Insights for Auto Port Ranking with LangChain-LLMs

Zehao Qian zehao.qian.cn@gmail.com

January 5, 2024

1 Task Schedule Model

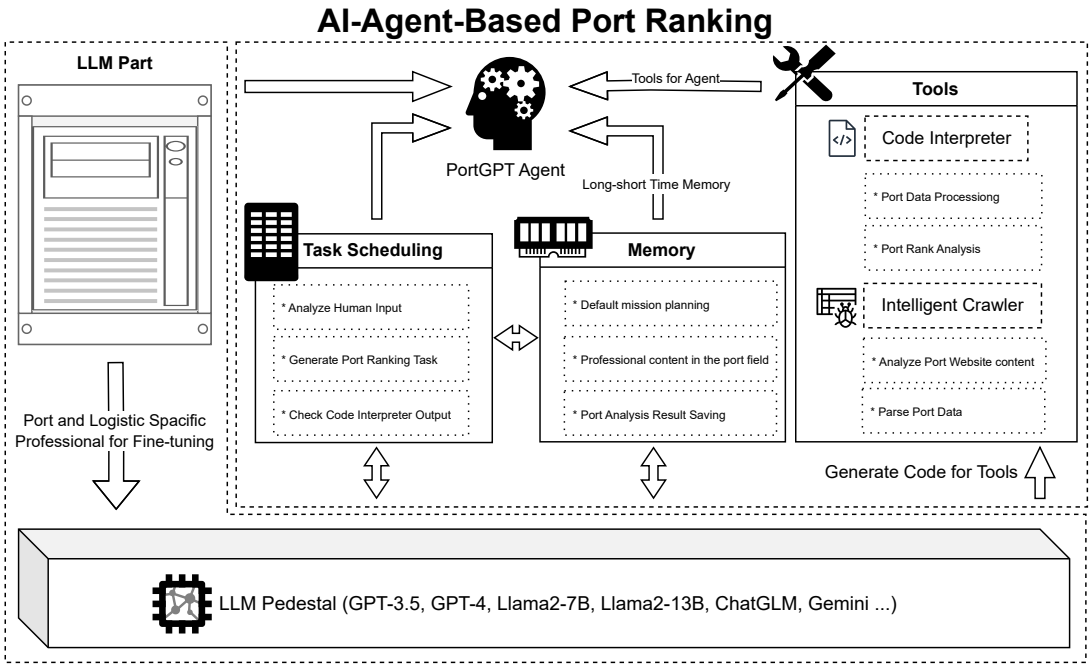

Figure 1: PortGPT Conceptual Architecture

2 LLM Fine-tuning and Prompt Engineering for Auto Port Ranking

3 Port Data Acquisition

Type	Advantages	Disadvantages

Type	Advantages	Disadvantages
Traditional Web Crawlers	- Fast speed - Low resource consumption - Simple and easy to use	- Poor flexibility - Easily blocked by anti-crawling mechanisms
Selenium-based Crawlers	- Highly flexible - Can handle complex scenarios	- Slower speed - High resource consumption - High complexity
LLM-based Intelligent Crawlers	- Strong understanding ability - Strong adaptability - Capable of complex interactions	- Potential speed limitations - Resource-intensive - Challenges in accuracy

4 Dealing with Outliers and Missing Values

4.1 Outliers

4.2 Missing Values

In modern port management and Operation Research, the integrity and accuracy of data is very important. Effective data analysis can not only reveal the current state of port operation, but also predict future trends and potential problems. However, data loss is a common problem in data collection, which can be caused by a number of factors, including technical failures, recording errors or delayed information updates. The purpose of this paper is to explore and solve this challenge in order to enhance the understanding and application of port data.

Firstly, the data collected from the port were examined to identify the patterns and possible causes of missing data. On this basis, we will explore various strategies to deal with missing data to ensure the accuracy and reliability of the analysis results. We will focus in particular on two major approaches: Imputation and Deletion. The interpolation method aims at estimating missing values and filling data gaps with these estimates, while the deletion method is to remove data records containing missing values.

Delete or Impute? When deciding whether to delete or impute missing data, consider the proportion and pattern of missingness, the nature of the data, the requirements of the analysis method, and the purpose of the study. Deletion is suitable for small proportions of randomly missing data and when it won't introduce bias, while imputation is preferred for large proportions of missing data, non-random missingness, to maintain dataset size and integrity, and to retain critical information. The decision should balance the characteristics of the data, the reasons for missingness, and the objectives of the analysis. [1]

4.2.1 Tranditional Missing Data Imputation Methods

Mean/Median/Mode Imputation [2] is a common and simple method where all missing values are replaced with the mean, median, or mode of the column. Although this approach is easy and fast, it has its drawbacks. It can skew the statistical nature of the data, underestimate variance, and distort histograms. This method is generally advisable only when data is missing completely at random (MCAR) or missing at random (MAR), but it is not suitable if data is missing not at random (MNAR)

Multiple Imputation [3] [4] involves creating multiple copies of the dataset, where missing values are replaced with imputed values based on their predictive distribution from observed data. This method uses standard statistical methods to fit models to each imputed dataset, and then averages the results to provide overall estimated associations. Multiple imputation, based on a Bayesian approach, accounts for all uncertainty in predicting missing values by introducing appropriate variability into the imputed values. It is essential in multiple imputation to model the distribution of each variable with missing values accurately. However, pitfalls can occur, such as omitting the outcome variable from the imputation process or dealing incorrectly with non-normally distributed variables. The validity of results from multiple imputation depends on careful and appropriate modeling.

KNN and linear regression imputing [5] [6] [7] [8]

4.2.2 Imputing Missing Data with Deep Learning and LLMs

The great potential of deep learning in imputation of missing values in data. In the field of port management, these methods can be used to process and analyze large amounts of complex data, such as cargo flows, vessel dynamics, weather conditions, and port operation efficiency. Data imputation using deep learning not only improves data completeness and accuracy, but also helps predict and optimize port operations, thereby improving efficiency and safety. [9] [10] [11] [12]

In order to overcome the challenge of missing values in port data, this study proposes a method to predict and interpolate these missing values using a pre-trained model Xi machine learning. We first use existing port data, including multi-dimensional information such as cargo throughput, vessel arrival frequency, weather conditions, and port facility usage, to train our machine Xi model. This data has been cleaned and preprocessed to ensure the quality of the data fed into the model. We chose a machine Xi model suitable for working with time series data, such as long short-term memory networks (LSTMs) or gated recurrent units (GRUs), because port data often have significant time-dependent and seasonal characteristics. The training of the model was performed on a rich historical dataset containing complete and missing instances of data collected from normal operations. In this way, the model learns Xi complex patterns and associations in the data, which can be used to predict missing values.

References

[1] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Medical Research Methodology*, 17(1):162, 2017. ID: Jakobsen2017.

[2] Ambar Kleinbort. Data imputation: Beyond mean, median, and mode. *Open Data Science*, 2020.

[3] Jonathan A C Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009.

[4] Shahidul Islam Khan and Abu Sayed Md Latiful Hoque. Sice: an improved missing data imputation technique. *Journal of Big Data*, 7(1):37, 2020. ID: Khan2020.

[5] Turki Aljrees. Improving prediction of cervical cancer using knn imputer and multi-model ensemble learning. *PLOS ONE*, 19(1):1–24, 01 2024.

[6] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology*, 8:1–19, 01 2017.

[7] Della Murti, Utomo Pujianto, Aji Wibawa, and Muhammad Akbar. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology*, pages 83–88, 10 2019.

[8] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. A survey on missing data in machine learning. *Journal of Big Data*, 8(1):140, 2021. ID: Emmanuel2021.

[9] Yinan Mei, Shaoxu Song, Chenguang Fang, Haifeng Yang, Jingyun Fang, and Jiang Long. Capturing semantics for imputation with pre-trained language models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 61–72, 2021.

[10] Zhenhua Wang, Olanrewaju Akande, Jason Poulos, and Fan Li. Are deep learning models superior for missing data imputation in large surveys? evidence from an empirical comparison, 2022.

[11] Jangho Park, Juliane Muller, Bhavna Arora, Boris Faybishenko, Gilberto Pastorello, Charuleka Varadharajan, Reetik Sahu, and Deborah Agarwal. Long-term missing value imputation for time series data using deep neural networks, 2022.

[12] Ramiro D. Camino, Christian A. Hammerschmidt, and Radu State. Improving missing data imputation with deep generative models, 2019.