Large language models are successful retail forecasters

Author: Andreas Laksberg
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
Large language models (LLMs) have made headlines since the release of ChatGPT to the public in 2022. Quick advancements in their capabilities have made them a potential disruptor in many industries. This study investigated whether large language models could create and improve retail forecasting models without human interference and proposed a theoretical framework for adapting LLMs into retail forecasting. An experiment was conducted where two of the currently most advanced LLMs, OpenAI's o4-mini-high and Gemini 2.5 Pro, were tasked with improving the accuracy of their forecast models over 10- and 20-attempt series based on Walmart sales data from the M5 forecasting competition. The results showed that ChatGPT beat the accuracy of the best performing benchmark model by 10.2%. Gemini outperformed most benchmarks but lost to the most accurate benchmark by 1.8%. Meanwhile, Gemini showed off its learning capabilities and achieved statistically significant improvements to accuracy over a series of attempts while ChatGPT failed to produce statistically significant improvements over time. This study has explored using LLMs in retail forecasting, highlighting the potential of LLMs being able to automate a significant amount of the forecasting process in the future.

Graduation Committee members:
Dr. M. de Visser
Dr. R. Effing

Keywords
Large language models, generative AI, retail sales forecasting, time series, machine learning, autonomous forecasting.

During the preparation of this work, the author used ChatGPT and Gemini to conduct the experiment and ChatGPT to help generate ideas. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

1. INTRODUCTION

Forecasting is the process of predicting the future based on historical and present data (Graefe et al., 2013). Forecasting in the retail sector will never be perfectly accurate, yet even the slightest improvement in predictions can give an advantage over one's competitors in the highly competitive market of retailing. Between 1990 and 2012, six out of the ten top retailers in the US have fallen and been replaced by new winners. One of these winners is Amazon, whose large investments into data analytics and machine learning for decision-making helped the company become the largest e-commerce platform in the world (MacKenzie et al., 2013). Artificial intelligence (AI), more specifically large language models (LLMs) such as ChatGPT, are considered one of the most influential innovations in the last decade (Boyko et al., 2023) and must therefore also be considered a potential disruptor in retail forecasting. Forecasting has its roots in ancient times and has thereon evolved from relying solely on expert opinion to using time-series methods to building regression models, growing more complex with advancements in mathematics and digitalisation (Petropoulos et al., 2022). The next large development in the field of forecasting may come from the adoption of LLMs into the process.

Businesses are increasingly data-driven with a growing emphasis on making decisions backed by supporting data analytics (Fildes et al., 2022). There is, however, more data collected which is currently not being used due to a lack of analytics proficiency among employees as well as processing resources (Johnson et al., 2021; Li et al., 2021). The labour market thereby faces a high demand for data analysts that it is struggling to meet (Almgerbi et al., 2021). The shortage of labour will be further pronounced by the effects of the ageing population crisis (Bloom et al., 2010). As high-skilled workers become more thinly spread across various sectors which compete for labour, each forecaster must be able to do more work in less time. So far, it has only been possible to automate a limited amount of work going into accurate forecasting due to every forecasting scenario being unique and needing a custom-tailored model for best results (Seaman, 2018). While advancements in complex forecasting methods such as neural networks and machine learning models have already made forecasting more accurate (Makridakis et al., 2022a), implementing these methods takes even more resources than traditional time series models.

This raises the question of whether it is possible to increase the accuracy of retail forecasts and utilising the growing amount of collected data while simultaneously alleviating labour shortages and compensating for the lack of experienced forecasters and data analysts.

LLMs present an opportunity to simplify and automate the process of creating forecasting models. Since its release in 2022, the frontrunner in generative AIs, ChatGPT, has already led to many researchers evaluating the abilities of LLMs in performing various tasks. These attempts range from grading exams papers (Flodén, 2025) to programming (Bucaioni et al., 2024) and financial analysis (Cheng et al., 2024). LLMs offer endless possibilities for improving the productivity of labour. The release of ChatGPT has also evoked an LLM investment boom, forcing many companies out of their comfort zone by having to include AI in some form or another in their operating processes to appease investors (Xexéo et al., 2024). These developments should be considered in the context of forecasting as a potential solution to the previously discussed difficulties.

On the other hand, adopting LLMs into forecasting also has some potential drawbacks which must be considered before widespread implementation. There is growing resistance towards AI adoption due to fears of it replacing jobs and lacking data protection, amongst other issues (Wach et al., 2023). Lack of confidence in the accuracy of AI, combined with non-optimal prompts and limited knowledge about how AI comes up with its responses are also threatening to slow down the adoption of AI into business processes, including retail forecasting (Singla et al., 2024).

The academic debate around using LLMs in forecasting has grown considerably, but several knowledge gaps remain. Most studies have so far focused on feeding input data to the LLMs and asking for a prediction output (Jin et al., 2023; Tang et al., 2025). This method becomes less useful in real-world applications where datasets are too large and complex to run on the allocated memory of an LLM chat. This paper will address that research gap by taking advantage of the programming capabilities of LLMs (Bucaioni et al., 2024) and asking them to output a Python code that can be run locally to forecast with larger datasets. There is also a lack of benchmarking standards for evaluating LLMs against traditional methods (Paleka et al., 2025), which this study tries to address by using a competition's dataset where the forecast accuracy of LLMs can be compared against widely accepted benchmark models and the best human forecasters. Forecasting across different domains varies by methods used, forecasting complexity, and available data. This paper will focus specifically on filling the research gap in retail forecasting.

1.1 Research objective and question

Large language models have paved the way to automating more complex tasks than so far possible. Maximising the benefits of LLMs will help alleviate the effects of labour shortages and negative demographic trends. Not to mention giving a competitive advantage to the firms seizing those opportunities first. This research aims to explore the possibilities and advantages of using generative AI tools in retail forecasting. Therefore, the following research question has been formulated:

Can large language models create and improve retail forecasting models without human interference?

1.1.1 Research sub-questions

Three sub-questions will be formulated to help answer the research question:

1. To what extent can an LLM understand given datasets and create forecasting models to exploit that data?
2. To what extent can an LLM improve the accuracy of a forecast model over multiple attempts when only receiving performance measure feedback?
3. Which LLM achieves the best result and how does it compare against professional forecasters?

1.2 Academic and practical relevance

The goal of this research is to deepen the understanding of LLMs capabilities and limitations in the field of forecasting and data analysis. This study will try to fill a gap in existing literature regarding the use of LLMs in retail forecasting. This is done by gathering a clear overview of the current most advanced LLM models and conducting an experiment to test their forecasting abilities. Furthermore, this paper will propose a theoretical framework for adapting LLMs into retail forecasting. This framework will be used to highlight the key elements of forecasting that could eventually be automated with LLMs. The experiment conducted in this paper will serve as a model on how to further evaluate LLMs abilities of forecasting in future studies and how to compare different LLMs.

The practical relevance of this study is to offer an alternative to the lack of supply of data analysts on the labour market by

potentially replacing some of those jobs with LLMs and improving the performance of the remaining data analysts. In addition, generative AI can be a useful tool to compensate the lack of experience and technical knowledge among employees by assisting the construction of data analysis and forecasting systems.

In the modern-day highly competitive retail sector, having an advantage over competitors in forecasting processes offers an opportunity to cut costs and offer lower prices. Due to the large investments and the fast pace of improvements in the LLM industry, companies must actively monitor developments in the sector. Being aware of the capabilities and limitations of the latest LLMs is critical for ensuring an optimal LLM selection process.

2. THEORETICAL FRAMEWORK

The theory chapter is structured as follows. Different types of forecasting models that are used in retail will first be examined that differ from each other by complexity, strengths and weaknesses. The second subchapter will explore how LLMs are built, this should give a better understanding of how they generate their answers and what are the strengths and limitations of LLMs. These two subchapters will then be connected by exploring the latest literature from relevant topics and discussing the different possibilities of how generative AI models could be useful in retail forecasting. A theoretical framework highlighting the different levels of generative AI integration in retail forecasting will be presented to conclude the chapter.

2.1 Forecasting in retail

Forecasting has a key role in retail operations. Accurate forecasts save costs, improve sales, and enable functioning just-in-time supply chains (Ma & Fildes, 2021). Inaccuracies on the other hand lead to waste creation through overestimated sales and customer dissatisfaction through empty shelves and underestimated demand (van Donselaar et al., 2006). As such, different forecasting models have been created to accustom varying data types, complexity and patterns (Geurts & Kelly, 1986). A single one-size-fits-all forecasting method has not been invented and while some forecasting models perform better in certain scenarios, others outperform them elsewhere. Retailers are constantly working on improving their forecasting models to gain a competitive advantage over their rivals, yet the types of models they use is often similar (Fildes et al., 2022). The most common retail forecasting methods will be explored in this chapter. Their advantages and disadvantages will be discussed, and recent scientific studies will be summarised.

2.1.1 Time series models

Time series models analyse historical data to identify patterns over time; basic models reveal trends in data by using moving averages, either simple or weighted, to smooth out fluctuations (Petropoulos et al., 2022). Exponential smoothing models give extra weight to recent observations but otherwise work the same (Geurts & Kelly, 1986). These are relatively easy to create models, more complex time series models are for example autoregressive integrated moving average (ARIMA) and seasonal ARIMA that attempt to also capture cycles and seasonality (Van Calster et al., 2017).

Time series models are most useful when forecasting with consistent historical patterns, also they are easy to understand and create. The main limitations of time series models are poor performance in volatile or dynamic environments, inconsideration of causal variables, and degraded performance when using sparse or irregular datasets (Alon et al., 2001).

Time series models are used in inventory management and sales forecasting, ARIMA is for example used in grocery stores for forecasting daily demand of perishable goods and reducing waste (van Donselaar et al., 2006). While seasonal exponential smoothing is used for planning holiday sales in department stores (Geurts & Kelly, 1986).

Gruver et al. (2023) studied the ability of GPT-3, LLaMA-2 and GPT-4 to make zero-shot time series forecasts and found that they perform equally well or better when compared to purpose-built models. Noguer I Alonso and Pereira Franklin (2024) analysed the performance of using LLMs to improve financial time series forecasting and their results emphasise the potential for accurate forecasting using LLMs.

2.1.2 Machine learning models

Machine learning (ML) models use algorithms to identify patterns in data, ML models that are used in retail include decision trees, random forests and gradient boosting machines such as XGBoost and LightGBM (Gai, 2025). Deep learning models are a complex version of ML models that have risen in popularity due to high-volume and data-rich e-commerce's search for more accurate forecasting (Loureiro et al., 2018). These models use artificial neural networks with multiple hidden layers, essentially analysing inputs through a black box and outputting a prediction (Gai, 2025).

ML models are great for finding patterns in volatile and dynamic environments, more advanced ML models can also integrate various types of structured and semi-structured data such as clickstream data and loyalty program data (Wang et al., 2021). The main disadvantages of ML models are their poor scalability, requirement for large volumes of clean data for training, specialised expertise needed for creating the models, and resource-intensive running and updating of the ML models (Fildes et al., 2022).

ML models are applied to a wide range of retail forecasting tasks. This includes demand forecasting, customer segmentation, price optimisation, customer churn analysis, and product recommendations (Wang et al., 2021). Hasan (2024) used ML to more effectively capture seasonality and trends in sales forecasting than traditional methods. Amir et al. (2023) showed the benefits of using convolutional neural networks for accurate sales predictions with significant seasonal and regional variations in the datasets. Wellens et al. (2024) showed that simplified tree-based methods can provide high accuracy and efficient computations for sales forecasting, which makes them suitable for large-scale retail datasets. Alice and Srivastava (2023) demonstrated that XGBoost can outperform many traditional models in sales forecasting by being able to handle complex relationships within large datasets.

2.1.3 Hybrid models

Hybrid models combine the previously described forecasting models to leverage each one's strengths and improve accuracy, for example, an ARIMA model can be effectively combined with a neural network model for improved results (Huber & Stuckenschmidt, 2020). Multiple ML models can be similarly combined, where each one trains on certain parts of the complete dataset or applies different tuning parameters, increasing overall accuracy of the forecast (Makridakis et al., 2022a).

The main advantages of hybrid models lie in their increased flexibility and improved accuracy, hybrid models are a balancing act between utilising different forecasting methods to maximise accuracy given the forecasting scenario while minimising the computational cost for faster modelling (Mediavilla et al., 2022; Petropoulos et al., 2022; Zhang, 2003). The disadvantages are mainly trade-offs of developing these capabilities, increasing the demand for highly skilled data analysts in the company while

also requiring more time and resources for creating and optimising these hybrid models (Petropoulos et al., 2022).

Tran et al. (2023) used a hybrid sales forecasting model that combined time series analysis with ML methods to capture both short-term fluctuations and long-term trends. Gandhi et al. (2023) created a novel hybrid approach for sales forecasting, their fuzzy pruning least square support vector machine (LS-SVM) model increased forecast accuracy by addressing non-linearity. Liu et al. (2023) created a combination model that used multi-angle feature extraction and social media to combine sentiment analysis with traditional sales data to significantly improve forecasting accuracy for electric vehicle sales.

2.2 Large language models

LLMs are advanced deep learning models which can perform a variety of language processing tasks. This is possible because LLMs are pre-trained on large datasets, and they are capable of learning complex patterns in that data (Naveed et al., 2023). More specifically, LLMs are defined as foundation models, which on a technical level are efficient because of transfer learning and the LLMs' scale (Bommasani et al., 2021). Foundation models are trained on a broad range of unlabelled

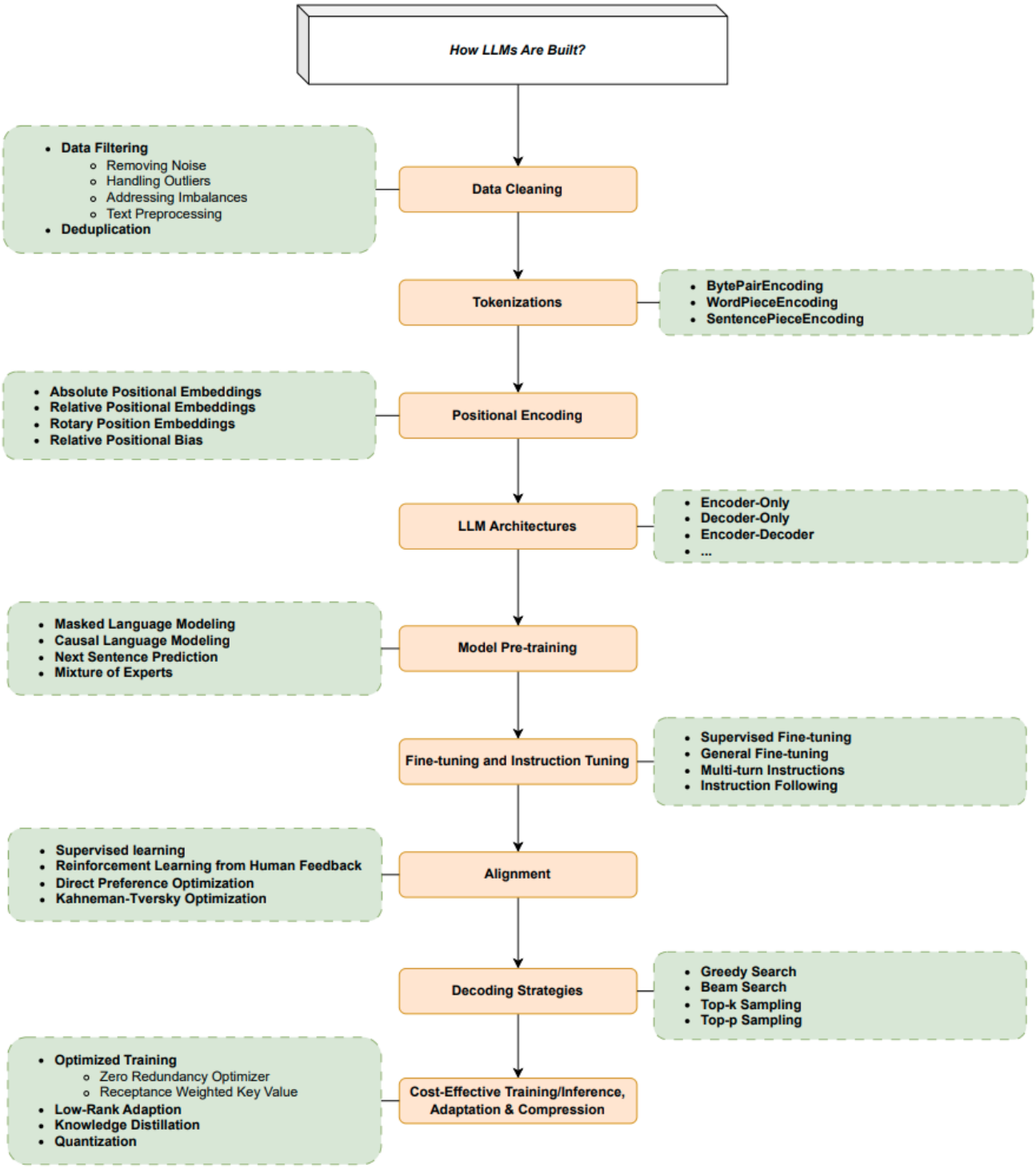

Figure 1. Different components of LLMs and how LLMs are built by Minaee et al. (2024).

data with minimal fine-tuning and can be used for many different tasks.

Minaee et al. (2024) created a clear overview of the necessary steps that are required to create a modern LLM as seen on Figure 1, the process also involves multiple critical decisions between various encoding, decoding, and architectural strategies. Most of the modern advanced LLM architectures are founded on the transformer architecture first introduced by Vaswani et al. (2017). This enabled significantly faster training times and superior performance over previous architectures using recurrent or convolutional layers. Vaswani et al. (2017) key innovation was the self-attention mechanism of transformers, allowing LLMs to weigh the importance of different tokens in an input, which enables them to understand context by identifying relevant connections between parts of the input and the token being processed.

LLM development involves two primary training phases, which are pre-training and fine-tuning. The pre-training phase helps LLMs develop language understanding capabilities and this phase is usually unsupervised or self-supervised (Minaee et al., 2024). Fine-tuning is then used to better adapt the general model to specific tasks or domains, fine-tuning is also used to align outputs with user preferences and stop the LLM from publishing potentially undesirable content and bias (Bommasani et al., 2021).

2.2.1 Limitations and risks of LLMs

LLMs are still a relatively new technology, so they exhibit many limitations and risks which must be considered by their users. Naveed et al. (2023) summarise a large variety of these challenges and some of those have already been solved by the latest LLMs. This chapter will shortly highlight the most relevant issues with regards to adopting LLMs in retail forecasting that are still prevailing:

- **Hallucinations**: LLMs sometimes display plausible sounding but false information.
- **Overfitting**: LLMs have great learning capabilities, yet noisy and peculiar patterns in their training data may lead to overfitting, which will cause illogical responses.
- **Reasoning and planning**: Some tasks, which might be doable for humans, might go beyond the logical reasoning and planning capabilities of current LLMs.
- **Long-term dependencies**: LLMs can fail to manage long-term dependencies and preserve context, especially in complex and long conversations or documents. This may result in incoherent or incorrect responses.
- **Prompt engineering**: The syntax and semantics of input prompts play a critical role in the output quality of models. Slight input variations can lead to wildly different outputs from the model.
- **Privacy concerns**: Using an LLM for retail forecasting might necessitate sharing confidential company data with the LLM to procure the best results. However, there are concerns that this private data is then used to train future models, leading to potential exploitation of this sensitive data by adversaries. This must be an important consideration for companies before deciding which LLM to incorporate into their operations.

2.3 LLMs in retail forecasting

The next step is to combine the capabilities of LLMs with the requirements of generating accurate forecast models. In this chapter, it will be discussed why LLMs could even be useful in retail forecasting and three possible use cases on how an LLM might be able to assist with retail forecasting will be highlighted. These situations grow gradually more complex and signify the increasingly difficult work that—if successfully implemented—LLMs can assist data analysts with.

2.3.1 Why LLMs could be useful in retail forecasting

Some work has already been done to evaluate the skills of LLMs in mathematical forecasting problems and the results have been promising. Lopez-Lira and Tang (2023) analysed whether LLMs can forecast stock price movements and concluded that more advanced LLMs like ChatGPT can effectively sort through complex information and ChatGPT's predictions outperform traditional methods. Lin et al. (2025) used ChatGPT to create a more accurate credit rating system and successfully improved the development of more accurate credit rating forecasts for small and medium-sized companies than traditional models. Tang et al. (2025) used LLMs for time series forecasting and showed that LLMs are great at predicting datasets with clear patterns and trends but were struggling when the data was lacking periodicity. Jin et al. (2023) reprogrammed LLMs into their TIME-LLMs that outperformed specialised forecasting models in time series forecasting. Zhang et al. (2024) fine-tuned LLMs into what they call LLMForecaster to incorporate unstructured information and historical data into an existing demand forecasting pipeline. This led to significant forecast improvements subject to holiday-driven demand surges. Ghasemloo and Moradi (2025) leveraged auxiliary knowledge to increase LLM performance in time series forecasting and showed that it significantly outperformed the baseline with no auxiliary information. They highlighted knowledge transfer strategies as a potential way to close the performance gap between LLMs and domain-specific forecasts. Park et al. (2025) also tried to improve the effectiveness of LLM zero-shot time series forecasting but eventually concluded that their sensitivity to noise limits their ability to achieve high accuracy.

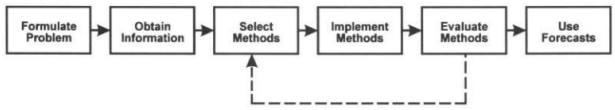

Figure 2. Generalised forecasting process by J. Scott Armstrong (2001).

As can be seen on Figure 2, J. Scott Armstrong (2001) formulated a generalised framework for forecasting which is an extended adaptation of the Box and Jenkins (1970) methodology. The first step is to formulate the problem and specify objectives. The second step of forecasting is to have clean usable data, ChatGPT has already been shown to be capable of cleaning and filtering, saving time and effort that data preparation usually takes before forecasting can even begin (Hassani & Silva, 2023; Zhang et al., 2023). As ChatGPT has already been shown to be competent and fast in data cleaning, this will not be evaluated in this paper, the focus will instead be on whether an LLM can successfully perform the learning cycle. This means that the LLM should be able to select, implement, and evaluate forecasting methods, and then repeat the process to improve forecast accuracy.

2.3.2 Creating forecast models with an LLM

It is important to understand the data, scenario and purpose of the forecast to choose an appropriate model (Armstrong, 2001). This is the first challenge the LLM must overcome. The chosen model must be created in a programming language. This code needs to correctly read the datasets, calculate a forecast and write an output file in the format requested by the user. The LLM needs to demonstrate a clear understanding of the problem and the

user's prompts to write a seamlessly working code which the user could run without prior modifications. The most advanced LLMs allow uploading different file types in addition to a prompt. They have been trained to be able to read and understand the structure and content of these files (*OpenAI Code Interpreter*, 2025), similarly to how they can understand various text structures. They combine this with their forecasting skills that they have acquired by training on the vast pool of forecasting knowledge available on the internet. This should allow LLMs to be capable of understanding datasets and creating forecast models to utilise the data.

2.3.3 Improving forecast models with an LLM

Improving a forecast model's accuracy is challenging, it requires testing through trial-and-error, understanding the weaknesses of a model requires logical reasoning. In more complex forecasting models, a balance has to be found between overfitting a model to the training data and not capturing relevant trends (Ulrich et al., 2022).

For an LLM to be able to improve forecast models, it must understand the effects of each change to the model and what they imply. It must also correctly evaluate which data is valuable for the forecast and which should be disregarded as noise.

2.3.4 Autonomous LLM forecasting

If an LLM could forecast autonomously, it would be able to analyse the data, decide which model to use, run the model by itself and improve that model based on feedback loops. This would require the data analyst to only write prompts, saving considerable time and energy to work on other tasks rather than manually testing slight adjustments to forecasting models.

2.4 Framework for adapting LLMs into retail forecasting

While Chapter 2.3.1 gave an overview of the latest studies using LLMs for forecasting, the abilities of LLMs in retail forecasting have so far not been evaluated in any published scientific literature. This gap will be filled by implementing the use cases of LLMs previously discussed into a framework. That framework will then be used as the foundation for this paper's experiment, designed to help answer the research question: "Can large language models create and improve retail forecasting models without human interference?" Figure 3 shows this paper's theoretical framework for the tasks an LLM must be able to complete for retail forecasting to advance towards automation. These tasks are increasingly complex in the sense that each requires more logical reasoning and comprehension of the process. That is where LLMs have shown unstable performance and hallucinations in the past (Laban et al., 2023; Su et al., 2024).

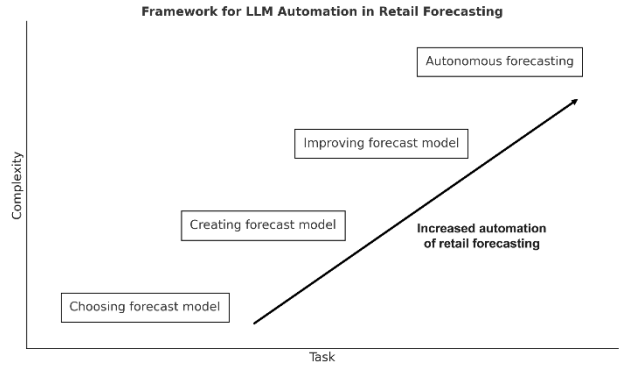

Figure 3. Framework for LLM automation in retail forecasting.

The first step is choosing a forecast model appropriate to the data and forecasters needs. LLMs can use their vast pool of online sources to compare which models are generally used for similar datasets as the one they have. A slightly more complex task is to then create the forecasting model so that the user can run the model without having to make any quick fixes and adjustments. Improving forecast models is the third task and is significantly more complex than the last two. This requires the LLM to exhibit logical reasoning and consider which variables might help it better forecast the data. When the adjusted forecast performs worse than its predecessor, the LLM must also be able to reason what made the model perform less efficiently and use this insight towards a more accurate forecast. Autonomous forecasting is the last and most complex task of this framework. This stage implies that the LLM only needs starting directions and an end goal from the user, after which it can independently work on improving the forecasting model until it can no longer find a way to improve the model's accuracy any further.

3. METHODOLOGY

3.1 Sample

The M5 forecasting competition public dataset will be used for this study. The dataset is created for academic research by Makridakis et al. (2022b) and is published on Kaggle (Howard et al., 2020). It contains Walmart's sales data of California, Texas, and Wisconsin. The data is gathered from 10 different stores from 29 January 2011 to 19 June 2016. Daily total sales of 3049 products are shared, additional datasets include calendar metadata with holidays and events, sample submission formatting, and sell prices of products per store and date.

There are many different LLMs on the market with more in development and each getting new iterations on a regular basis. This makes it unfeasible to benchmark all the models within the scope of this research paper, in addition, the main goal of this paper is not to necessarily compare LLMs against each other but to evaluate the forecasting abilities of the current-best generative AIs. The constant release of newer models means that previous literature on which models to benchmark is scarce and already outdated. Two generative AIs will be chosen for the analysis.

Previous ChatGPT versions have consistently outperformed their competitors (Abolghasemi et al., 2025; Liu et al., 2024) making it the first AI chosen for the analysis. More specifically, the **OpenAI o4-mini-high model** will be used as it is the newest version available that is recommended for coding and advanced reasoning (*OpenAI Models*, 2025). The second AI is Google's Gemini that has played catch-up on OpenAI since 2022 but has shown strong performance with its latest **Gemini 2.5 Pro model** (Artificial Analysis, 2025; Vals AI, 2025).

3.2 Method

This study aims to evaluate LLMs' abilities to improve the accuracy of forecast models. An LLM is given a set of four dataset samples, and it is then asked to create and improve a forecasting model to be as accurate as possible within 10 and 20 test runs. After each test the LLM will get feedback on their accuracy. Following the M5 competition structure, LLMs will receive feedback in a weighted variant of root mean squared scaled error (WRMSSE). WRMSSE is suitable for the evaluation of this forecast because it can be safely computed for all series and aligns with the objective of trying to accurately forecast average sales (Makridakis et al., 2022a). WRMSSE translates into forecast accuracy and is the target (dependent) variable of the forecasts, the goal for the LLMs is to get their WRMSSE score as low as possible.

The starting prompts given to the LLMs are found in Appendix A. The LLMs were additionally given a small sample file of every dataset, so they could inspect the structure of data and adapt their codes to work seamlessly. These datasets had slightly modified features (e.g. different file name, IDs of products) compared to the original datasets found on Kaggle to better suit the test and to avoid LLMs submitting unmodified code scraped from Kaggle forums. The first dataset contains daily sales data for 3049 products over five years. Each unique product ID is created by merging state, store, category, department, and item ID, this enables the use of easily connecting variables from the other datasets to improve forecasts. Calendar dataset includes information about holidays, potentially influential events (e.g. Super Bowl), and which states had their Supplemental Nutrition Assistance Program (SNAP) active on which dates. The third dataset includes weekly sale prices for every product and the fourth one is a sample submission file to make sure LLMs prepare a uniform output each time.

To ensure comparability, user inputs to the LLMs were kept to a minimum after the initial prompt and they are specifically told to work independently and make decisions based on its own reasoning, so to not ask for the user's opinion. This is established with the following section of the starting prompt: "I will run your code and return results. Do not ask me for which improvements to make, do what you think will be the most accurate" (Appendix A).

3.3 Analysis

LLMs will be asked to write their forecast model into a Python script. The forecast models made by the LLMs will be copied and ran locally on a computer. Outputs of these models will be reformatted to the M5 format by a separate script and uploaded to the Kaggle M5 competition submission page to use their evaluation algorithm. The resulting WRMSSE will then be returned to the LLMs as feedback after which they will submit an updated script and so on. In the case a script returns an error, this will be copied back to the LLM, so they can fix their code and resubmit a working version. Both ChatGPT and Gemini will go through three series with 10 attempts to improve and three series with 20 attempts, totalling 180 forecast models.

To answer the first research sub-question: "To what extent can an LLM understand given datasets and create forecasting models to exploit that data?" It has already been shown in previous studies that LLMs are already quite capable of analysing datasets and generating forecasting models, as was explored in the theory chapter. Various positive and negative experiences regarding this question will nevertheless be discussed based on this paper's experiment as well. Observations will be given about the LLMs performance and stability during the experiment and common errors, including their potential causes, will be considered.

The second research sub-question was: "To what extent can an LLM improve the accuracy of a forecast model over multiple attempts when only receiving performance measure feedback?" So do LLMs learn from their previous tries and improve over time. If they do improve, then is there a limited number of runs with increased accuracy or do they continue to improve consistently within the limits of this test. The strength of the relationship between forecast accuracy (dependent variable) and number of attempts (independent variable) will first be tested with Pearson's correlation coefficient (r). In the M5 competition, 24 forecasting methods were used as benchmarks to compare the performance of competitors to easily implementable forecasting models (Makridakis et al., 2022a). The best performing benchmark was exponential smoothing with bottom-up reconciliation (ES_bu). The performance of LLMs will additionally be compared to those benchmarks to help answer the second sub-question.

To answer the third research sub-question "Which LLM achieves the best result and how does it compare against professional forecasters?" The performance of the different LLM models will be evaluated against each other. Pearson's correlation coefficient will be used to compare the learning capabilities of LLM, while the highest accuracy models of each LLM are compared to determine the best performing LLM in retail forecasting. These tests will help establish which one is currently the most suitable LLM for improving forecasting models. The accuracy of the LLM models will also be compared to the results of people who originally competed at the M5 forecasting competition hosted by Makridakis et al. (2022b) and Walmart. This provides insights into potential strengths and limitations that LLMs have when compared to human forecasters.

4. RESULTS

4.1 Forecast creation

After an initial prompt engineering to guarantee the LLMs would clearly understand the task and be able to function without any other feedback than the forecasting scores, both LLMs were able to successfully finish all six series of forecasting with the starting prompts shown in Appendix A. Both OpenAI's o4-mini-high and Gemini 2.5 Pro were able to easily understand the task and create forecasting models in the Python programming language. When errors did occur in their code, both LLMs were able to fix the mistakes after receiving the traceback and error output from the terminal.

Both models tried implementing both simpler time series models and advancing to more complex LightGBM machine learning models. While both were able to create time series models without any serious complications, the large datasets led unoptimised ML models to return a memory error. Trying to keep millions of cells of data in memory while training the model is not a feasible solution for any forecasting task that deals with large amounts of data. The LLMs dealt with this error in one of two ways. Sometimes they decided that the datasets are too large for ML models and went back to optimising their time series model. Around half the times they decided to instead stick to implementing the ML model and tried to optimise it by either limiting the training data or employing other methods of memory optimisation.

There were also some negative observations that must be mentioned. Retrieving data from the calendar dataset highlighted on multiple occasions how LLMs can rely too heavily on their training data and forget the user's input. For example, an important column's ID was "wm_yr_wk", but the LLMs sometimes tried to fetch the data with IDs like "week" and "date", which are generally common IDs in calendar datasets. After returning the error to LLMs, they scanned through the sample file again and were able to fix the mistake. Sometimes they also tried to use functions with unsuitable value types which they were then able to fix by converting the value types. The longer and more complex the code became, especially when the LLMs tried to implement new forecasting factors, the more error codes and tweaks the LLMs were likely to go through before reaching a functioning code that returned an output. Some of these issues arose from the LLM trying to use outdated function variables from Python's packages. This is likely a consequence of the LLMs' training data also including older code repositories that are no longer applicable in the newest version. Most other issues arose from seemingly human mistakes where the LLM forgot to define a variable or made a typo in the code. On a few occasions, the LLMs ran into a new error when trying to fix a

previous error multiple times in a row, and once the LLM eventually had to revert to a time series model when not being able to fix its ML forecast model. All these mistakes only happened in the longer and more complex forecast models.

4.2 Forecast improvement

In the first test, the LLMs were given 10 attempts to improve their forecasting accuracy. Figure 4 shows how these results changed over the attempts for both ChatGPT and Gemini. Additionally, three guiding benchmarks from Makridakis et al. (2022a) are shown for comparison. The green line is a simple naïve benchmark that forecasts values as equal to the last known time series. This benchmark achieved a WRMSSE of 1.752. The orange line is a seasonal naïve benchmark that can also capture some possible seasonal variations. It received a WRMSSE of 0.847 which is a significant improvement and beats 16 out of the 24 benchmarks used. The red line is an exponential smoothing with bottom-up reconciliation benchmark that was the most accurate out of all the benchmarks employed with a WRMSSE of 0.671.

In the 10-attempt scenarios, ChatGPT managed to outperform the best benchmark ES_bu on all three occasions with the best scores of three series being 0.665, 0.603, and 0.662, respectively. The best ChatGPT score was 10.2% more accurate than ES_bu. In the first run, ChatGPT started with a very simple time series model and continuously tried adding new features to it until it finally reached its best score in round 9 using multiple variables from the calendar data in addition to the sales data. In the second run, ChatGPT decided to start with a LightGBM model from the get-go. The overall best result 0.603 from the second run was achieved on the third attempt of the series with the previous two attempts scoring 1.005 and 0.615. The third run achieved its best result only on the very last attempt. While in the first series ChatGPT stuck to time series models and in the second series to LightGBM models, in the third series it started with the seasonal naïve model used in the benchmark. It then tried and got a terrible result of 3.381 with a LightGBM model after which it went back to time series models until attempt 7 when it decided to give ML another try. By the last round it got the LightGBM model optimised enough to barely beat the ES_bu benchmark.

The best scores that Gemini achieved during its three 10-attempt scenarios were 0.683, 0.8034, and 0.701, respectively. ES_bu was only 1.8% more accurate than Gemini's best result. Although Gemini did not manage to beat the ES_bu benchmark on any of its attempts, it showed consistent improvements over time as seen on Figure 4. In the first two series, Gemini stuck to time series forecasts, steadily implementing additional features and tweaking their weights to optimise the model. Only by the

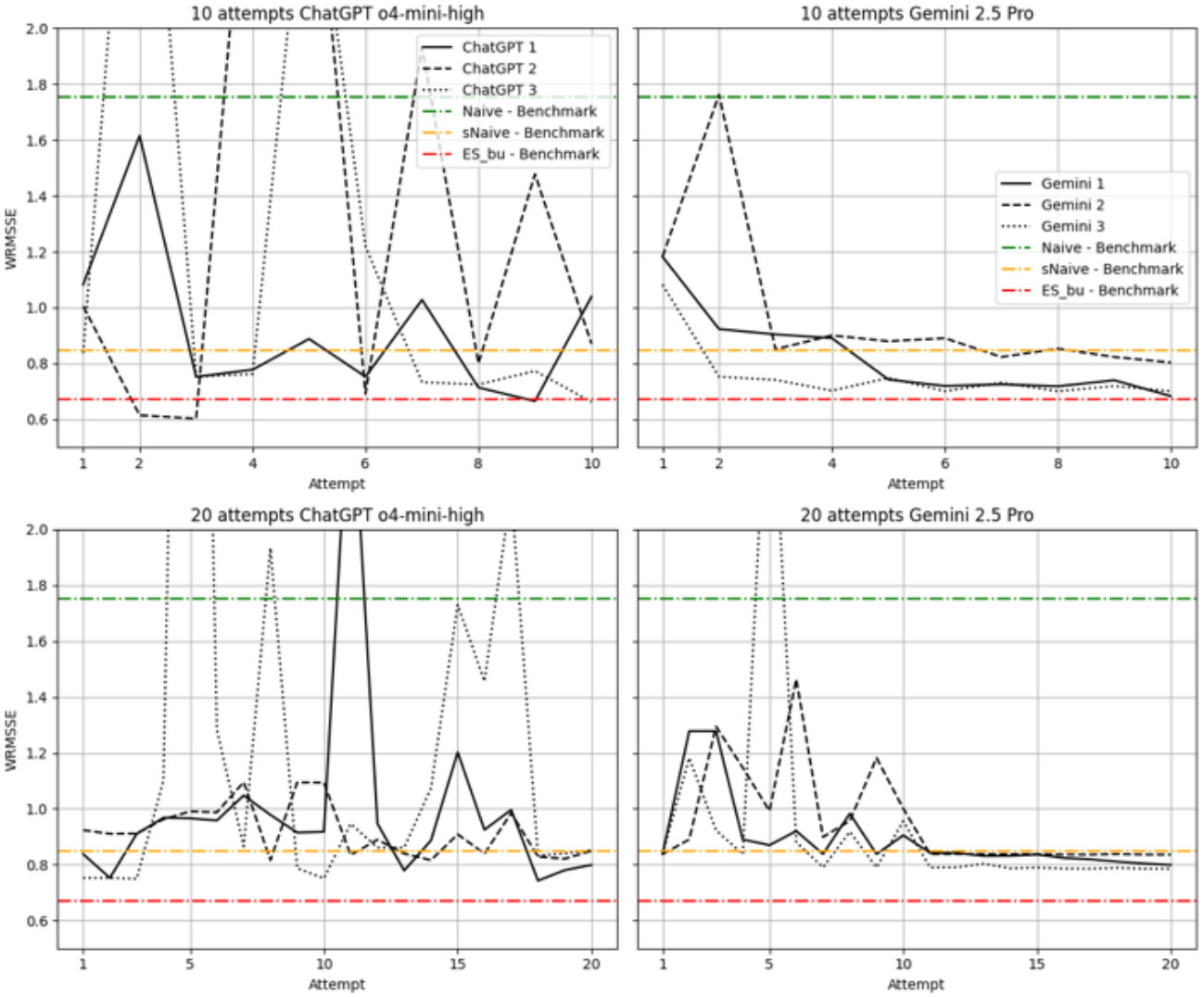

Figure 4. Forecast accuracy of ChatGPT and Gemini over 10 and 20 attempts compared to three benchmark models. Lower is better.

fourth attempt in the third series did Gemini decide to try implementing a LightGBM model and did not retreat to time series models despite receiving a memory allocation error at first. In the first two series, Gemini earned its best result by attempt 10, showing gradual improvement of its time series model over time. In the third series, Gemini's LightGBM model reached its peak result by attempt 8.

Table 1. Learning performance in 10-attempt series

Model	Slope	Intercept	Pearson r	p-value
ChatGPT	-0.056	1.496	-0.338	0.3393
Gemini	-0.043	1.089	-0.819**	0.0038

Table 1 shows the learning performance of ChatGPT and Gemini during the 10-attempt series. A negative slope and regression mean that the WRMSSE decreases over time which might indicate learning capabilities over multiple attempts. ChatGPT had a slope of -0.056 and a Pearson's r of -0.338, however the p-value was 0.3393 which means the correlation is not statistically significant. Gemini had a slightly smaller slope of -0.043 but a much stronger correlation with Pearson's r reaching -0.819. Gemini's p-value is also 0.0051, which makes the correlation statistically significant.

20-attempt scenarios were less successful for both ChatGPT and Gemini. Both managed to beat the seasonal naïve benchmark but neither got sub 0.7 results. ChatGPT's best results from each series were 0.742, 0.816, and 0.749, respectively. ChatGPT only used time series models in the 20-attempt runs. It tried to run LightGBM models but after receiving a memory error decided to return to time series models and implement unnecessary memory optimisations to those. Gemini's best results from these three series were 0.799, 0.836, and 0.785, respectively. Interestingly, Gemini started its first attempt with a seasonal naïve model on all three runs. In the first run, Gemini decided to keep using time series models till the end. This changed in the second and third run. It used a LightGBM model in attempt 7 of the second run but after that failed to fix a slew of errors that popped up with its next LightGBM model, making it eventually return to time series models. Gemini also tried to use LightGBM in the fifth and sixth attempt of the third run. It scored a poor 2.750 for the fifth attempt but improved the LightGBM model to a 0.879 WRMSSE in the next attempt. It then decided that because the LightGBM model did not beat the best performing simple model, it should return to refining the time series model instead and did not try LightGBM again. ES_bu was 9.6% more accurate than the best ChatGPT model and 14.6% more accurate than the top performing Gemini model in the 20-attempt series.

Table 2. Learning performance in 20-attempt series

Model	Slope	Intercept	Pearson r	p-value
ChatGPT	-0.011	1.193	-0.147	0.5364
Gemini	-0.019	1.122	-0.601**	0.0051

Table 2 shows the learning performance of ChatGPT and Gemini during the 20-attempt series. Both models have smaller average slopes in these series, -0.011 for ChatGPT and -0.019 for Gemini. ChatGPT has a Pearson r of -0.147 but it is still not significant with a p-value of 0.5364. Gemini also has a weaker correlation than in the 10-attempt series, yet it is still statistically significant with an r of -0.601 and a p-value of 0.0051.

When comparing 10-attempt learning series to 20-attempt series, for some reason, almost everything about the LLMs performance was better during the 10-attempt runs. The shorter runs developed more accurate forecasting models, handled errors more effectively, and had better learning performance. Gemini's learning performance slope was over two times steeper (-0.043 to -0.019), with a better intercept value (1.089 to 1.122), and with a higher correlation (-0.819 to -0.601).

Figure 5 shows the total distribution of the accuracy achieved by the LLMs during all the tests compared with all the benchmarks implemented by Makridakis et al. (2022a). The peak of the distribution curve for the LLMs was equal to the seasonal naïve benchmark. Five of the twelve runs started from there; some tested its accuracy later in the series and others achieved very similar accuracy with different methods. Majority of the benchmarks are to the right of the LLMs density curve's peak, which means that most of the LLMs attempts beat these benchmarks. Gemini's best scores overlap the best performing benchmarks while ChatGPT's best scores make up the left tail of the density curve that outperformed all benchmark forecasts.

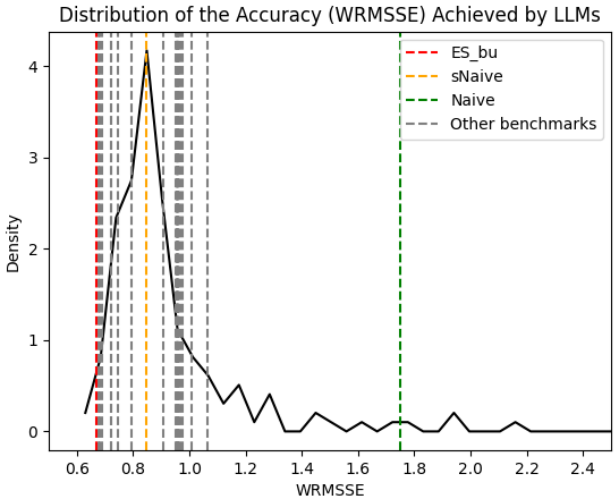

Figure 5. Distribution of LLMs' forecasting accuracy.

4.3 Forecaster accuracy comparison

Answering the third research sub-question requires direct comparisons between the performance of ChatGPT and Gemini. OpenAI's o4-mini-high managed to clearly outperform Gemini 2.5 Pro in terms of best forecast accuracy. In the 10-attempt series, ChatGPT's all three runs achieved a better high score than the best Gemini attempt. ChatGPT's best score was 11.7% more accurate than the best Gemini result. In the 20-attempt series, ChatGPT had two runs outperform Gemini with the best outperforming Gemini by 5.4%. One of ChatGPT's runs managed to only beat the worst performing Gemini series and lost to the other two. On the other hand, when comparing stability and continuous improvement of the LLMs forecasting abilities, then here ChatGPT struggles, and Gemini is the clear winner.

Figure 6 shows a modified version of the previous distribution graph, where ChatGPT (black curve) and Gemini (blue curve) are shown separately, and the lines are converted into Gaussian kernel density estimations (KDE). ChatGPT has a flatter bell-shaped distribution with a peak that is slightly worse than the seasonal naïve benchmark. Both of its tails are higher than Gemini's, revealing better top performance but also a higher number of badly performing forecast models. Gemini shows a tighter KDE peaking at a slightly better WRMSSE than the seasonal naïve benchmark. It produced more high-quality forecasts than ChatGPT and less badly performing forecasts, but at the same time Gemini failed to generate such top performing models as ChatGPT did.

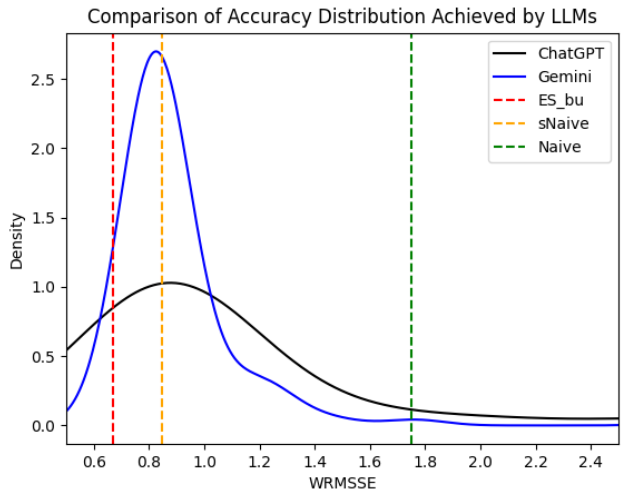

Figure 6. Comparison of LLMs' accuracy distribution.

In the original competition by Makridakis et al. (2022a), 2666 (48.4%) of participating teams managed to outperform the naïve benchmark, 1972 (35.8%) beat the seasonal naïve benchmark, and only 415 (7.5%) outperformed the ES_bu benchmark. The authors offer multiple valid explanations as to why only such a little amount of people managed to beat the best performing benchmark. Some comparisons can still be made between the results obtained by humans at the competition and the results of LLMs. Using these statistics, the best ChatGPT models outperformed more than 92.5% of human forecasters participating at the competition. The winning team of the competition managed a WRMSSE of only 0.520, this is a 22.4% improvement over ES_bu, a 13.7% improvement over the best ChatGPT result, and 23.9% more accurate than Gemini. It is important to consider that unlike the LLM forecasts, the competitors were able rely on extensive feature engineering, data analysis, additional feedback mechanisms, and combining the results of multiple forecast models. The winner of the competition, In and Jung (2022) used 10 per store models, 30 per store-category models, and 70 per store-department models. He also considered two variations per model, so in total 220 models were built. For each series in the data, the average of six models using different learning approaches and training sets were used, with additional fine-tuning done before choosing a final solution (Makridakis et al., 2022a).

5. DISCUSSION

5.1 Conclusion

This study set out to answer the research question: "Can large language models create and improve retail forecasting models without human interference?" To help answer this, the question was divided into three sub-questions focusing on (a) how well can LLMs understand datasets and create forecasting models, (b) to what extent are LLMs capable of improving the accuracy of forecast models with performance measure feedback, and (c) how the performances of different LLMs compare between each other and to humans.

Within the literature review, it already became clear that LLMs should be capable of understanding different datasets and writing code to create a forecasting model. This was further reinforced by the findings—both ChatGPT and Gemini could easily understand the structure of data and created forecasting models ranging from the most basic time series models to complex machine learning models composed of hundreds of lines of code. This process was not entirely seamless, however, as issues such as the LLMs using a generic calendar ID instead of the ID in the calendar dataset occurred at times, and had to be fixed by the LLM after encountering a runtime error. This aligns with one of the common limitations of LLMs, which was also discussed in the theory chapter, to potentially overfit training data and return illogical answers. In the longer and more complex forecast models, the LLMs encountered more frequent errors which they had to fix for the code to work. This indicates that LLMs still sometimes struggle with long-term dependencies, another issue highlighted in the literature review.

Answering the second research sub-question is more complicated. ChatGPT managed to beat all the benchmark models with its best forecasts with the highest performing one beating ES_bu by 10.2%. However, there was no clear improvement across the attempts. ChatGPT showed rather inconsistent performance, creating a well-performing model but then managing to ruin it after an attempt or two when adding extra features. Meanwhile, Gemini's best results were slightly less accurate than the best forecast, losing to it by only 1.8% of accuracy. Where Gemini did stand out was showing consistent improvement over time. It had a statistically significant reduction to WRMSSE across its attempts, achieved by methodically testing out the effects of adding new features to the accuracy, and then building on these results. Interestingly, both LLMs performed significantly worse when given 20 attempts to improve their forecasts instead of 10 attempts. This could again be explained by the LLMs' weakness with long-term dependencies, although it is not clear what made the LLMs both less accurate and more inconsistent in the longer series.

ChatGPT and Gemini were both able to handle the task of creating and improving forecast models. While ChatGPT achieved higher top scores, it also had many more highly unsuccessful attempts where the WRMSSE hiked up considerably. Gemini, on the other hand, showed a much more consistent performance, testing out different forecast features and weights between features to find a well-balanced forecasting model. This means that there is no clear winner between ChatGPT and Gemini—one got a better accuracy high score, while the other exhibited the ability of learning and improving over time, which was one of the primary focuses of the overall study. Both LLMs managed to beat most human competitors at the M5 forecast accuracy competition. The top performers at the competition, meanwhile, clearly outperformed the LLMs' results. This was an unfair fight, however, as the humans were able to combine hundreds of ML models into a complex hybrid model, with each ML model predicting different variables using different training sets and learning approaches. LLMs would benefit from an experiment where they have a more hands-on operator to test whether they are able to accomplish the creation of such complex hybrid forecasting systems.

To answer the overall research question, it can be concluded that large language models are indeed capable of creating and improving retail forecasting models without human interference to a certain degree. They can create time series models ranging from the simplest to more complex models that use multiple datasets with various forecasting features. LLMs are also able to create various ML models, but with both types of models, the more complex they get, the more likely it is that the LLMs find themselves struggling to fix mistakes in the generated code. The specific LLM model used has a large impact on the achieved accuracy of the forecast, and different models have varying consistency regarding their learning capabilities and forecast improvement across multiple attempts. It is not completely clear what is an optimal number of attempts an LLM should be given to achieve the highest forecasting accuracy.

5.2 Practical implications

This study has introduced LLMs into retail forecasting, offering a new potential tool to automate significant parts of the forecasting process, which is crucial due to the shortage of experienced data analysts. In the modern-day highly competitive retail sector, having an advantage over competitors in forecasting processes can offer an opportunity to cut costs and offer lower prices.

Another important implication is that different LLMs are better at different aspects of retail forecasting, so depending on the needs of the company, an LLM or a combination of LLMs should be chosen to fit these needs. Due to the large volume of investments and the fast pace of improvements in the LLM industry, companies must actively monitor these developments. Being aware of the capabilities and limitations of the latest LLMs is critical to ensure an optimal LLM selection process.

5.3 Theoretical implications

This paper has expanded the scientific literature on LLM capabilities. While the number of papers has grown exponentially since the introduction of ChatGPT, there are still large gaps in the literature. Each new iteration of LLMs is introduced with new capabilities and functionalities. This makes LLMs a potential solution for increasingly more business applications, further exacerbating the problem of missing literature. While most literature has so far focused on giving input data to LLMs and asking for a prediction output, this paper has expanded the literature to using LLMs for external model creation which can be run locally to forecast with larger datasets and benefit from further improvement by human forecasters. The experiment has additionally highlighted a way to improve the benchmarking standards of LLM forecasting abilities by utilising a large set of easily adaptable benchmark results and using forecasting competitions to get a comparable human performance measure to test the LLMs against.

There is also the research gap in using LLMs for various domain-specific forecasting applications. This study tries to fill that gap in retail forecasting by gathering a clear overview of the current most advanced LLM models and conducting an experiment to test their forecasting abilities. Furthermore, this paper proposed a theoretical framework for adapting LLMs into retail forecasting. This framework highlights the key elements of forecasting that could eventually be automated with LLMs and helps assess the progress of automating forecasting in the future. The experiment conducted in this paper can be used as an example on how to further evaluate LLMs' abilities of forecasting in future studies and how to compare the proficiency of different LLMs. With small adjustments, this type of experiment can also be applied to testing other aspects and skills of various LLMs.

5.4 Limitations

This study was made using public datasets with a large volume of available online discussions and repositories focused on creating solutions for this forecasting problem. It is likely that at least some of these solutions are part of the LLMs' training data and it is unclear what effects this might have had on the performance of the LLMs compared to forecasting with private sales data.

As also highlighted in the literature review on the subject of LLM limitations, prompt engineering has a large effect on the performance and output of LLMs. This study did not delve too extensively into creating the perfect prompt for the LLMs, instead only making it good enough for the LLM to understand and effectively perform the task on a stable basis. More extensive prompt engineering might considerably affect the performance of LLMs' forecasting accuracy and learning consistency.

Two LLMs were chosen for this study based on the results of previous literature and online benchmarks of different LLMs. Older papers did not use the same models that are available now and online benchmarks test LLMs on specific tasks to create their rankings. This does not guarantee that the models chosen really were the best LLMs for creating forecasting models. The chosen LLMs were asked to create and improve their forecasts to be as accurate as possible within 10 or 20 attempts. Since 20-attempt series performed worse, it is not clear what might be the optimal number of attempts to give them or whether it would be best not to clarify how many times they can improve in the first place.

5.5 Future research

Testing LLMs abilities with other datasets, particularly with private data, will help clarify the possible effect of LLMs having access to potential pre-existing solutions. Because companies often do not wish to publish their sales data lightly, this might be best done through an internship research project in a retail company. Repeating this experiment without telling the LLM how many attempts it will have to improve or testing for an optimal number of given attempts is another potential research area to solve a limitation of this study.

Many companies are quickly developing their own LLMs. Future research around half a year or a year later that involves as many LLMs as possible would give a clearer picture on both the speed of advancements during that time gap and on which models are most suitable for retail forecasting. This is a developing field and repeating similar research can yield significantly different results after only a few months of new advancements.

Comparison between human and LLM capabilities in retail forecasting was not comprehensively answered during this research. Future experiments where humans with and without LLMs go head-to-head should provide a better understanding of how useful LLMs will be in a real business context. Another option is creating an experiment with a more hands-on human operator guiding the LLM to perform feature engineering and build hybrid models to evaluate whether those can compete with the best performing models crafted by humans.

6. ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. M. de Visser, who helped me throughout the process of writing this thesis. Thank you for your guidance, enlightening questions, and helpful comments. I also express my gratitude to my brother who planted the initial seed of an idea that led to this thesis topic. I next want to thank my thesis circle, who helped me shape the idea up to a useable research topic.

7. REFERENCES

Abolghasemi, M., Ganbold, O., & Rotaru, K. (2025). Humans vs. large language models: Judgmental forecasting in an era of advanced AI. *International Journal of Forecasting*, *41*(2), 631-648.

Alice, K., & Srivastava, S. A. (2023). Sales Forecasting using XGBoost. *Authorea Preprints*.

Almgerbi, M., De Mauro, A., Kahlawi, A., & Poggioni, V. (2021). A Systematic Review of Data Analytics Job Requirements and Online-Courses. *Journal of Computer Information Systems*, *62*, 1-13. https://doi.org/10.1080/08874417.2021.1971579

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, *8*(3), 147-156.

https://doi.org/https://doi.org/10.1016/S0969-6989(00)00011-4

Amir, W., Soom, A. B. M., Jasin, A. M., Ismail, J., Asmat, A., & Rahman, R. A. (2023). Sales Forecasting Using Convolution Neural Network. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *30*(3), 290-301.

Armstrong, J. S. (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners. *International Series in Operations Research & Management Science*. https://doi.org/10.1007/978-0-306-47630-3

Artificial Analysis. (2025). *Independent analysis of AI models and API providers*. https://artificialanalysis.ai/

Bloom, D. E., Canning, D., & Fink, G. (2010). Implications of population ageing for economic growth [Article]. *Oxford Review of Economic Policy*, *26*(4), 583-612. https://doi.org/10.1093/oxrep/grq038

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., & Brunskill, E. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Box, G., & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Boyko, J., Cohen, J., Fox, N., Veiga, M. H., Li, J. I., Liu, J., Modenesi, B., Rauch, A. H., Reid, K. N., & Tribedi, S. (2023). An interdisciplinary outlook on large language models for scientific research. *arXiv preprint arXiv:2311.04929*.

Bucaioni, A., Ekedahl, H., Helander, V., & Nguyen, P. T. (2024). Programming with ChatGPT: How far can we go? *Machine Learning with Applications*, *15*, 100526. https://doi.org/https://doi.org/10.1016/j.mlwa.2024.100526

Cheng, Y., Zeng, Y., & Zou, J. (2024). Harnessing ChatGPT for predictive financial factor generation: A new frontier in financial analysis and forecasting. *The British Accounting Review*, 101507. https://doi.org/https://doi.org/10.1016/j.bar.2024.101507

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283-1318. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.06.004

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201-224. https://doi.org/https://doi.org/10.1002/berj.4069

Gai, T. (2025, 2025//). Machine Learning Approaches for Accurate Sales Forecasting in Supermarket Chains. Advances in Computational Vision and Robotics, Cham.

Gandhi, M. A., Maharram, V. K., Raja, G., Sellapaandi, S., Rathor, K., & Singh, K. (2023). A novel method for exploring the store sales forecasting using fuzzy Pruning LS-SVM approach. 2023 2nd International Conference on Edge Computing and Applications (ICECAA),

Geurts, M. D., & Kelly, J. P. (1986). Forecasting retail sales using alternative models. *International Journal of Forecasting*, *2*(3), 261-272. https://doi.org/https://doi.org/10.1016/0169-2070(86)90046-4

Ghasemloo, M., & Moradi, A. (2025). Informed Forecasting: Leveraging Auxiliary Knowledge to Boost LLM Performance on Time Series Forecasting. *arXiv preprint arXiv:2505.10213*.

Graefe, A., Green, K. C., & Armstrong, J. S. (2013). Forecasting. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of Operations Research and Management Science* (pp. 593-604). Springer US. https://doi.org/10.1007/978-1-4419-1153-7_357

Gruver, N., Finzi, M., Qiu, S., & Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, *36*, 19622-19635.

Hasan, M. (2024). Addressing seasonality and trend detection in predictive sales forecasting: A machine learning perspective. *Journal of Business and Management Studies*, *6*(2), 100-109.

Hassani, H., & Silva, E. S. (2023). The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing*, *7*(2), 62. https://www.mdpi.com/2504-2289/7/2/62

Howard, A., inversion, Makridakis, S., & vangelis. (2020). *M5 Forecasting - Accuracy*. Kaggle. https://kaggle.com/competitions/m5-forecasting-accuracy

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420-1438. https://doi.org/https://doi.org/10.1016/j.ijforecast.2020.02.005

In, Y., & Jung, J.-Y. (2022). Simple averaging of direct and recursive forecasts via partial pooling using machine learning. *International Journal of Forecasting*, *38*(4), 1386-1399. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.11.007

Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., & Pan, S. (2023). Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Johnson, M., Jain, R., Brennan-Tonetta, P., Swartz, E., Silver, D., Paolini, J., Mamonov, S., & Hill, C. (2021). Impact of Big Data and Artificial Intelligence on Industry: Developing a Workforce Roadmap for a Data Driven Economy [Article]. *Global Journal of Flexible Systems Management*, *22*(3), 197-217. https://doi.org/10.1007/s40171-021-00272-y

Laban, P., Kryściński, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., & Wu, C.-S. (2023). Llms as factual reasoners: Insights from existing benchmarks and beyond. *arXiv preprint arXiv:2305.14540*.

Li, G., Yuan, C., Kamarthi, S., Moghaddam, M., & Jin, X. (2021). Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis [Article]. *Journal of Manufacturing Systems*, *60*, 692-706. https://doi.org/10.1016/j.jmsy.2021.07.007

Lin, J., Lai, S., Yu, H., Liang, R., & Yen, J. (2025). ChatGPT based credit rating and default forecasting. *Journal of Data, Information and Management*, *7*(1), 69-92. https://doi.org/10.1007/s42488-025-00143-6

Liu, J., Chen, L., Luo, R., & Zhu, J. (2023). A combination model based on multi-angle feature extraction and sentiment analysis: Application to EVs sales forecasting. *Expert Systems with Applications*, *224*, 119986.

Liu, X., Wu, Z., Wu, X., Lu, P., Chang, K.-W., & Feng, Y. (2024). Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv preprint arXiv:2402.17644*.

Lopez-Lira, A., & Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Loureiro, A. L. D., Miguéis, V. L., & da Silva, L. F. M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81-93. https://doi.org/https://doi.org/10.1016/j.dss.2018.08.010

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111-128. https://doi.org/https://doi.org/10.1016/j.ejor.2020.05.038

MacKenzie, I., Meyer, C., & Noble, S. (2013). *How retailers can keep up with consumers*. McKinsey. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). M5 accuracy competition: Results, findings, and conclusions [Article]. *International Journal of Forecasting*, *38*(4), 1346-1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*(4), 1325-1336. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.07.007

Mediavilla, M. A., Dietrich, F., & Palm, D. (2022). Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. *Procedia CIRP*, *107*, 1126-1131. https://doi.org/https://doi.org/10.1016/j.procir.2022.05.119

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey, 2024. *arXiv preprint arXiv:2402.06196*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Noguer I Alonso, M., & Pereira Franklin, R. (2024). *Large Language Models for Financial Time Series Forecasting*. Elsevier BV. https://dx.doi.org/10.2139/ssrn.4988022

OpenAI Code Interpreter. (2025). OpenAI. https://platform.openai.com/docs/assistants/tools/code-interpreter

OpenAI Models. (2025). OpenAI. https://platform.openai.com/docs/models

Paleka, D., Goel, S., Geiping, J., & Tramèr, F. (2025). Pitfalls in Evaluating Language Model Forecasters. *arXiv preprint arXiv:2506.00723*.

Park, J., Lee, H., Lee, D., Gwak, D., & Choo, J. (2025). Revisiting LLMs as Zero-Shot Time-Series Forecasters: Small Noise Can Break Large Models. *arXiv preprint arXiv:2506.00457*.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A.,…Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705-871. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.11.001

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*(4), 822-829. https://doi.org/https://doi.org/10.1016/j.ijforecast.2018.03.001

Singla, A., Sukharevsky, A., Yee, L., & Chui, M. (2024). *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. McKinsey. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024). Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.

Tang, H., Zhang, C., Jin, M., Yu, Q., Wang, Z., Jin, X., Zhang, Y., & Du, M. (2025). Time series forecasting with llms: Understanding and enhancing model capabilities. *ACM SIGKDD Explorations Newsletter*, *26*(2), 109-118.

Tran, D. T., Huh, J.-H., & Kim, J.-H. (2023). Building a Lucy hybrid model for grocery sales forecasting based on time series. *The Journal of Supercomputing*, *79*(4), 4048-4083.

Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., & Senge, R. (2022). Classification-based model selection in retail demand forecasting. *International Journal of Forecasting*, *38*(1), 209-223. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.05.010

Vals AI. (2025). *MMLU Pro Benchmark*. https://www.vals.ai/benchmarks/mmlu_pro-04-04-2025

Van Calster, T., Baesens, B., & Lemahieu, W. (2017). ProfARIMA: A profit-driven order identification algorithm for ARIMA models in sales forecasting. *Applied Soft Computing*, *60*, 775-785. https://doi.org/https://doi.org/10.1016/j.asoc.2017.02.011

van Donselaar, K., van Woensel, T., Broekmeulen, R., & Fransoo, J. (2006). Inventory control of perishables in supermarkets. *International Journal of Production Economics*, *104*(2), 462-472. https://doi.org/https://doi.org/10.1016/j.ijpe.2004.10.019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT [Article]. *Entrepreneurial Business and Economics Review*, *11*(2), 7-30. https://doi.org/10.15678/EBER.2023.110201

Wang, X., Ryoo, J. H., Bendle, N., & Kopalle, P. K. (2021). The role of machine learning analytics and metrics in retailing research. *Journal of Retailing*, *97*(4), 658-675.

https://doi.org/https://doi.org/10.1016/j.jretai.2020.12.001

Wellens, A. P., Boute, R. N., & Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research*, *314*(2), 523-539. https://doi.org/https://doi.org/10.1016/j.ejor.2023.10.039

Xexéo, G., Braida, F., Parreiras, M., & Xavier, P. (2024). The economic implications of large language model selection on earnings and return on investment: A decision theoretic model. *arXiv preprint arXiv:2405.17637*.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175. https://doi.org/https://doi.org/10.1016/S0925-2312(01)00702-0

Zhang, H., Arvin, C., Efimov, D., Mahoney, M. W., Perrault-Joncas, D., Ramasubramanian, S., Wilson, A. G., & Wolff, M. (2024). LLMForecaster: Improving seasonal event forecasts with unstructured textual data. *arXiv preprint arXiv:2412.02525*.

Zhang, H., Dong, Y., Xiao, C., & Oyamada, M. (2023). Large language models as data preprocessors. *arXiv preprint arXiv:2308.16361*.

APPENDIX A

ChatGPT prompt:

You are a retail forecasting expert. You need to forecast the unit sales of various products at stores in various locations for a 28-day period.

You will be evaluated based on Weighted Root Mean Squared Scaled Error (RMSSE).

You have [10 or 20] attempts to improve your forecast model to be as accurate as possible, after each attempt you will be given your RMSSE score.

The dataset includes 4 csv files:

sell_prices.csv – Contains information about the price of the products sold per store and date.

calendar.csv – Contains information about the dates on which the products are sold, including holidays and special events.

sales_train.csv – Contains the historical daily unit sales data per product and store [d_1 – d_1913]

sample_submission.csv – The correct format for submissions. Each row contains an id that is a concatenation of an item_id and a store_id. You are predicting 28 forecast days (F1 – F28) of items sold for each row. This corresponds to d_1914 – d_1941.

Answer with a Python script that outputs the forecast into submission_[test]_[name of AI].csv. Datasets will be in the same working directory as the script.

I will run your code and return results. Do not ask me for which improvements to make, do what you think will be the most accurate.

Included are sample csv files of these datasets.

Gemini prompt:

You are a retail forecasting expert. You need to forecast the unit sales of various products at stores in various locations for a 28-day period.

You will be evaluated based on Weighted Root Mean Squared Scaled Error (RMSSE).

You have [10 or 20] attempts to improve your forecast model to be as accurate as possible, after each attempt you will be given your RMSSE score.

The dataset includes 4 csv files:

sell_prices.csv – Contains information about the price of the products sold per store and date.

calendar.csv – Contains information about the dates on which the products are sold, including holidays and special events.

sales_train.csv – Contains the historical daily unit sales data per product and store [d_1 – d_1913]

sample_submission.csv – The correct format for submissions. Each row contains an id that is a concatenation of an item_id and a store_id. You are predicting 28 forecast days (F1 – F28) of items sold for each row. This corresponds to d_1914 – d_1941.

Answer with a Python script that outputs the forecast into submission_[test]_ [name of AI].csv. Datasets will be in the same working directory as the script.

I will run your code and return results. Do not ask me for which improvements to make, do what you think will be the most accurate.

Included are sample txt files of these datasets.

Next rounds:

Score is: [WRMSSE]