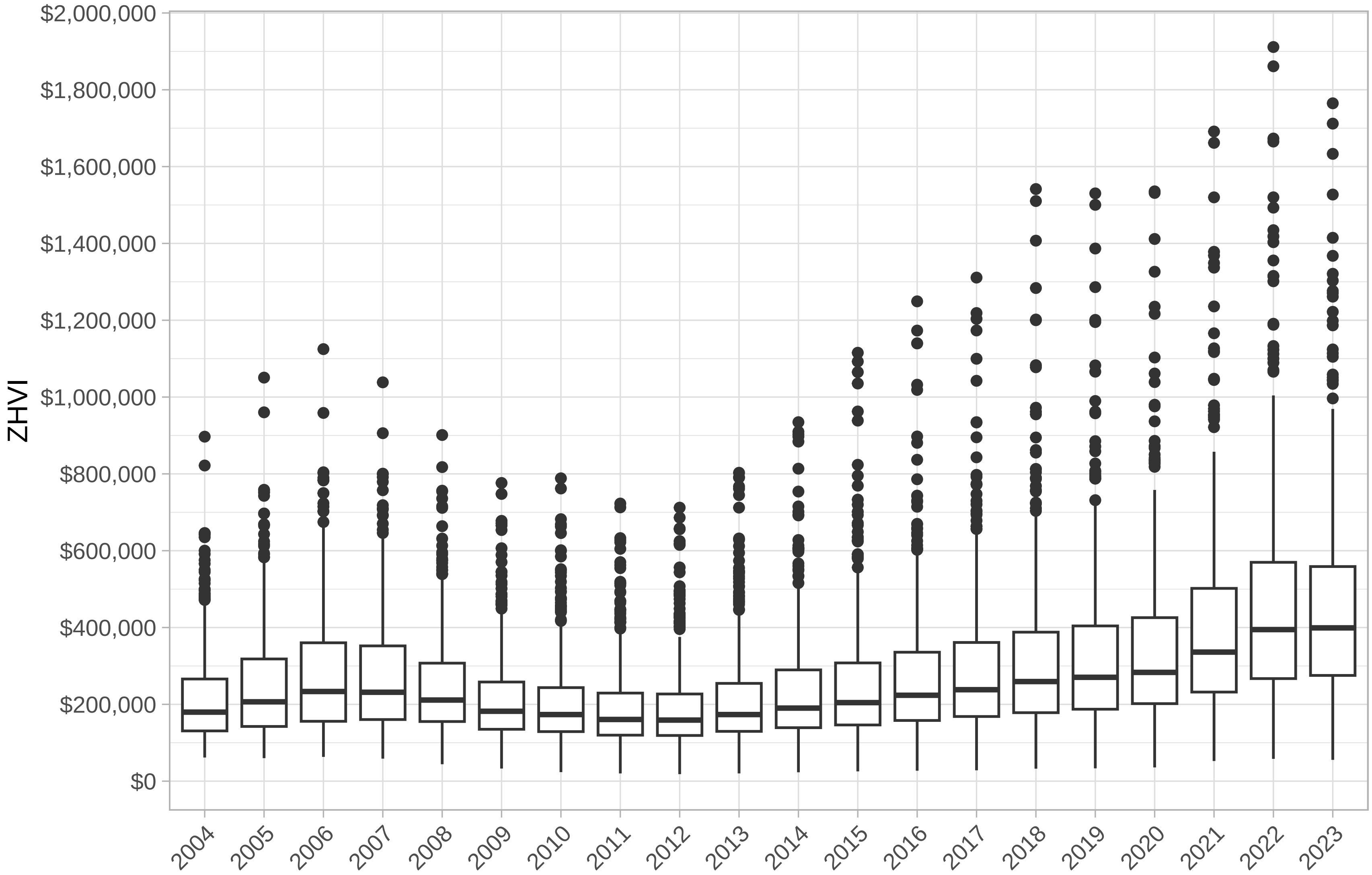


Figure 2: Zillow Home Value Index (ZHVI) – Benchmarks

Each curve represents a monthly time series of ZHVIs. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series.

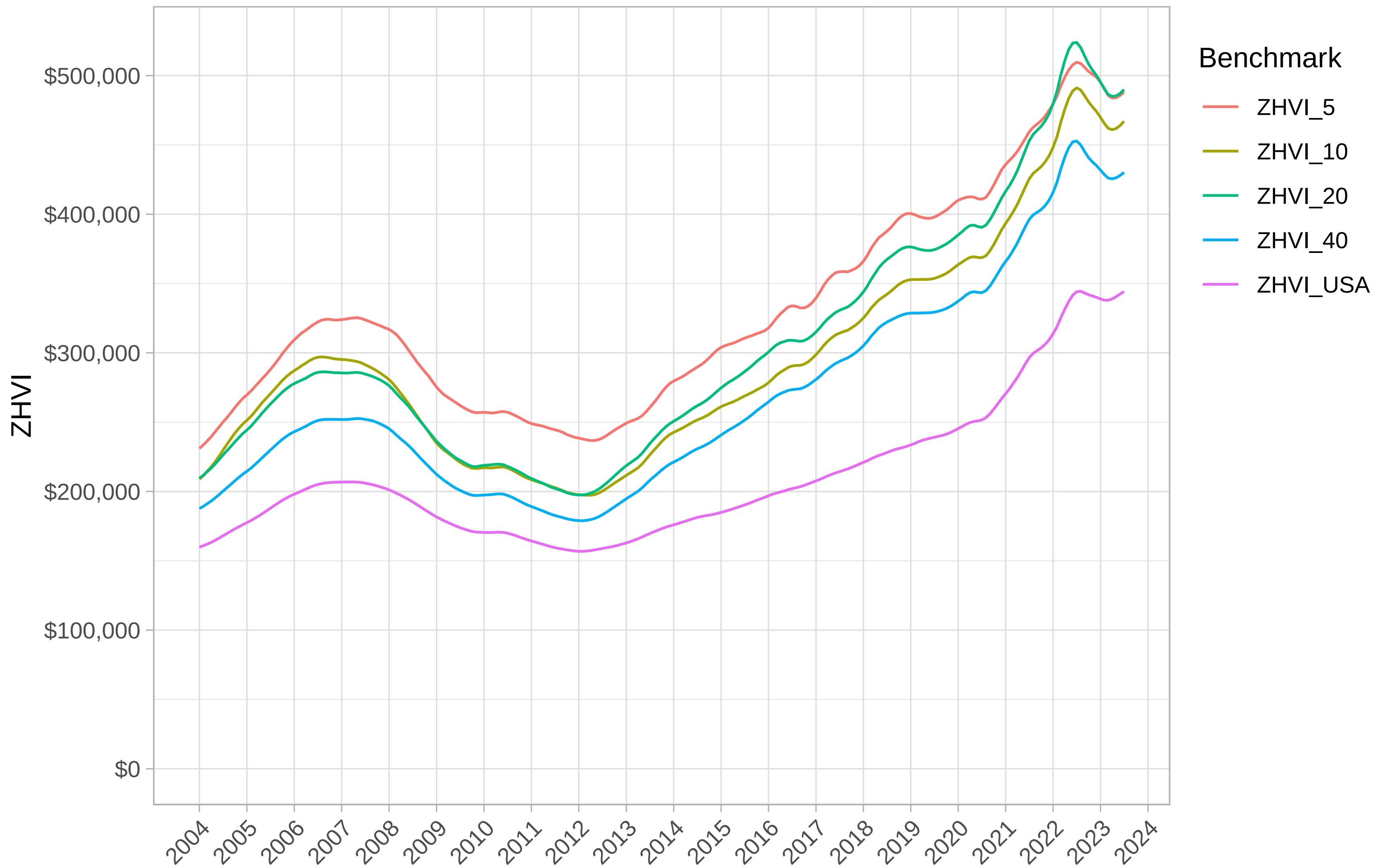


Figure 3: Census Populations – Annual City-Level Boxplots

The city-month-level populations were first averaged to the city-year level. Then each annual cross-section is summarized with a boxplot. The y-axis is presented in log-scale to better depict the distributions.

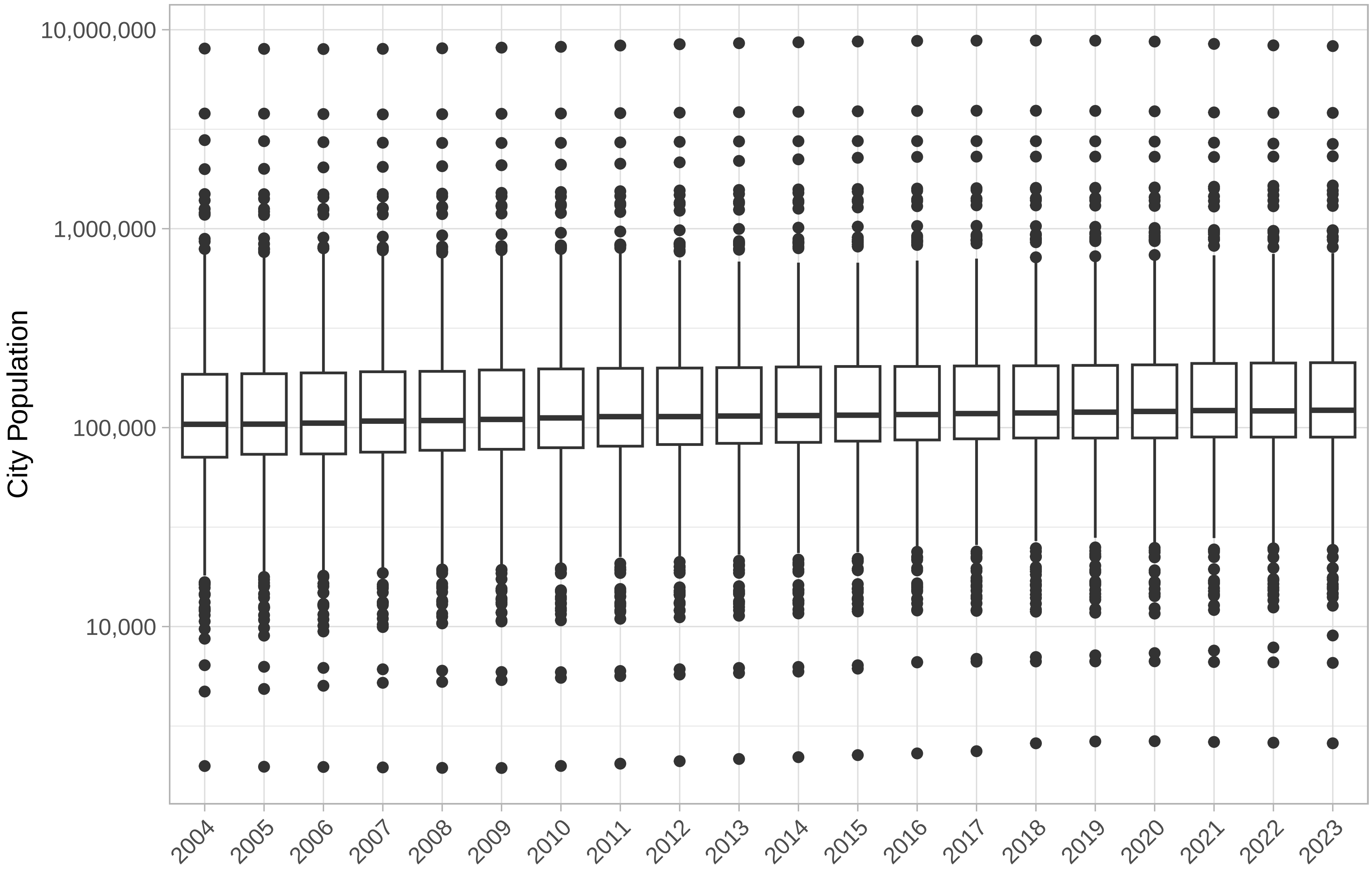


Figure 4: Unemployment Rates – Annual State-Level Boxplots

The state-month-level unemployment rates were first averaged to the state-year level. Then each annual cross-section is summarized with a boxplot.

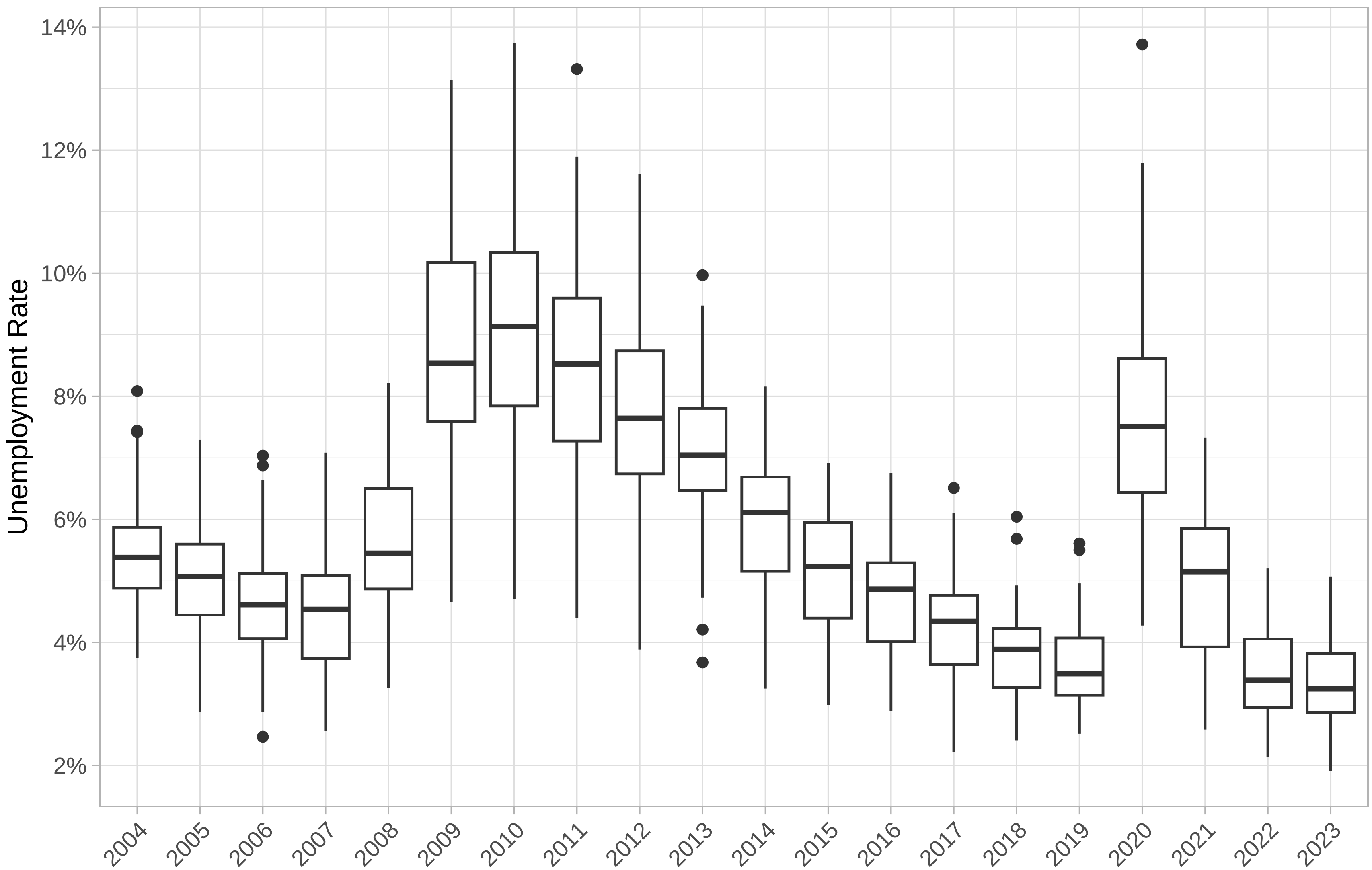


Figure 5: Freddie Mac 30-Year Fixed Mortgage Rates

This line chart depicts the monthly series of MORTGAGE30US in FRED, which is the 30-year fixed mortgage rate from Freddie Mac.

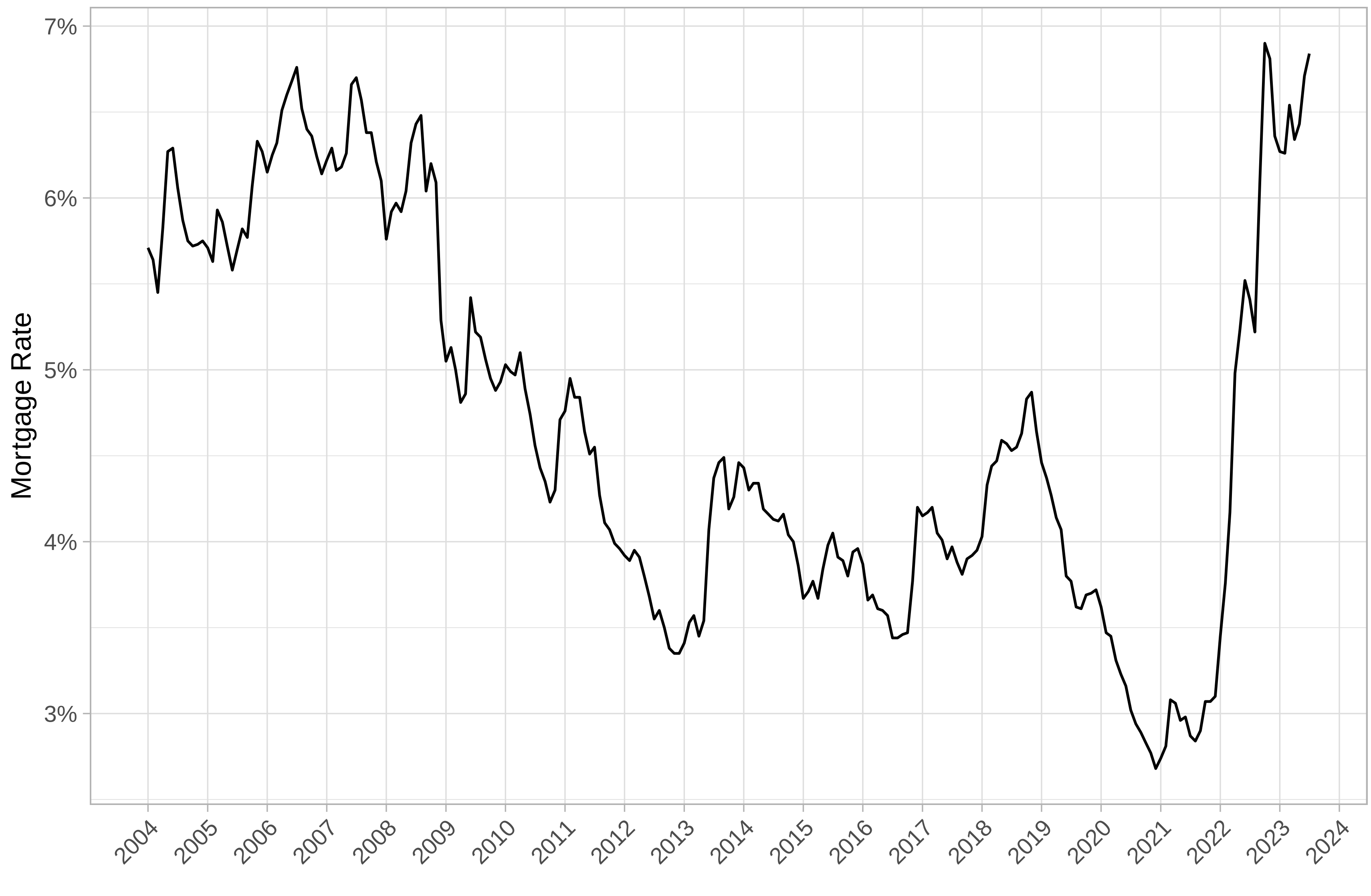


Figure 6: Google Search Volume Index (GSVI) – Annual City-Level Boxplots – “Real Estate” Category

The city-month-level GSVIs were first averaged to the city-year level. Then each annual cross-section is summarized with a boxplot.

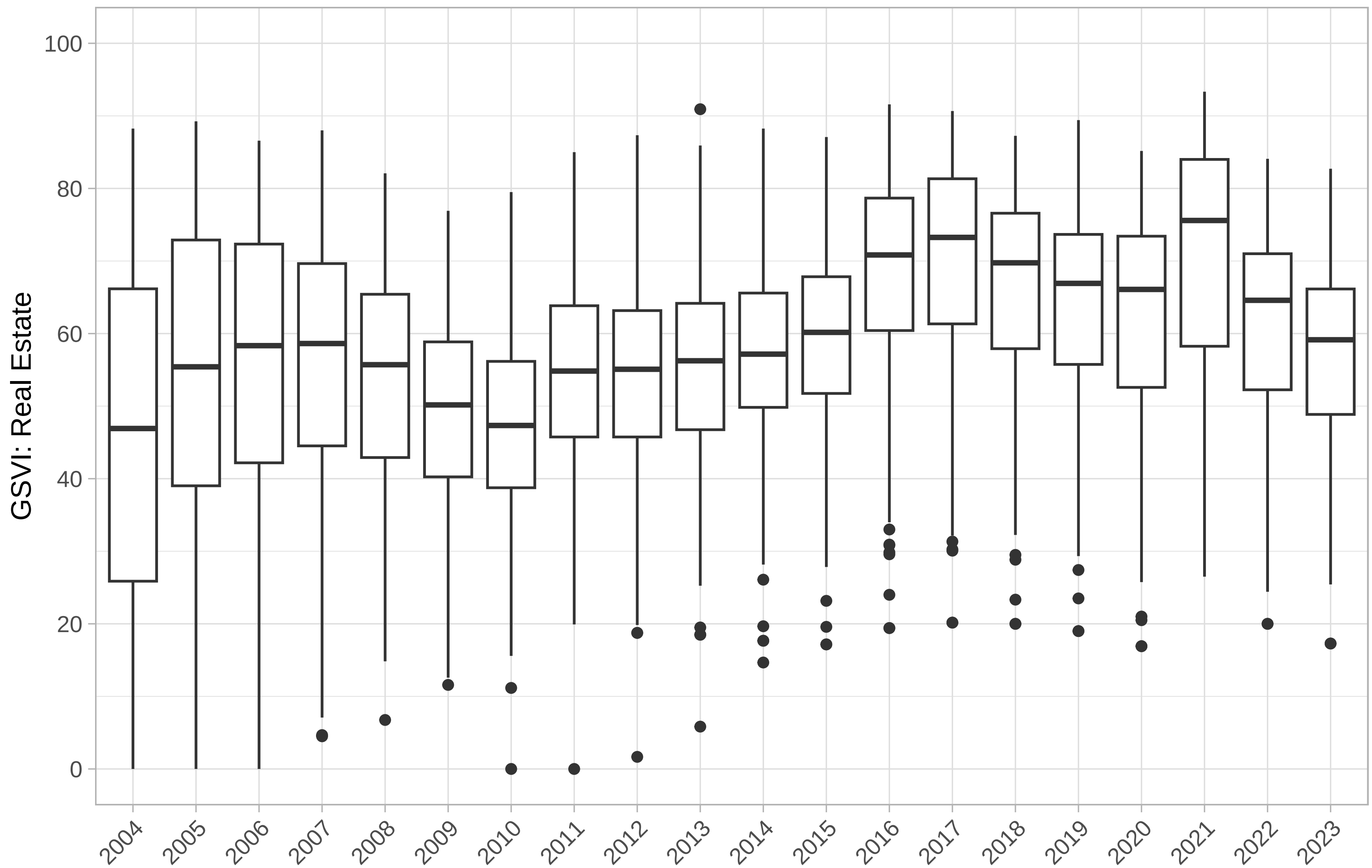


Figure 7: Google Search Volume Index (GSVI) – Annual City-Level Boxplots – “Real Estate Listings” Category

The city-month-level GSVIs were first averaged to the city-year level. Then each annual cross-section is summarized with a boxplot.

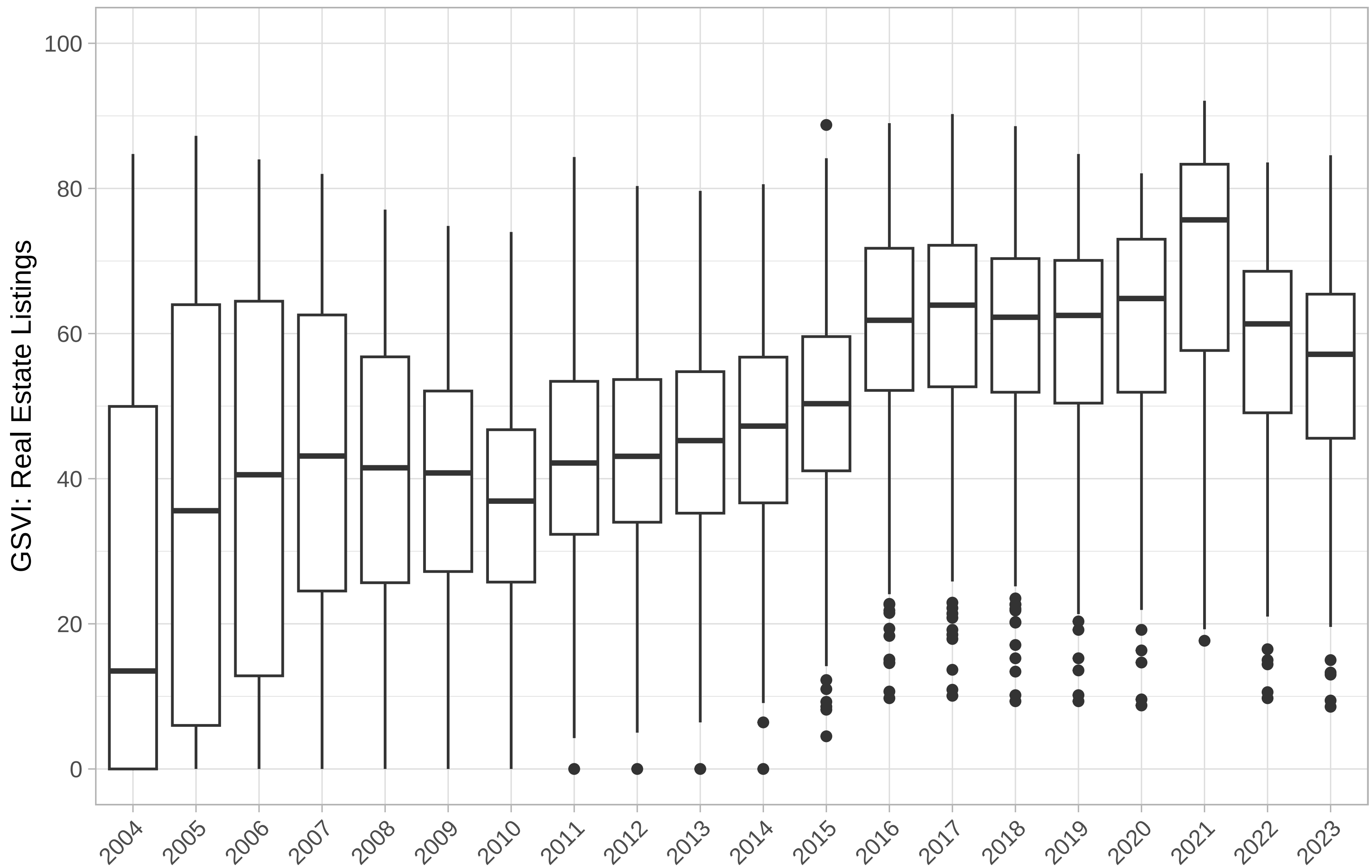


Figure 8: Scaled Values – All Benchmarks

Each of the benchmarks is scaled by its first observation in January 2004 to create a common starting point for the curves. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S.

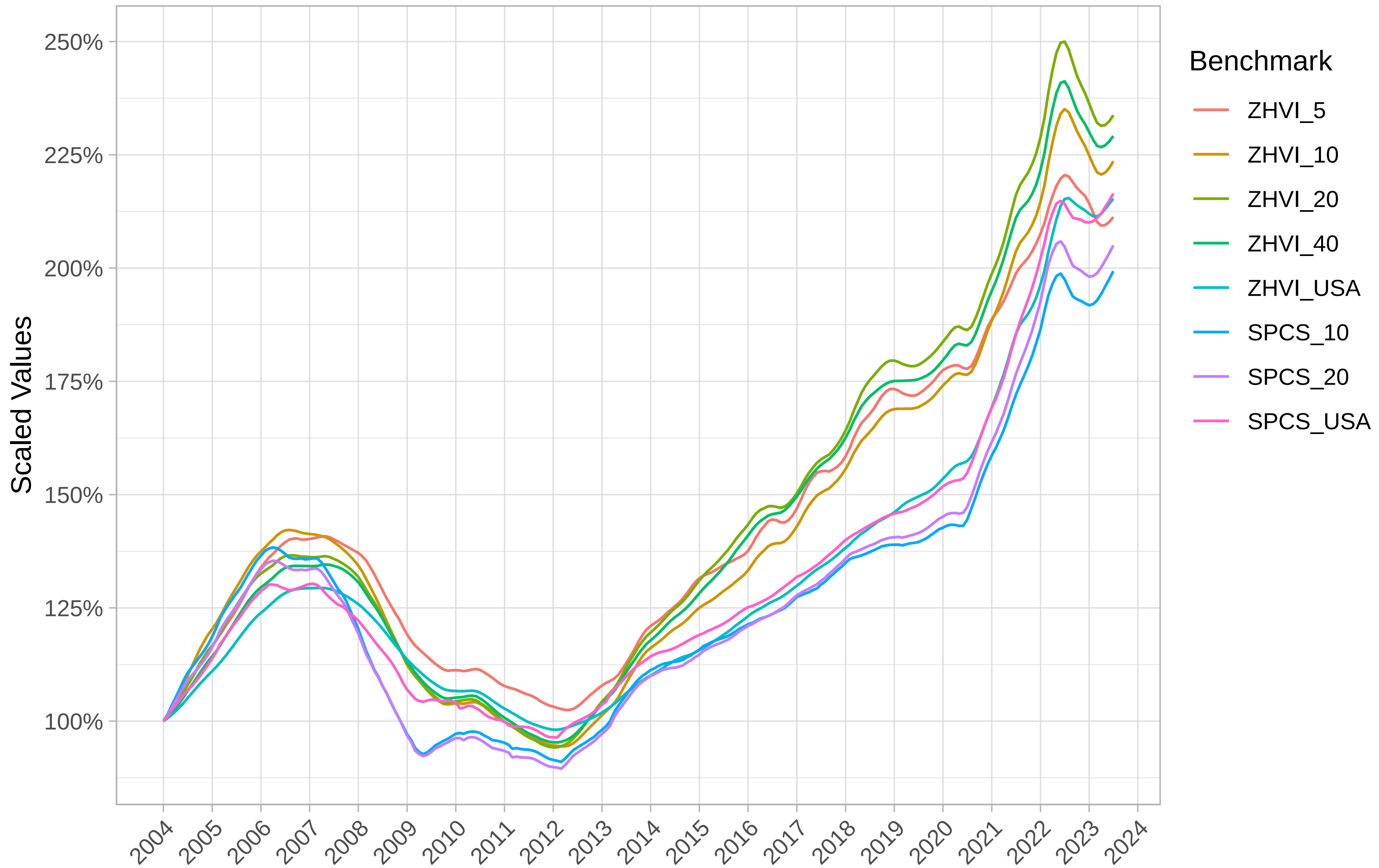


Figure 9: Unobfuscated vs. Obfuscated Training Data Comparisons – City Match Rates

This plot depicts the match rates between the city selections from the unobfuscated (GPT_N) and the obfuscated (OGPT_N) training data. The x-axis traces the five-year training windows, and the y-axis measures the percentage of cities that are selected by the AI in both versions of our training dataset. The low match rates suggest that the AI is selecting different cities when provided obfuscated data vs. unobfuscated data.

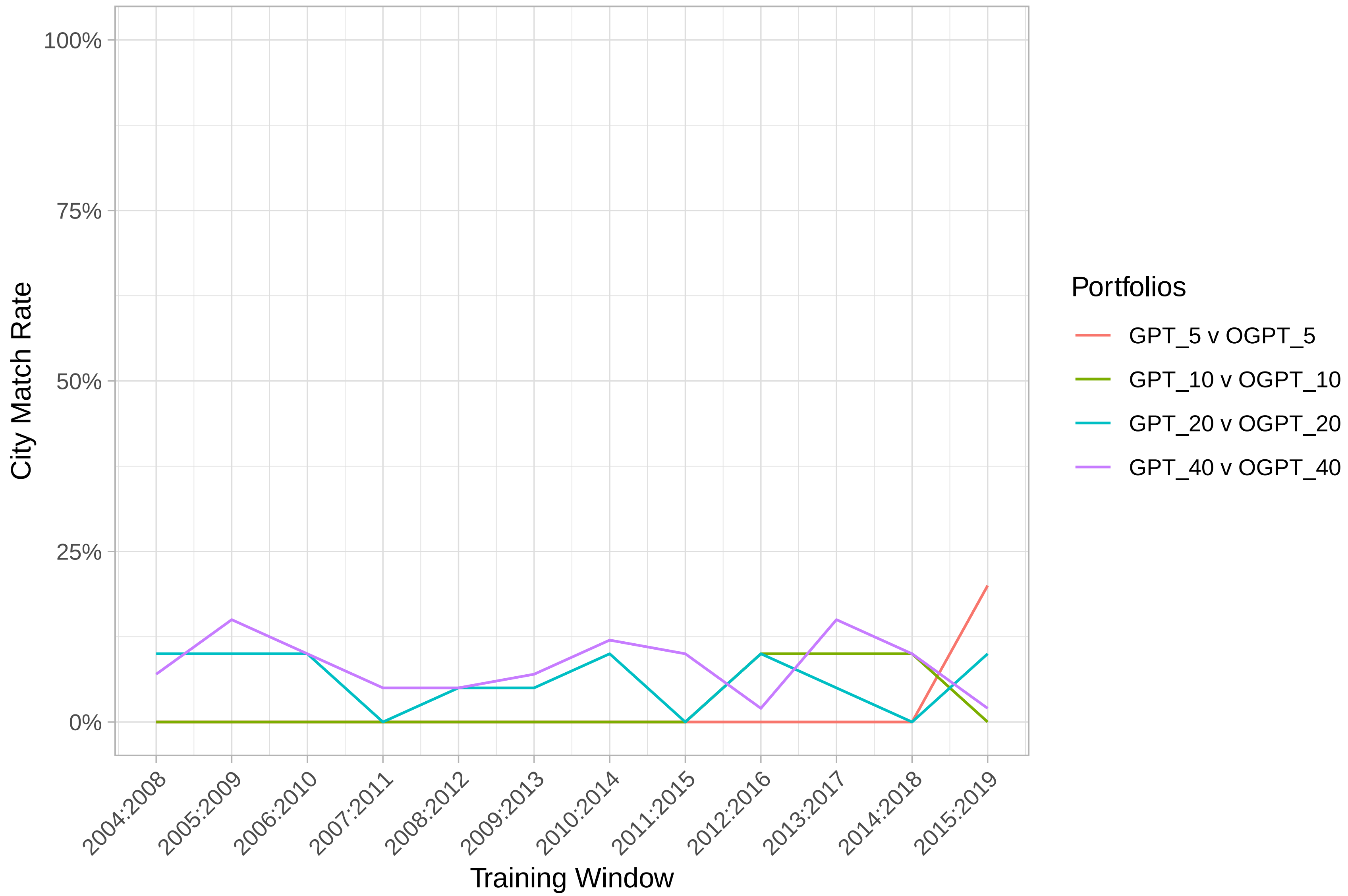


Figure 10: Backtesting – Horse Races

The x-axis traces the three-year testing windows, and the y-axis measures the cumulative returns of each portfolio. Portfolio colors and line types are the same as Figure 11 with the AI-generated portfolios as solid lines, ZHVI benchmarks as dashed lines, and SPCS benchmarks as dotted lines.

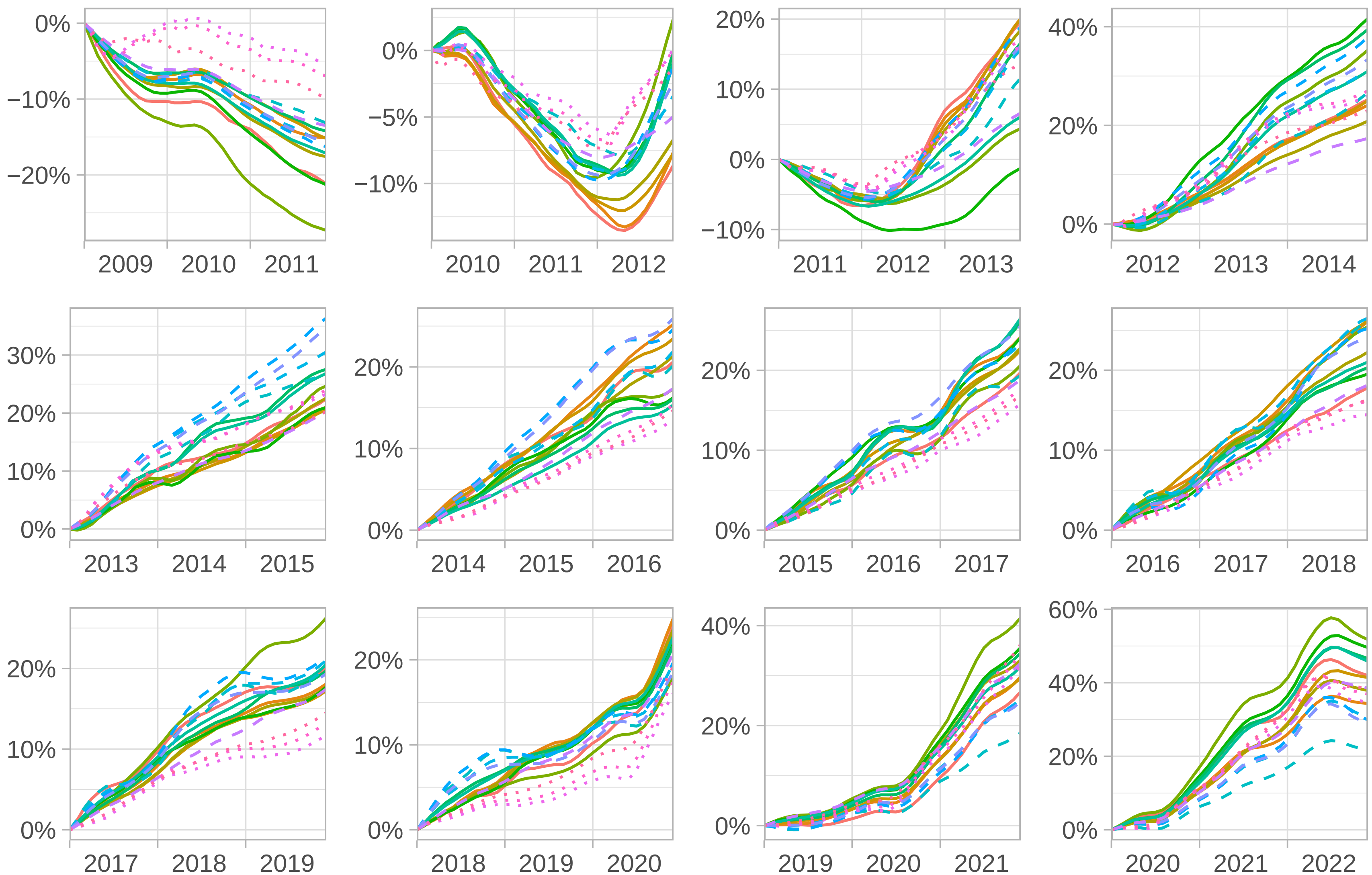


Figure 11: Backtesting – Total Returns

The x-axis traces the three-year testing windows, and the y-axis measures the cumulative returns of each portfolio-window pair. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S.

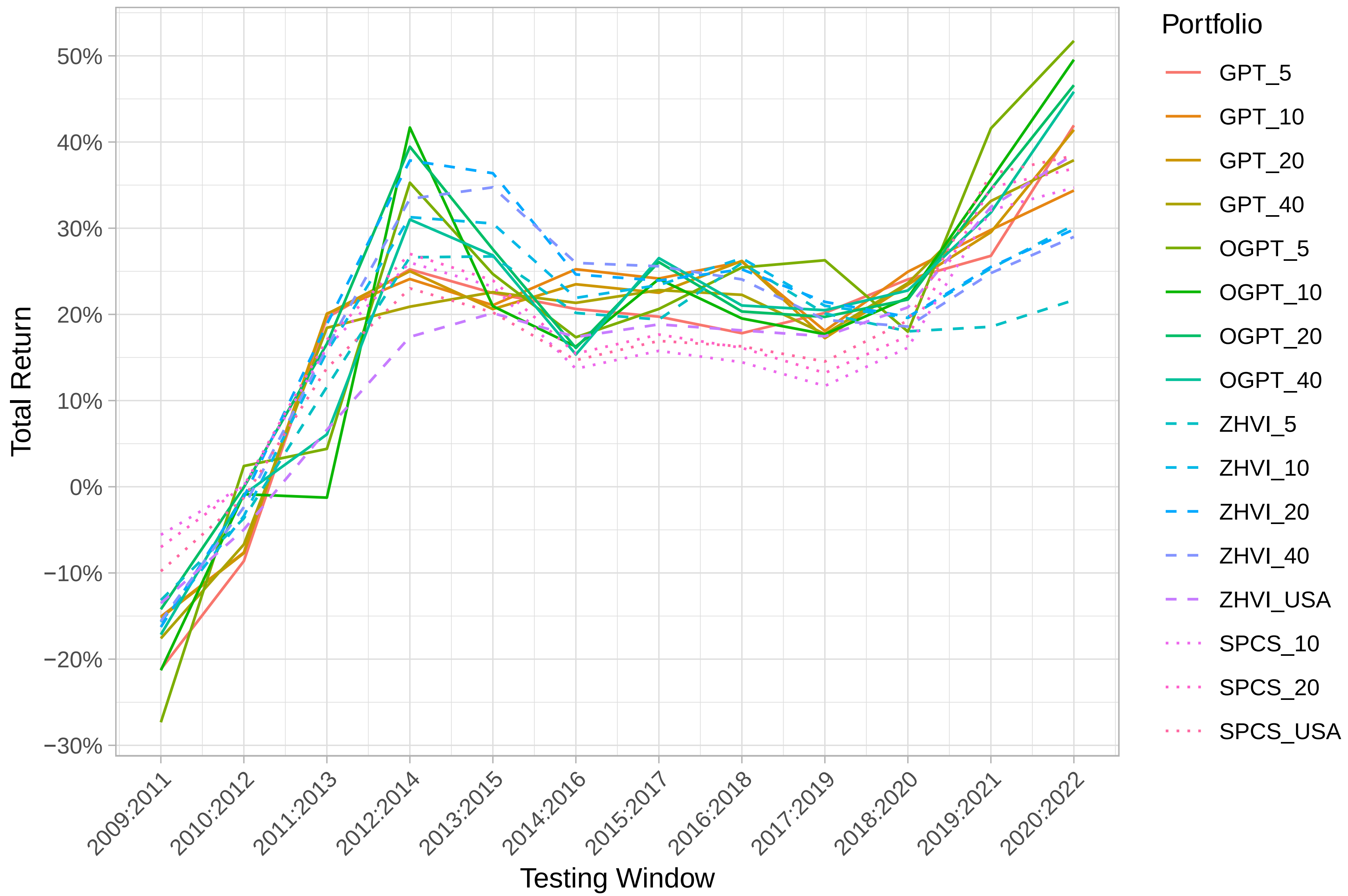


Figure 12: Unobfuscated vs. Obfuscated Training Data Comparisons – Total Return Differences

The x-axis traces the three-year testing windows, and the y-axis measures the difference in total returns between the unobfuscated and obfuscated AI-generated portfolios for each backtesting window. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The differences are calculated as GPT_N − OGPT_N. So positive values indicate better performance from the unobfuscated portfolio compared to the obfuscated portfolio, and negative values represent the opposite.

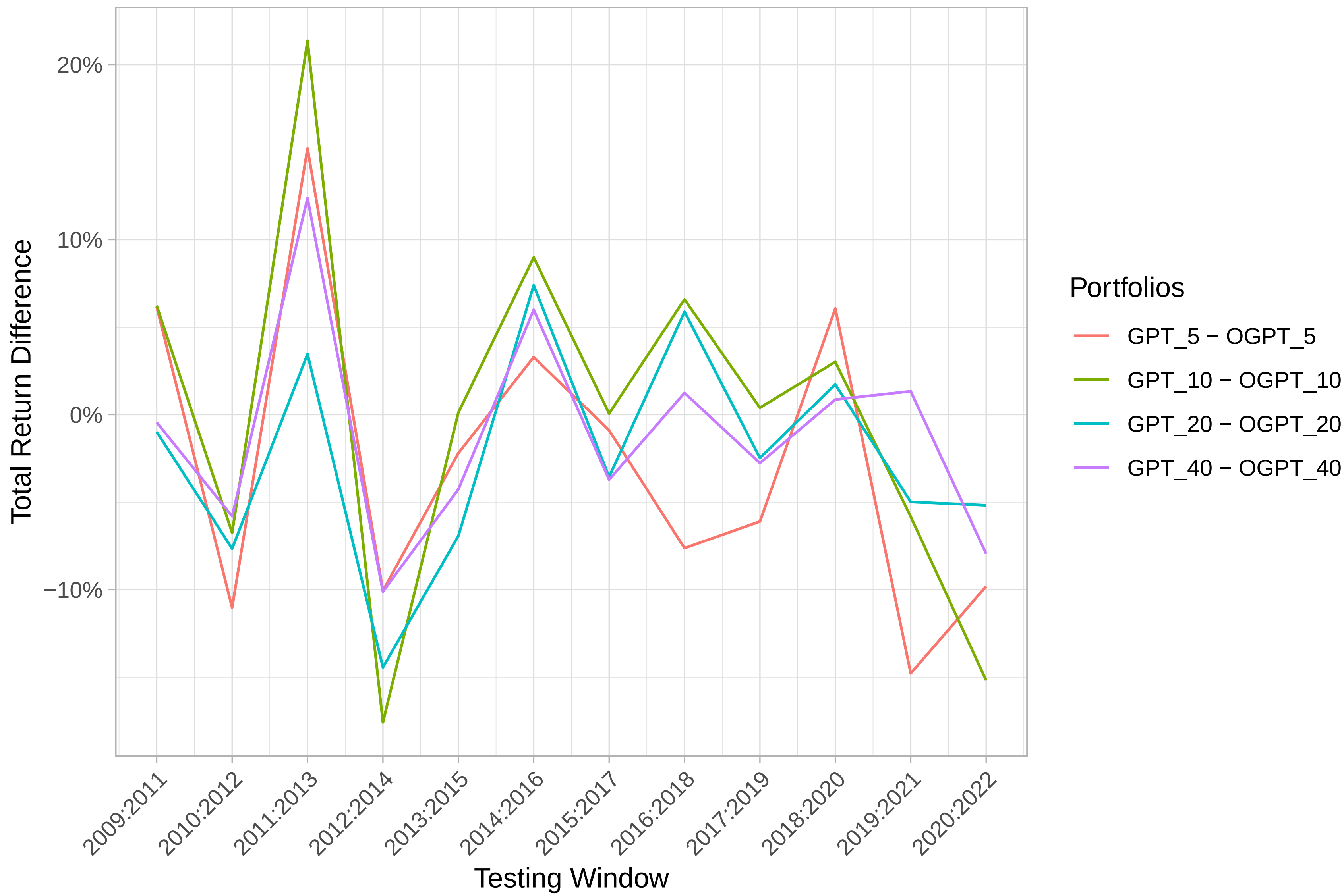


Figure 13: Backtesting – Sharpe Ratios

The x-axis traces the three-year testing windows, and the y-axis measures the Sharpe ratios of each portfolio-window pair. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S.

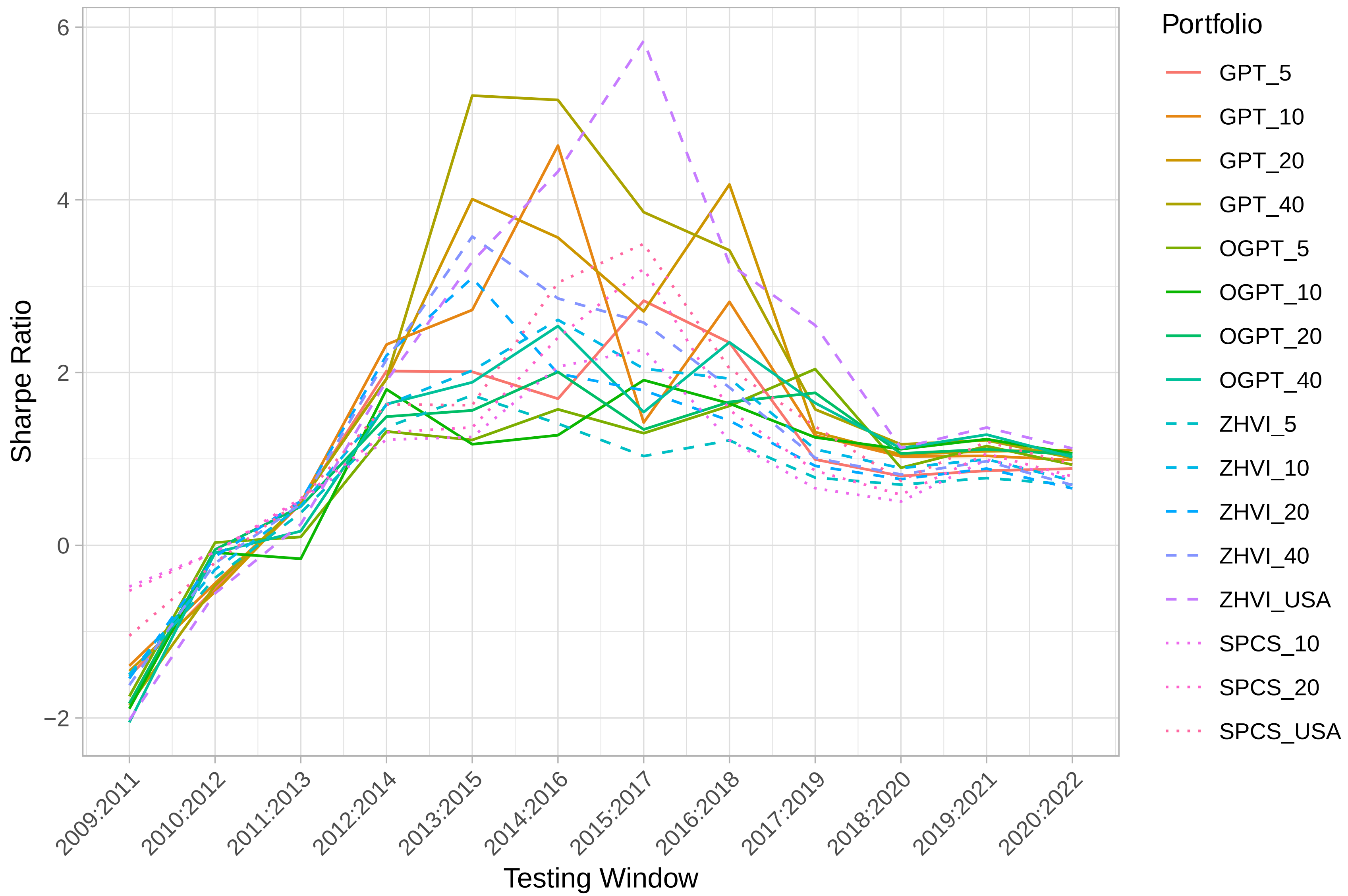


Figure 14: Unobfuscated vs. Obfuscated Training Data Comparisons – Sharpe Ratio Differences

The x-axis traces the three-year testing windows, and the y-axis measures the difference in Sharpe ratios between the unobfuscated and obfuscated AI-generated portfolios for each backtesting window. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The differences are calculated as GPT_N − OGPT_N. So positive values indicate better performance from the unobfuscated portfolio compared to the obfuscated portfolio, and negative values represent the opposite.

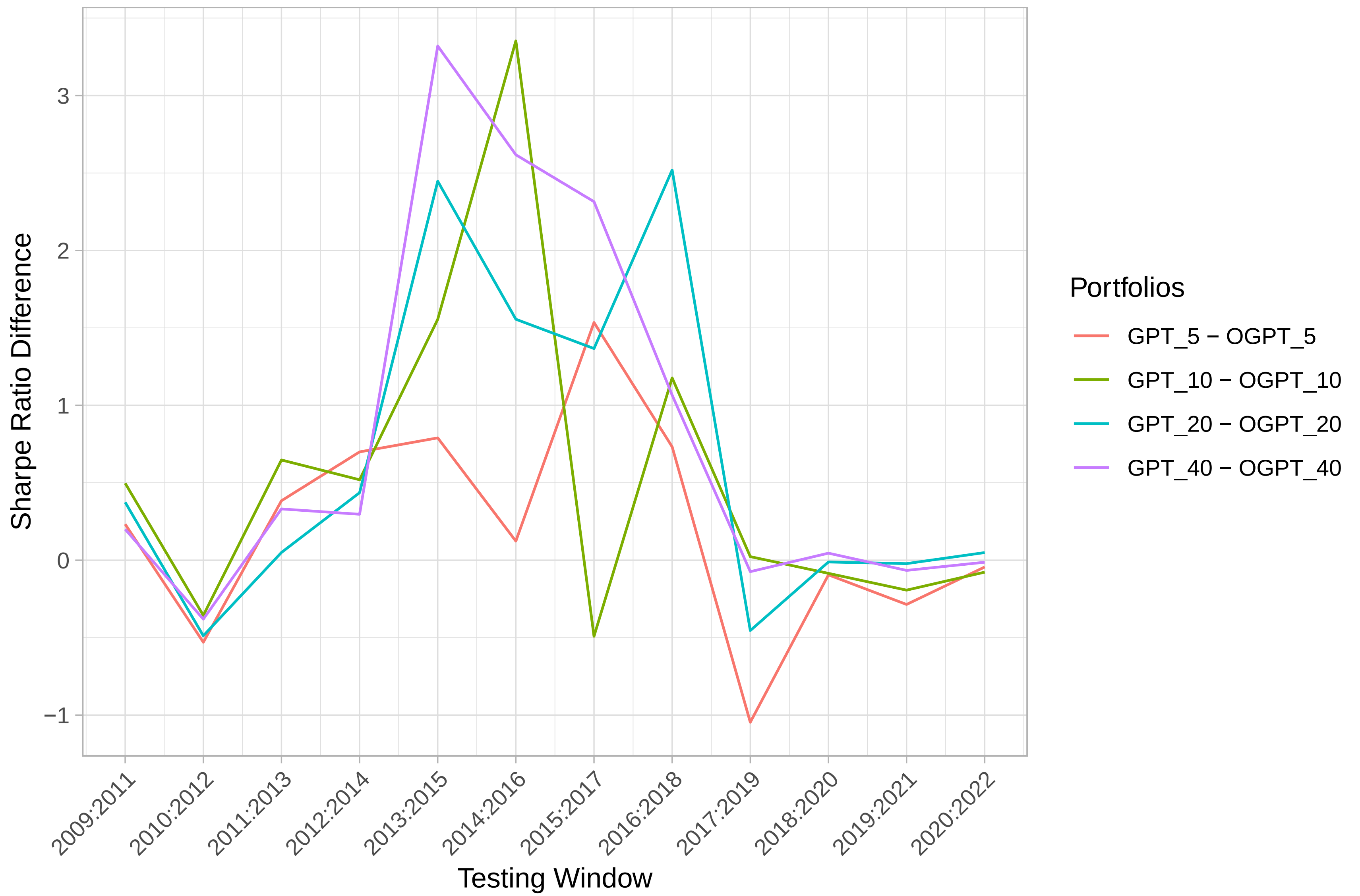


Table 1: Summary Statistics for the Final Training Dataset

ZHVI is the Zillow Home Value Index for 433 of the 500 largest U.S. cities from January 2004 – July 2023. POP is the residential population estimates from the Census. UnemployRate is the state-level unemployment rates from BLS, and MortgageRate is the nationwide 30-year fixed mortgage rate from Freddie Mac. The GSVI variables represent the city-specific search interest in the respective real estate categories. These values are provided by Google Trends as whole numbers between 0 and 100, where 100 represents the peak search volume for each city and 0 represents no meaningful search interest. See Section 3 for more details.

Statistic	Mean	St. Dev.	Min	Median	Max	N
ZHVI/1,000	282.04	197.77	17.85	221.14	1,998.93	101,521
POP/1,000	221.10	504.28	1.94	114.91	8,826.38	101,521
UnemployRate	6.08	2.54	1.80	5.40	30.60	101,521
MortgageRate	4.68	1.13	2.68	4.43	6.90	101,521
GSVI:RealEstate	57.93	19.17	0	59	100	101,521
GSVI:RealEstateListings	48.40	23.71	0	50	100	101,521

*Table generated with the stargazer R package (Hlavac, 2022).

Table 2: Unobfuscated vs. Obfuscated Training Data Comparisons – City Match Counts

This table presents the number of shared cities between the AI-generated portfolios for the unobfuscated and obfuscated training data. Each column represents a training window with the 04:08 column corresponding with training data from 2004–2008. The Mean column represents the average number of city matches across all windows. The low match counts suggest that the AI is selecting different cities when provided obfuscated data vs. unobfuscated data. See Figure 9 for a visual representation of these results in the form of city match rates.

Portfolios	04:08	05:09	06:10	07:11	08:12	09:13	10:14	11:15	12:16	13:17	14:18	15:19	Mean
GPT_5 v OGPT_5	0	0	0	0	0	0	0	0	0	0	0	1	0.08
GPT_10 v OGPT_10	0	0	0	0	0	0	0	0	1	1	1	0	0.25
GPT_20 v OGPT_20	2	2	2	0	1	1	2	0	2	1	0	2	1.25
GPT_40 v OGPT_40	3	6	4	2	2	3	5	4	1	6	4	1	3.42

*Table generated with the stargazer R package (Hlavac, 2022).

Table 3: Unobfuscated vs. Obfuscated Training Data Comparisons – Top 10 Most Selected Cities

This table contains two sub-tables that each contain the 10 most frequently selected cities from the AI-generated portfolios. Panel (a) shows the top 10 cities generated from the unobfuscated training data, and Panel (b) does the same for the obfuscated data. CityID's were randomly assigned. RegionName and State show the city names and states. The Count column represents the total number of times each city was selected across all testing windows. Thus, each city has 48 potential opportunities to be selected (4 portfolios per window × 12 testing windows).

(a) Unobfuscated Training Data – Top 10 Cities

CityID	RegionName	State	Count
15	Edmond	OK	9
353	New York	NY	7
399	Los Angeles	CA	7
341	Chesapeake	VA	7
9	Overland Park	KS	7
432	Fredericksburg	VA	7
22	Erie	PA	7
241	Sunnyvale	CA	7
39	Lake Charles	LA	7
47	Indianapolis	IN	6

*Table generated with the stargazer R package (Hlavac, 2022).

(b) Obfuscated Training Data – Top 10 Cities

CityID	RegionName	State	Count
126	San Francisco	CA	15
353	New York	NY	12
221	San Diego	CA	11
399	Los Angeles	CA	10
163	San Jose	CA	10
88	Pasadena	CA	9
357	Detroit	MI	8
65	Warren	MI	8
105	Berkeley	CA	8
296	Chicago	IL	7

*Table generated with the stargazer R package (Hlavac, 2022).

Table 4: Backtesting Comparisons – Total Returns (in %)

This table summarizes the total (cumulative) returns for the AI-generated portfolios and benchmarks across each of the back-testing windows. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S. Each column represents a testing window with the 09:11 column corresponding with training data from 2004–2008, and then evaluating the performance over the 2009–2011 period. In addition to presenting the cumulative returns for each window, the Mean column provides the averages across all windows. The portfolios (rows) have also been sorted by the Mean column. See Figure 11 for a visual representation of these results.

Portfolio	09:11	10:12	11:13	12:14	13:15	14:16	15:17	16:18	17:19	18:20	19:21	20:22	Mean
OGPT_20	−14.22	−0.05	16.58	39.46	27.55	16.11	26.07	20.33	19.72	21.73	34.53	46.60	21.20
ZHVI_20	−16.28	−1.19	18.86	37.86	36.38	24.63	23.93	25.23	21.46	19.71	25.51	29.86	20.50
OGPT_5	−27.32	2.41	4.40	35.28	24.68	17.35	20.63	25.44	26.28	18.00	41.58	51.74	20.04
ZHVI_40	−15.59	−2.37	16.16	33.40	34.75	25.98	25.59	24.06	19.36	18.59	24.79	29.00	19.48
OGPT_40	−17.15	−0.88	6.07	31.01	26.85	15.35	26.53	21.03	20.49	22.77	31.82	45.84	19.14
GPT_20	−15.20	−7.70	20.04	25.02	20.61	23.50	22.51	26.21	17.24	23.44	29.54	41.42	18.89
ZHVI_10	−15.69	−3.40	15.66	31.28	30.55	21.92	23.40	26.52	21.00	19.59	25.33	30.37	18.88
GPT_10	−15.07	−7.62	20.09	24.11	21.07	25.24	24.16	26.11	18.09	24.95	29.80	34.37	18.78
OGPT_10	−21.30	−0.87	−1.26	41.68	20.96	16.26	24.10	19.53	17.70	21.93	35.64	49.55	18.66
GPT_40	−17.60	−6.69	18.44	20.90	22.58	21.34	22.82	22.27	17.72	23.64	33.15	37.89	18.04
SPCS_20	−7.00	0.12	17.06	27.02	23.95	15.45	17.66	16.12	13.20	17.49	34.75	36.92	17.73
GPT_5	−21.17	−8.62	19.61	25.23	22.48	20.63	19.74	17.82	20.18	24.06	26.80	41.93	17.39
SPCS_USA	−9.79	−1.29	13.76	23.02	20.23	14.69	16.92	16.29	14.53	19.28	36.29	38.43	16.86
SPCS_10	−5.58	0.14	16.14	26.01	23.23	13.70	15.77	14.45	11.70	16.14	32.09	34.71	16.54
ZHVI_5	−13.16	−3.66	11.59	26.63	26.73	20.17	19.40	25.97	20.11	18.03	18.56	21.66	16.00
ZHVI_USA	−13.51	−4.97	6.61	17.39	20.15	17.37	18.84	18.14	17.46	20.82	32.47	38.58	15.78

*Table generated with the stargazer R package (Hlavac, 2022).

Table 5: Backtesting Comparisons – Total Return Ranks

This table presents the total return ranks for each backtesting window, as well as the mean ranks. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S. Each column represents a testing window with the 09:11 column corresponding with training data from 2004–2008, and then evaluating the performance over the 2009–2011 period. The portfolios (rows) have also been sorted by the Mean column.

Portfolio	09:11	10:12	11:13	12:14	13:15	14:16	15:17	16:18	17:19	18:20	19:21	20:22	Mean
OGPT_20	6	4	7	2	4	12	2	10	7	7	5	3	5.75
ZHVI_20	11	7	4	3	1	3	6	6	2	9	13	14	6.58
OGPT_5	16	1	15	4	7	10	10	5	1	14	1	1	7.08
GPT_10	7	14	1	13	12	2	4	3	9	1	10	12	7.33
OGPT_40	12	6	14	7	5	14	1	9	4	5	9	4	7.50
ZHVI_10	10	10	10	6	3	5	7	1	3	10	14	13	7.67
ZHVI_40	9	9	8	5	2	1	3	7	8	12	15	15	7.83
OGPT_10	15	5	16	1	13	11	5	11	11	6	3	2	8.25
GPT_20	8	15	2	12	14	4	9	2	13	4	11	6	8.33
GPT_40	13	13	5	15	10	6	8	8	10	3	6	9	8.83
GPT_5	14	16	3	11	11	7	11	13	5	2	12	5	9.17
SPCS_20	2	3	6	8	8	13	14	15	15	15	4	10	9.42
ZHVI_5	4	11	12	9	6	8	12	4	6	13	16	16	9.75
SPCS_10	1	2	9	10	9	16	16	16	16	16	8	11	10.83
SPCS_USA	3	8	11	14	15	15	15	14	14	11	2	8	10.83
ZHVI_USA	5	12	13	16	16	9	13	12	12	8	7	7	10.83

*Table generated with the stargazer R package (Hlavac, 2022).

Table 6: Backtesting Comparisons – Sharpe Ratios

This table summarizes the Sharpe ratios for the AI-generated portfolios and benchmarks across each of the backtesting windows. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S. Each column represents a testing window with the 09:11 column corresponding with training data from 2004–2008, and then evaluating the performance over the 2009–2011 period. In addition to presenting the Sharpe ratios for each window, the Mean column provides the averages across all windows. The portfolios (rows) have also been sorted by the Mean column. See Figure 13 for a visual representation of these results.

Portfolio	09:11	10:12	11:13	12:14	13:15	14:16	15:17	16:18	17:19	18:20	19:21	20:22	Mean
GPT_40	−1.85	−0.47	0.49	1.93	5.21	5.16	3.86	3.42	1.57	1.17	1.22	1.01	1.89
ZHVI_USA	−2.02	−0.56	0.24	1.90	3.29	4.33	5.84	3.26	2.55	1.13	1.36	1.12	1.87
GPT_20	−1.46	−0.54	0.50	1.93	4.01	3.56	2.71	4.18	1.31	1.05	1.09	1.10	1.62
GPT_10	−1.40	−0.44	0.49	2.32	2.72	4.63	1.42	2.82	1.27	1.03	1.04	0.99	1.41
SPCS_USA	−1.05	−0.20	0.52	1.63	1.62	3.04	3.49	2.05	1.38	0.75	1.20	0.94	1.28
ZHVI_40	−1.62	−0.21	0.48	2.15	3.57	2.86	2.58	1.83	1.01	0.82	0.97	0.70	1.26
SPCS_20	−0.53	−0.08	0.54	1.31	1.36	2.40	3.20	1.57	0.86	0.58	1.06	0.80	1.09
OGPT_40	−2.05	−0.09	0.16	1.63	1.89	2.54	1.54	2.35	1.65	1.12	1.28	1.02	1.09
GPT_5	−1.52	−0.50	0.48	2.02	2.01	1.70	2.83	2.35	0.99	0.80	0.86	0.89	1.08
ZHVI_10	−1.50	−0.28	0.45	1.63	2.02	2.61	2.05	1.93	1.11	0.89	1.00	0.75	1.06
ZHVI_20	−1.54	−0.13	0.51	2.20	3.10	1.99	1.79	1.44	0.92	0.76	0.89	0.66	1.05
OGPT_20	−1.83	−0.05	0.45	1.49	1.56	2.01	1.34	1.66	1.76	1.06	1.11	1.05	0.97
SPCS_10	−0.48	−0.08	0.52	1.22	1.25	2.07	2.26	1.22	0.66	0.51	0.99	0.77	0.91
OGPT_5	−1.75	0.03	0.10	1.32	1.22	1.57	1.30	1.61	2.04	0.90	1.15	0.93	0.87
OGPT_10	−1.89	−0.08	−0.16	1.81	1.17	1.28	1.91	1.64	1.25	1.11	1.23	1.06	0.86
ZHVI_5	−1.51	−0.38	0.38	1.37	1.74	1.41	1.03	1.21	0.78	0.70	0.78	0.70	0.68

*Table generated with the stargazer R package (Hlavac, 2022).

Table 7: Backtesting Comparisons – Sharpe Ratio Ranks

This table presents the portfolio ranks for each backtesting window, as well as the mean ranks. The GPT_N portfolios are selected by AI from unobfuscated training data. The OGPT_N portfolios are selected by AI from obfuscated training data. The ZHVI_N benchmarks are the average ZHVI across the N largest cities, as per Zillow's SizeRank variable. ZHVI_USA is the national ZHVI series. The SPCS benchmarks are the S&P CoreLogic Case-Shiller Composite Home Price Indices for 10-cities, 20-cities, and the entire U.S. Each column represents a testing window with the 09:11 column corresponding with training data from 2004–2008, and then evaluating the performance over the 2009–2011 period. The portfolios (rows) have also been sorted by the Mean column.

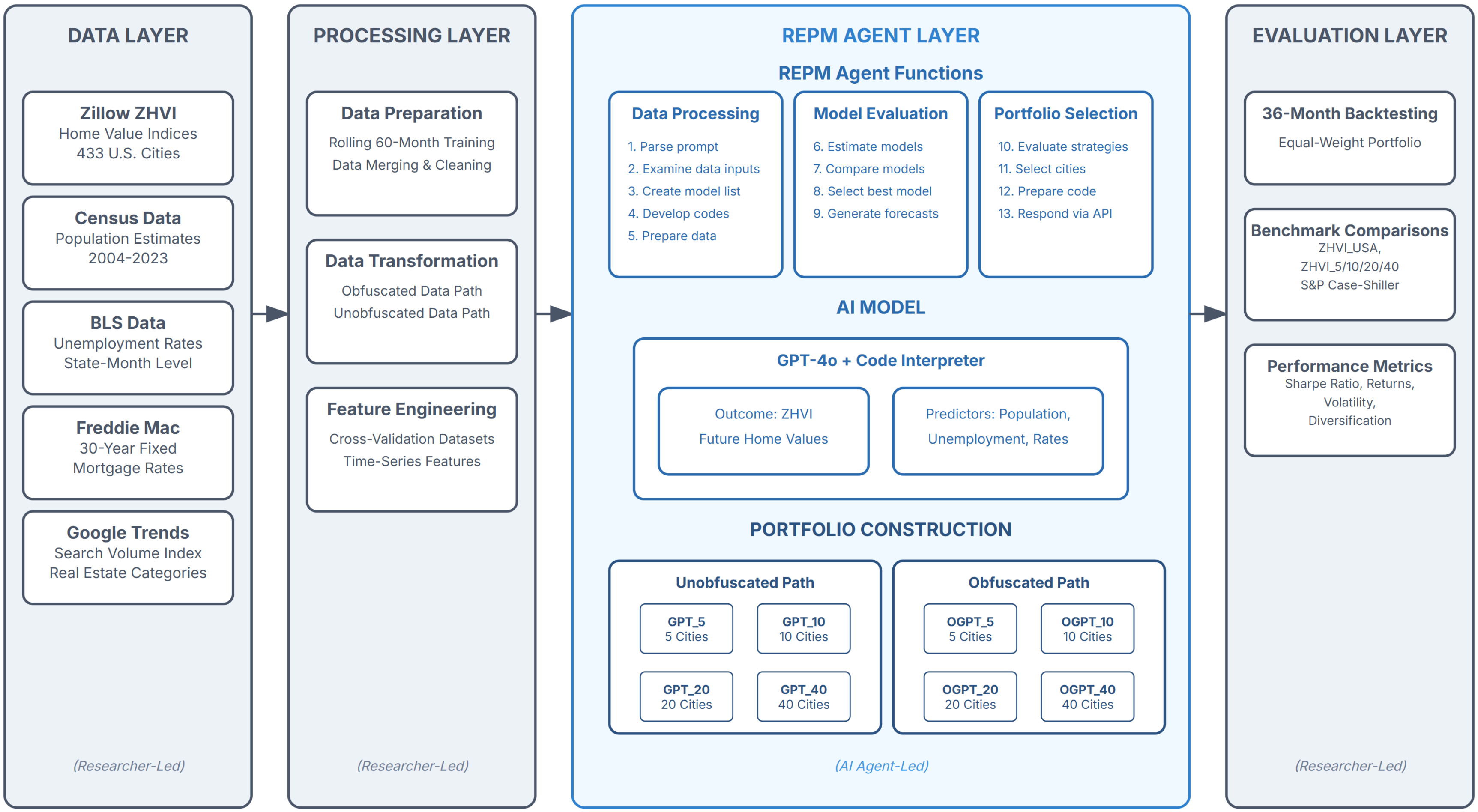
Portfolio	09:11	10:12	11:13	12:14	13:15	14:16	15:17	16:18	17:19	18:20	19:21	20:22	Mean
GPT_40	13	13	6	5	1	1	2	2	5	1	4	6	4.92
GPT_20	5	15	5	6	2	4	6	1	7	6	8	2	5.58
ZHVI_USA	15	16	13	7	4	3	1	3	1	2	1	1	5.58
GPT_10	4	12	7	1	6	2	13	4	8	7	10	7	6.75
SPCS_USA	3	8	2	10	11	5	3	7	6	13	5	8	6.75
OGPT_40	16	6	14	9	9	8	12	5	4	3	2	5	7.75
OGPT_20	12	2	11	12	12	11	14	10	3	5	7	4	8.58
ZHVI_40	10	9	9	3	3	6	7	9	11	10	13	15	8.75
SPCS_20	2	3	1	15	13	9	4	13	14	15	9	11	9.08
ZHVI_10	6	10	10	11	7	7	9	8	10	9	11	13	9.25
GPT_5	8	14	8	4	8	13	5	6	12	11	15	10	9.50
OGPT_10	14	5	16	8	16	16	10	11	9	4	3	3	9.58
ZHVI_20	9	7	4	2	5	12	11	14	13	12	14	16	9.92
OGPT_5	11	1	15	14	15	14	15	12	2	8	6	9	10.17
SPCS_10	1	4	3	16	14	10	8	15	16	16	12	12	10.58
ZHVI_5	7	11	12	13	10	15	16	16	15	14	16	14	13.25

*Table generated with the stargazer R package (Hlavac, 2022).

A Real Estate Portfolio Management (REPM) Agent Architecture

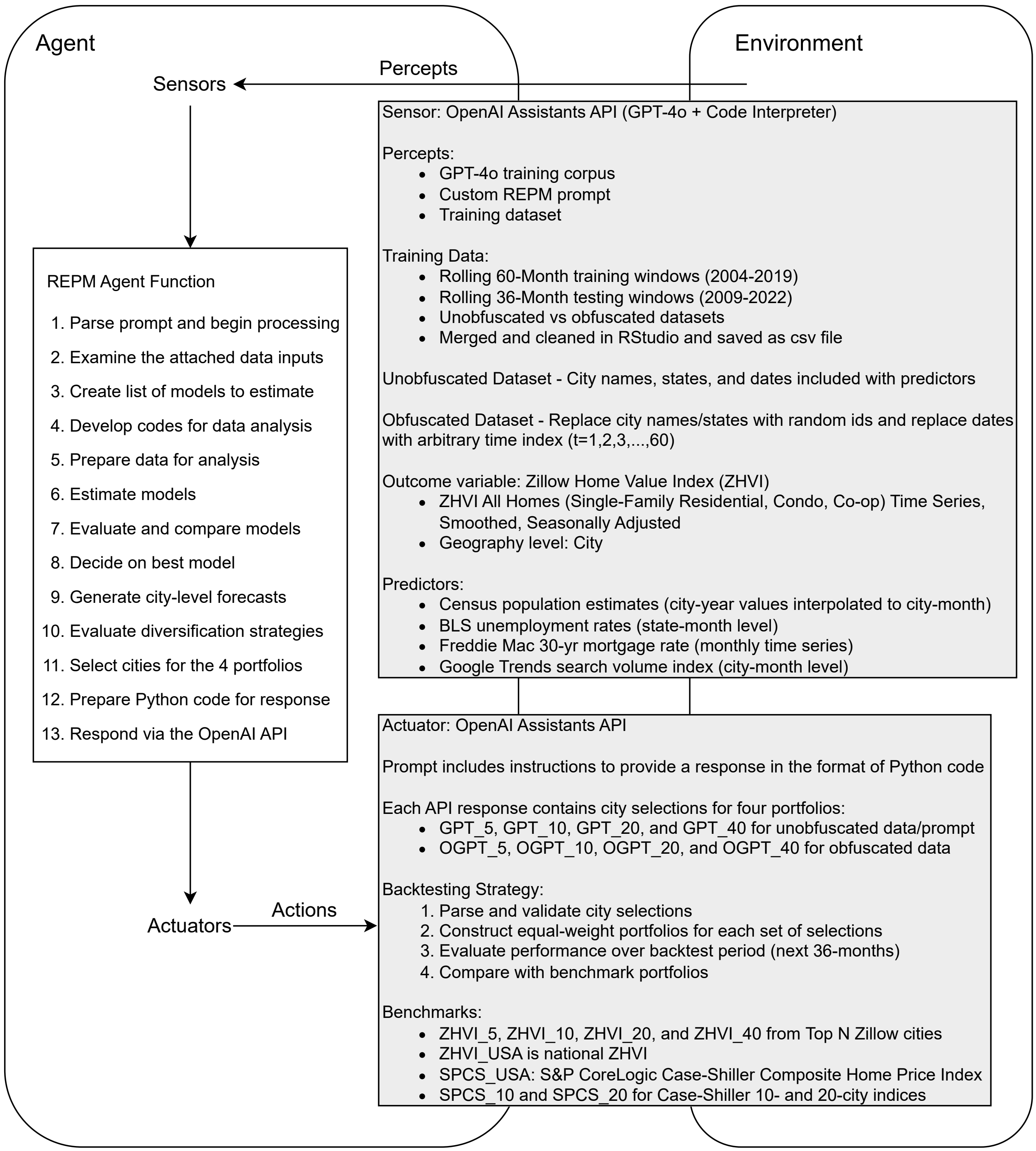
Figures 15–16 below summarize the experimental design of this research and the agent function for the OpenAI assistant that we used as the subject of the experiment. The REPM Agent diagram in Figure 15 provides a simplified overview of the data flows and portfolio construction process with an emphasis on researcher-led and AI agent-led tasks.

Figure 15: REPM Agent Portfolio Construction Workflow



Alternatively, Figure 16 below is inspired by the generic agent diagram in Figure 2.1 of Russell and Norvig (2020) and provides a detailed description of the inputs (percepts) and outputs (actions) of our REPM agent, as well as the agent function for its responses.

Figure 16: Detailed REPM Agent Structure Diagram



B Full Prompt and Example Response

The following prompt was provided to GPT-4o through the OpenAI API with a goal of generating investment recommendations based on the associated training dataset. See Section 3 for more details on the training dataset, and see Section 4.3 for more details on the backtesting process. This prompt was designed to guide the AI through the creation of predictive models and using them to generate investment decisions. Below is the full text of the prompt with some minor formatting changes, and below that is an example response.

> **Role Specification:**
>
> You are my expert academic research assistant specializing in finance and real estate. Your tasks must be performed at the highest level, akin to the standards expected at a PhD level in these fields. We are focused on accurate investment analysis and predictions, and your output should reflect this expertise.
>
> **Task Description:**
>
> This dataset contains a year-month panel of 433 U.S. cities with house price data from Zillow. The dataset includes demographic variables and Google Trends search indices, with descriptions of each variable below:
>
> - CityID: A unique numeric identifier for each city in the dataset
> - YearMonthID: A unique numeric identifier for the time period
> - RegionName: Name of the city **excluded from obfuscated version*
> - State: State of the city **excluded from obfuscated version*
> - Date: Calendar date of the observation **excluded from obfuscated version*
> - ZHVI: Zillow Home Value Index for the region
> - POP: Population of the city/metropolitan area
> - UnemployRate: Unemployment rate (%)
> - MortgageRate: Average 30-year fixed mortgage rate
> - GSVI:RealEstate: Google search intensity for general real estate terms
> - GSVI:RealEstateListings: Google search intensity for real estate listings
>
> An important note regarding the GSVI data is that each time series is independently scaled by its peak search volume. Therefore, direct comparisons between values across different series (companies and categories) are not appropriate; instead, comparisons between the dynamics (changes in GSVI) are more relevant. It is the change in these GSVI index values within a series that matters.

Your Task:

Model Development:

Use the provided data to create a variety of predictive models. Train all models using the 5 years of data in the dataset. The data is monthly. Implement cross-validation on this training data to evaluate the performance of these models. Ensure that the models are sophisticated enough to capture the nuances in the data, considering the high level of analysis required in the finance and real estate domains. Please use a variety of modeling approaches, ranging from simple linear regression models to the most complex modeling approaches available in this environment. Aim to estimate and compare around 4–5 specifications that incorporate domain-level knowledge about how these economic variables interact with each other.

Model Evaluation:

Evaluate each model using metrics such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), as well as any other relevant measures of model fit and parsimony. Additionally, compute the average mean squared errors (MSEs) for the testing sets in your cross-validation process. Based on these criteria, select the one best-performing model to use as your forecasting model.

Forecasting and Investment Selection:

Using your selected forecasting model, generate forecasts for the next 3 years following the training period. The data for this testing period is not provided to you, but will begin immediately after your most recent observation. Based on these forecasts, your next task is to create some portfolios by selecting cities to invest in over these 3 years. You will form four portfolios: GPT_5cities, GPT_10cities, GPT_20cities, and GPT_40cities.

IMPORTANT: City selections must be based on quantitative model forecasts, NOT sequential numbering. Never simply select cities in numerical order. Each portfolio must show clear evidence of distinct selection criteria. Portfolios will be equal-weighted and compared against benchmarks. When selecting cities for the portfolios, make the selections independently for each portfolio. You should aim to create portfolios that will achieve the largest risk-adjusted return (Sharpe ratio) over the testing period.

Return ONLY the final investment selections in Python code, each line referencing the chosen CityIDs. Here are all valid CityIDs you are allowed to use: Integers 1 - 433. Your overall response must have exactly 4 lines (one per portfolio). The selection can vary across models if different strategies are suggested by the forecasts. Each portfolio should be selected independently.

Investment Selections:

After your analysis, return your city selections (portfolios) in the format of the following Python code, with no other text or explanations:

```
GPT_5cities=[city#, city#, ..., city#] #Selected by [some model]
GPT_10cities=[city#, city#, ..., city#] #Selected by [some model]
GPT_20cities=[city#, city#, ..., city#] #Selected by [some model]
GPT_40cities=[city#, city#, ..., city#] #Selected by [some model]
```

Where for each line:

- Replace city#, city#, etc. with **actual** CityID numbers that appear in the dataset.
- The array must contain exactly the required number of unique CityIDs for that portfolio.
- Each comment must specify which model was used for the selection.
- Here are all valid CityIDs you are allowed to use: Integers between 1 and 433

IMPORTANT:

- Return ONLY those 4 lines of code with your selections (no additional output).
- Each portfolio must contain exactly the specified number of integers (no placeholders, zeros, or duplicates).
- All CityIDs should be an integer between 1 and 433.
- Never send empty portfolios
- Never send portfolios that are all identical
- No sequential/consecutive CityIDs
- Do not include any extra text, disclaimers, or explanations outside the code block.

END OF PROMPT

Example GPT-4o Response:

CONVERSATION:

Role: assistant

Type: text

Content:

```
GPT_5cities = [12, 54, 105, 230, 387] # Selected by Random Forest
   model demonstrating strong predictors of growth from GSVI
   changes
GPT_10cities = [17, 38, 82, 116, 145, 197, 262, 310, 377, 421] #
   Cluster analysis model optimizing POP and ZHVI variations
GPT_20cities = [5, 26, 33, 47, 68, 79, 92, 107, 129, 144, 150,
   174, 199, 211, 243, 269, 303, 328, 360, 404] # Time-series
   ARIMA identifying upward trends in ZHVI and POP forecasts
GPT_40cities = [3, 8, 11, 21, 23, 35, 40, 44, 56, 61, 64, 74, 81,
   83, 90, 98, 101, 113, 126, 132, 138, 152, 167, 183, 187, 202,
   215, 223, 237, 246, 259, 265, 283, 296, 311, 329, 345, 356,
   371, 423] # Extensive ensemble learning combining multiple
   factors for maximized Sharpe ratio
```