POI GPT : Extracting POI Information from Social Media Text Data

Hyebin Kim, Sugie Lee*

M.S. student, Dept. of Urban Planning and Engineering, Hanyang University, Seoul, Korea, khb723@hanyang.ac.kr
Dept. of Urban Planning and Engineering, Hanyang University, Seoul, Korea, sugielee@hanyang.ac.kr

Keywords: Large Language Model(LLM), Point of Interest(POI), ChatGPT, Social Media, Named Entity Recognition(NER)

Abstract:

Point of Interest (POI) is an important intermediary connecting geo data and text data in smart cities, widely used to extract and identify urban functional areas. While computer uses numerical coordinates, human uses places names or addresses to find location, leading to spatial-semantic ambiguities. However, traditional methods of extracting POIs are time-consuming and costly, and has the limitation of the lack of integration of functionalities such as information extraction(IE), information searching. Also, previous models have low accessibility and high barriers for users. With the advent of Large Language Models(LLMs) we propose a method that connects LLM models and POI information based on social media text data. By employing two steps, named entities recognition(NER) and POI information searching, we introduce POI GPT, the specialized model for providing precise location of POIs in social media text data. We compared its results with those obtained by human experts, NER model and zero-shot prompts. The findings show that our model effectively found the POI and precise location from social media text data. In result, POI GPT is a effective model that solves the existing POI extraction problems. We provide new extraction technique of POI GPT which is a new paradigm in traditional urban research methodologies and be actively utilized in urban studies in the future.

1. INTRODUCTION

Point of Interest (POI) is an important intermediary connecting geo data and text data in smart cities, widely used to extract and identify urban functional areas. Most human activities occur at POIs, and their lightweight and accessible nature allows for effective and accurate understanding of trends in the relationship between humans and cities (Gao et al., 2017; Zhang et al., 2021). Typically, POI is provided as georeferenced information in dataset and place name, address is included in POI information. Georeferencing by place names (known as toponyms) is the most common way of associating textual information with a geographic location(Kuai et al., 2020). While computers use numerical coordinates(latitude-longitude) to recognize places, people use place names or addresses to find it. This, however, makes confusion between finding exact location since coordinates represent only one location but place names do not. Some places that share same name, such as franchise, may be commonly identified with several locations(Zhu et al., 2016), leading to spatial-semantic ambiguities. In an effort to reduce this gap, numerous studies have been conducted to extract accurate place names from text data, for example NER. However, traditional methods of extracting POIs and searching each information are time-consuming and costly, and extracting exact POIs from text is very difficult unless the text is geotagged.

Meanwhile, the accumulation of various big data from social media has led to the accumulation of text data containing the opinions and experiences of urban residents. This facilitates the analysis of urban residents' perceptions and urban space utilization. Big data from social media has been widely used for urban vitality, epidemic detection, satisfaction analysis (Park et al., 2021; Wakamiya et al., 2018; Kwon & Lee, 2023), etc. Especially since the outbreak of COVID-19, non-face-to-face online platforms have been activated and people are actively communicating online, making social media data-based analysis seem useful (Kim, 2023). Due to the sharp increase in online activities following COVID-19, there is a need to use social media text data to analyse residents' experiences and capture urban activities.

On the other hand, with the advent of Generative Pre-trained Transformer-based Large Language Models (LLMs) in recent years, the field of Natural Language Processing (NLP) has rapidly advanced. Notable models include GPT-4, LLaMA, and PaLM. These LLMs, with their large language models and training datasets, have shown powerful and impressive performance in understanding and generating meaning, as well as in some complex tasks. Especially, OpenAI recently released 'GPTs' which provides custom ChatGPT creation. It supports knowledge files, web browsing, API use, and instructions for creating custom ChatGPTs (OpenAI, 2023). This demonstrates the potential for tools that not only extract POI place names but also provide information about the POIs. Despite of amazing performance and potential of ChatGPT, there are still several critical limitations that impede its reliability and broad adoption(Fu et al., 2024). First, ChatGPT's performance compared with humans and other NLP techniques, is currently unknown. Second, ChatGPT's performance is highly sensitive on how prompts are crafted.

Accordingly, this study utilizes social media data with text geotagging and narrows the gap between spatial location and text data containing place names. We developed a model called POI GPT, using ChatGPT to extract and find additional information of POIs. The model utilizes OpenAI's ChatGPT, which has recently gained prominence for its exceptional performance and in particular, we chosen GPTs to provide a more convenient user interface. By leveraging ChatGPT's inherent understanding, learning, and generative capabilities, POI GPT can overcome the limitations of existing POI

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-113-2024

extraction methodologies through a method that extracts POIs from social media text data. Specifically, POI GPT is developed as a new framework that can autonomously extract the places visited by the text authors and derive addresses and POI classifications by inputting natural language text data without any preprocessing step. That is, POI GPT can understand the places visited by the authors based solely on the input natural language text data and output appropriate addresses and POI classifications based on defined actions and knowledge.

The goal of our paper is to present a new GPT-based model, POI GPT, that accurately extracts POI and find information of it from raw, unprocessed social media text data. Also, we aim to verify whether such LLM models are capable of understanding text and addressing the existing POI extraction problems by comparing its results with expert's results, ChatGPT's zero-shot results and conventional NLP results. Our model is available in GPT store, https://chat.openai.com/g/g-KCAVctWUh-poigpt.

2. LITERATURE REVIEW

2.1 Information Extraction

Information Extraction(IE) is the task of automatically extracting structured information from unstructured text data. IE techniques are intended to test how well a machine can understand the text written in natural language(Chuang, 2016). Previous approaches to the IE categorized into three types by web pages' structure: structured, semi-structured and unstructured(Zhang et al., 2010). First, structured approach is about structured webpage which has format of predefined and strict. This kind of information can be easily be extracted using frame model. Second, semi-structured approach is an intermediate between structured format and unstructured text. This approach helps disambiguation on free texts. Lastly, unstructured approach is about unstructured webpage with free text of natural language. Generally used natural language techniques and rules are typically based on patterns involving syntactic relations between words, part-of-speech, and phrases or named entities(Downey et al., 2007).

There are various kinds of IE tasks such as postal address extraction, named entity recognition(NER). Postal address extraction is task about extracting complete addresses based on certain geographic terms(Sanderson & Kohler, 2004). NER is another major task in IE that seeks text's location and classify in pre-defined categories, such as location, personal name, organization(Sun et al., 2018). Early NER used rule-based approaches, such as lexical, contextual, morphological lexicon rules, but no longer mainstream approaches since built with less of no labelled corpus have been allowed. Such technological advancements have brought machine learning-based NER to the forefront. As the necessity for large volumes of language knowledge, in the form of lexicons and feature engineering, to achieve higher performance has increased, deep learning-based NER has concurrently risen to prominence(Sun et al., 2018).

2.2 LLM Application in Urban Studies

Given rapid development of LLMs recently, urban studies using LLM has also increased. Previous research covers a broad range of topics, such as urban planning(Fu et al., 2024; Zhu et al., 2024; Zhou et al., 2024), POI recommendation(Feng et al., 2024; Yin et al., 2023), NER(Yao et al., 2024; Hu et al., 2023).

LLMs have primarily been used in the field of urban planning, extensively for purposes such as evaluation and planning support system(PSS) across a broad spectrum. Fu et al.(2024) analyzed public feedback data in urban planning by employing ChatGPT and standard NLP techniques. As compared results shown, ChatGPT had few limitations but presented a transformative opportunity for planners especially on public feedback big data. Others proposed planning multi-agent collaboration framework by applying LLMs in original urban for participatory urban planning (Zhou et al., 2024) or by advanced tooling capabilities(Zhu et al., 2024). In terms of model proposing, Feng et al. (2024) developed POI recommendation system based on ChatGPT prompting strategies based on considering the geographical influence and sequential transitions. As mentioned earlier, with the increase in machine learning and deep learning approaches to NER models, applications of LLMs have also been active. Various studies have identified place names based on Twitter messages in disaster situations(Hu et al., 2023) and from drone delivery addresses(Yao et al., 2024). All these studies utilized NER to convert raw user addresses into precise locations using language models. In the same vein, our study aims to utilize LLMs to recognize POI place names from social media text data, converting users' raw natural language place names into precise addresses.

Although interest in LLM applications is rapidly growing, there are two major gaps in this emerging field within the information extraction. The first gap identified is the lack of integration of functionalities in the use of LLMs. While fields such as Information Extraction(IE) and appropriate information retrieval have been developing independently, attempts to combine these have been scarce. However, with the advent of LLMs and advances in custom GPT development, it is now possible to merge traditional IE capabilities with LLM's information retrieval and text generation skills. This integration allows for the creation of models that can perform POI extraction and POI information search, and explain the findings in a manner understandable to users. Second gap is the accessibility of NLP techniques for users. Previous models have been critiqued for being technology-driven without considering the user experience(Geertman and Stillwell, 2020). Recognizing that a model becomes meaningful when used by individuals, there is a pressing need for NLP models that consider user experience to improve accessibility. Consequently, this study employs GPTs using the ChatGPT Interface to develop a more user-friendly model, thereby overcoming the accessibility limitations that have been challenging for users to navigate.

3. METHOD

The objective of this study is to benchmark our POI GPT against experts, NER model and zero-prompt model with Chat GPT to accomplish extraction of POI information from blog text data.

3.1 Dataset

In this study, we selected a corpus of crowdsourced travel blog, *travelblog.org*. It is a worldwide popular travel blogging website and in each post, author describes their travel experiences(Ballatore and Adams, 2015; Davydova, 2012). The main locations are organized by hierarchical vocabulary in order of continent, country, administrative regions. Blog dates, number of words are also provided.

To collect the data, we used web crawling in Python with Selenium package and BeautifulSoup library. In total, we collected 4,290 travel blogs text data from 2019 to 2023. For the purpose of including only text data that contains POIs within, a preprocessing procedure was undertaken. This process involves retaining data with a word count exceeding 100 and below 300. Also to retain only reviews written in English, we employed the langdetect library in Python to remove reviews that were not in English. Through this procedure, total 63 travel blogs remained in dataset. Finally, we examined the data distribution by continent to check data's continent distribution (Figure 1). Example of acquired travel blog text data is provide in Table 1.

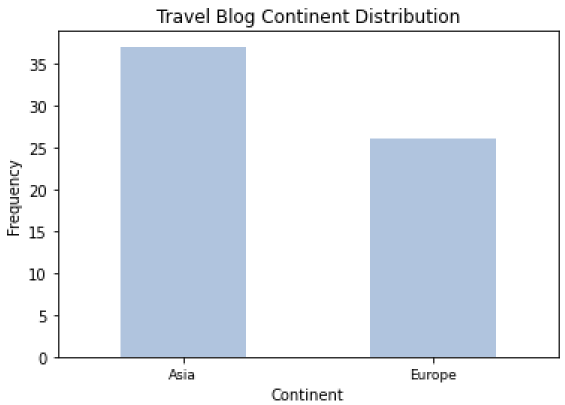

Figure 1. Travel blog continent distribution result.

3.2 GPT model

Our model is done through 'GPTs', which provides custom ChatGPT creation. publicly shared services in ChatGPT for customers and allows easy integration with users' data. Despite the existence of the GPT-4 APIs, we utilized GPTs over the API due to their provision of a GPT interface. Since GPTs offers a more user-friendly interface, it enhances applicability. This approach is significant in that it overcomes the limitations of supply-driven(technology-driven) urban technologies (Geertman and Stillwell, 2020), such as PSS, by offering a demand-driven, user-friendly environment adopting GPTs. Specifically, GPTs offers functionalities such as support for knowledge files, web browsing capabilities, API usage, and instructions for creating custom ChatGPTs to specific purposes (OpenAI, 2023).

The model in this study is composed of two steps. First, POI GPT performs Named Entity Recognition(NER) to identify POIs within text data. Second, it finds information for each POI, such as POI categorization, longitude, and latitude. Ultimately, the identified POIs and their associated information are compiled and presented to the user. Figure 2 provides an overview of our model.

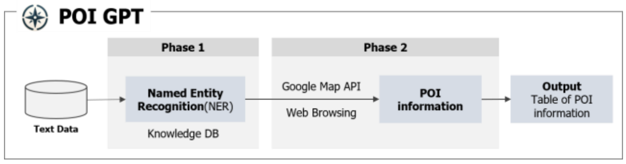

Figure 2. Framework of POI GPT.

3.2.1 POI extraction : Large language models can extract key information from various textual sources, including travel blog text data, public comments, and academic studies(Zhu et al., 2024). This process is about finding the POI names from the travel blog text data.

3.2.2 POI information : GPTs also offers web browsing and utilizing API by actions. While our model uses web browsing to access in public big data, to enhance accuracy and reliability, we employed GoogleMap API from SerpAPI to find exact information of address, POI category and coordinates. and then to answer the same questions as those asked in the zero-shot prompts.

3.3 Evaluation

This comparison allows us to find ChatGPT's effectiveness and develop prompting guidelines, providing insights for to effectively use this tool in analysing social media text data. Since this model has two phases, NER and POI information search, we also evaluate in two phases. First, we evaluate about our model's NER performance with expert's, NER model's, and zero-shot prompting's results. Second, we will evaluate the accuracy of searched POI information with expert's and zero-shot prompting's results. Since NER model does not provides information of POI, it cannot be used for evaluating the entire process of this model and can only be utilized for assessing the first phase, NER part. We employed accuracy, precision, recall and F1-score for the evaluation metrics.

3.3.1 Expert : To validate and assess our POI GPT's POI extraction and searching information performance, we asked two POI experts to extract POI from travel blog text data and search information about each POI. This process mirrors the operations conducted by our model, POI GPT.

Dates	Country	Contents
2023-11-06	United-States	We have recently returned from an amazing trip to Oahu and Big IslandWe booked our flights with British Airways from Manchester via Heathrow and Seattle. … .A very Comfortable room with a kitchenette as we knew Hawaii was expensive!
2024-01-23	Spain	It was nice coming back to Tarifa. It would be a nice place to spend the wi ter.I stayed two nights to give me a chance to cycle to La Janda. I was hoping to see back with get kites again. No luck, but I did get decent views of red legged partridge.
2019-04-01	United-Arab-Emirates	The Burj Khalifa is currently the world's tallest building at over 800 metres. We too the brave step to firstly spend money and then to soar to level 148 for a sunset view over Dubai. … only to be met by an influx of people coming back from the dancing fountains.

Table 1. Example of travel blog text data acquired

3.3.2 Named Entity Recognition(NER) : NER is the task to identify mentions of rigid designators from text belonging to predefined semantic types such as person, location, organization etc(Li et al., 2020). While NER is a traditional task of NLP, we adopted NER model to evaluate our model's NER performance. We used WikiNEuRal-Multilingual NER model from Tedeschi et al.(2021). WikiNEuRal is neural BERT-based model and shown substantial performance improvements over previous state-of-the-art approaches. To solely identify POIs, we extracted only the 'LOC(Location)' tags from among the various tags available.

3.3.3 Zero-shot prompting : We initially used zero-shot prompts with ChatGPT, which means giving prompts without specific instructions or a data set. This is the most widely used method(Fu et al., 2024) and can also represent the generalization of ChatGPT. We asked ChatGPT to "extract the POI in this text and find each POI's address, latitude, longitude, POI classification." The outputs were typically in a consistent format.

4. RESULTS

4.1 Result Interface

Our model POI GPT implements an interface using GPTs. An example of our model implementation is shown in Figure 3. Specifically, when the raw text data of a travel blog is input as a query, it automatically undergoes preprocessing, and if necessary, calls an API to provide the Place Name, Address, POI classification, Latitude, and Longitude in a table format. Additionally, our model mentions the Lab's website at the end of the response.

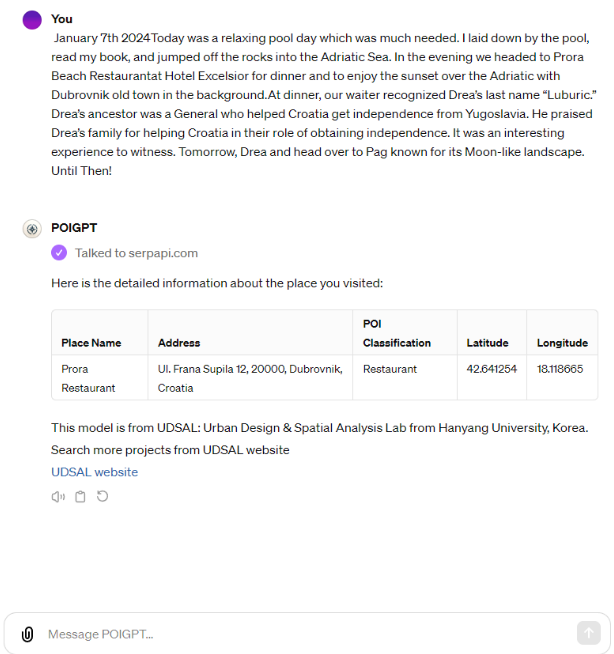

Figure 3. Example of POI GPT implementation.

POI Name	Address	POI Category	Latitude	Longitude
Umeda Sky Building	1 Chome-1-88 Oyodonaka, Kita Ward, Osaka, 531-6023, Japan	Architectural Landmark	34.7055	135.4907
Osaka Castle	1-1 Osakajo, Chuo Ward, Osaka, 540-0002, Japan	Historical Landmark	34.6873	135.526
Nagasaki Atomic Bomb Museum	7-8 Hirano-machi, Nagasaki, 852-8117, Japan	Museum	32.7732	129.8636

Table 2. Example outcomes of POI GPT

4.2 Evaluation on extracting POI

Initially, we conducted an evaluation of POI extraction. Our results, as sown in Table 3, confirm that our model, POI GPT, exhibited the best performance. The evaluation benchmark was based on Expert's POI extraction, focusing solely on locations that the author of the text actually visited. As a result, NER models that recognized all place names showed the lowest accuracy. This was also due to the extraction of incorrect POIs, often inaccurately recognizing terms like 'North' and 'South'. These are considered limitations of the NER model and highlight the need for utilizing GPT. In the case of zero-shot prompting, our model, POI GPT, showed high accuracy with only a slight difference.

Model	Number of POIs	Accuracy	Precision	Recall	F1-score
Expert	119	-	-	-	-
NER	423	30.52%	23.17%	82.35%	36.16%
Zero-shot	192	77.31%	51.58%	82.35%	63.43%
POI GPT	**188**	**78.51%**	**53.19%**	**84.03%**	**65.14%**

Table 3. Evaluation results of models on extracting POI

4.3 Evaluation on POI information accuracy

Next, we evaluated the second phase, which involves the collection of information for each POI. The results are presented in Table 4 and indicate that our model, POI GPT, consistently demonstrated the best performance. The evaluation benchmark was based on POI information directly collected by an expert using Google Maps, including address, latitude, longitude, and POI classification. The assessment was conducted only on POIs present in all three models, total 72 POIs. We verified the accuracy of information among the models through manual review. Although the differences were slight, our model, POI GPT, showed higher accuracy compared to zero-prompting. The overall lower accuracy can be attributed to the small sample size and challenges in clearly identifying addresses when POIs were closely located or involved broad districts.

Model	Accuracy
Zero-shot	59.72%
POI GPT	**62.5%**

Table 4. Evaluation results of POI information accuracy

5. DISCUSSION

After reviewing the literature on information extraction and the application of LLMs in urban studies, we identified two major gaps and developed a LLM-based model named POI GPT to overcome them. Our model extracts POIs from raw social media text data and collects location information for each POI, integrating functionalities of existing models while also providing a user-friendly interface. Given the ongoing increase in online activities and the availability of online text big data post-COVID-19, efforts to utilize this resource should continue, aligning with our research's focus on integration and convergence.

This study advances beyond previous research by combining POI extraction and information collection functionalities, demonstrating higher accuracy and usability compared to other models. However, there are limitations as the study used a small sample, and it does not fully represent the entire blog text. Additionally, the text data distribution was limited to Asia and Europe, not considering other continents. Future work should aim to acquire extensive blog data from all continents for further analysis. Moreover, while our study has improved interface and visual-accessibility limitations of previous research by using GPTs with a user-friendly interface, it faces challenges in processing large volumes of data simultaneously and only accessible to paid users of ChatGPT. Future research will endeavor to enhance user accessibility by simultaneously considering APIs.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(*MSIT) (No.2018R1A5A7059549). *Ministry of Science of ICT

References

Chuang, H. M., Chang, C. H., Kao, T. Y., Cheng, C. T., Huang, Y. Y., & Cheong, K. P. (2016). Enabling maps/location searches on mobile devices: Constructing a POI database via focused crawling and information extraction. *International Journal of Geographical Information Science*, 30(7), 1405-1425.

Davydova, E. (2012). The role of Travel Blogs in the process of Travel Planning.

D. Downey,M. Broadhead, and O. Etzioni. Locating Complex Named Entities in Web Text. In Proc. of IJCAI, 2007.

Feng, S., Lyu, H., Chen, C., & Ong, Y. S. (2024). Where to Move Next: Zero-shot Generalization of LLMs for Next POI Recommendation. arXiv preprint arXiv:2404.01855.

Fu, X., Sanchez, T. W., Li, C., & Reu Junqueira, J. (2024). Deciphering Public Voices in the Digital Era: Benchmarking ChatGPT for Analyzing Citizen Feedback in Hamilton, New Zealand. *Journal of the American Planning Association*, 1-14.

Gao, S., Janowicz, K., & Couclelis, H. (2017). Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3), 446-467.

Geertman, S., & Stillwell, J. (2020). Planning support science: Developments and challenges. *Environment and Planning B: Urban Analytics and City Science*, 47(8), 1326-1342.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., ... & Joseph, K. (2023). Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37(11), 2289-2318.

Kuai, X., Guo, R., Zhang, Z., He, B., Zhao, Z., & Guo, H. (2020). Spatial context-based local toponym extraction and Chinese textual address segmentation from urban POI data. *ISPRS International Journal of Geo-Information*, 9(3), 147.

Kweon, J. and Lee, S. 2023. Analysis of Resident's Satisfaction and Its Determining Factors on Residential Environment: Using Zigbang's Apartment Review Bigdata and Deeplearning-based BERT Model. *Journal of the Korean Regional Science Association*, 39(2):47-61.

Kim, Y. (2023). A Study on Corona Depression Using Big Data. *Health and Social Welfare Review*, 43(2), 29-47.

OpenAI. (2023). Introducing GPTs. https://openai.com/blog/introducing-gpts

Park, Y., Kim, M., & Seong, K. (2021). Happy neighborhoods: Investigating neighborhood conditions and sentiments of a shrinking city with Twitter data. *Growth and Change*, 52(1), 539-566.

Sun, P., Yang, X., Zhao, X., & Wang, Z. (2018, November). An overview of named entity recognition. In 2018 International Conference on Asian Language Processing (IALP) (pp. 273-278). IEEE.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021, November). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2521-2533).

Wakamiya, S., Kawai, Y., & Aramaki, E. (2018). Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. *JMIR Public Health and Surveillance*, 4(3), e8627.

Yao, Y., Luo, S., Zhao, H., Deng, G., & Song, L. (2024). Can LLM Substitute Human Labeling? A Case Study of Fine-grained Chinese Address Entity Recognition Dataset for UAV Delivery. arXiv preprint arXiv:2403.06097.

Yin, B., Xie, J., Qin, Y., Ding, Z., Feng, Z., Li, X., & Lin, W. (2023, September). Heterogeneous knowledge fusion: A novel approach for personalized recommendation via llm. In Proceedings of the 17th ACM Conference on Recommender Systems (pp. 599-601).

Zhang, H. P., Mo, Q., & Huang, H. (2010, October). Structured Poi data extraction from Internet news. In 2010 4th

International Universal Communication Symposium (pp. 116-122). IEEE.

Zhang, C., Xu, L., Yan, Z., & Wu, S. (2021). A glove-based poi type embedding model for extracting and identifying urban functional regions. *ISPRS International Journal of Geo-Information*, 10(6), 372.

Zhou, Z., Lin, Y., Jin, D., & Li, Y. (2024). Large Language Model for Participatory Urban Planning. arXiv preprint arXiv:2402.17161.

Zhu, H., Zhang, W., Huang, N., Li, B., Niu, L., Fan, Z., ... & Liu, X. (2024). PlanGPT: Enhancing Urban Planning with Tailored Language Model and Efficient Retrieval. arXiv preprint arXiv:2402.19273.

Zhu, R., Hu, Y., Janowicz, K., & McKenzie, G. (2016). Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*, 20(3), 333-355.