



Contents lists available at ScienceDirect

# Information Processing and Management

journal homepage: [www.elsevier.com/locate/ipm](http://www.elsevier.com/locate/ipm)

## BB-GeoGPT: A framework for learning a large language model for geographic information science

Yifan Zhang<sup>a</sup>, Zhiyun Wang<sup>a</sup>, Zhengting He<sup>a</sup>, Jingxuan Li<sup>a</sup>, Gengchen Mai<sup>b,c</sup>, Jianfeng Lin<sup>d</sup>, Cheng Wei<sup>a</sup>, Wenhao Yu<sup>a,e,\*</sup>

<sup>a</sup> School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430078, China

<sup>b</sup> SEAI Lab, Department of Geography and the Environment, University of Texas at Austin, Austin, 78712, TX, USA

<sup>c</sup> SEAI Lab, Department of Geography, University of Georgia, Athens, 30602, GA, USA

<sup>d</sup> Meituan, Beijing, 100102, China

<sup>e</sup> National Engineering Research Center for Geographic Information System, China University of Geosciences, Wuhan, 430078, China



### ARTICLE INFO

Dataset link: <https://github.com/AGI-GIS/BB-GeoGPT>

**Keywords:**

Large language model  
GIS knowledge corpus  
Domain adaptation  
Self-instruct instructions

### ABSTRACT

Large language models (LLMs) exhibit impressive capabilities across diverse tasks in natural language processing. Nevertheless, challenges arise such as large model parameter size and limited model accessibility through APIs such as ChatGPT and GPT-4, which prohibits the model deployment on mobile devices and domain adaptation or fine-tuning. Moreover, while LLMs excel in general domains, their performance in specialized fields such as GIS may not always align with the expectations of domain experts. This is primarily attributed to the diverse disciplinary origins of the training data, which often lack comprehensive coverage and treatment of knowledge specific to individual disciplines (e.g., GIS). Therefore, there is a crucial need to train and adapt LLMs specifically designed for different professional fields. In this paper, our focus is on the GIS domain, where we introduce BB(BaBy)-GeoGPT, a large language model with GIS-specific knowledge. To achieve this goal, we curated a comprehensive set of resources, comprising model pretraining data (BB-GeoPT, 26,907 documents), supervised fine-tuning data (BB-GeoSFT, 35,876 instructions), and evaluation data (BB-GeoEval, 600 objective questions and 150 subjective questions). BB-GeoGPT is developed by first adapting an open-source general-domain LLM, the LLaMA-2-7B model, to our pretraining data. Subsequently, we use instruction tuning to further fine-tune the model on our BB-GeoSFT. Through extensive experiments on the evaluation dataset, BB-GeoGPT demonstrates improvements ranging from 10.55% to 47.57% for objective questions and from 7.87% to 27.73% for subjective questions, when compared to general LLMs of similar size in terms of accuracy. Moreover, our data collection strategy and the amassed data can serve as a foundation for advancing LLM research in the GIS domain, fostering further development.

### 1. Introduction

The emergence of large language models (LLMs) has significantly advanced the field of natural language processing (NLP) (Min et al., 2023; Zhao et al., 2023b). The commercial deployment of LLMs, such as ChatGPT (Brown et al., 2020; Ouyang et al., 2022) and ERNIE (Wang et al., 2021), also allows the public to directly perceive the innovations and enhancements introduced

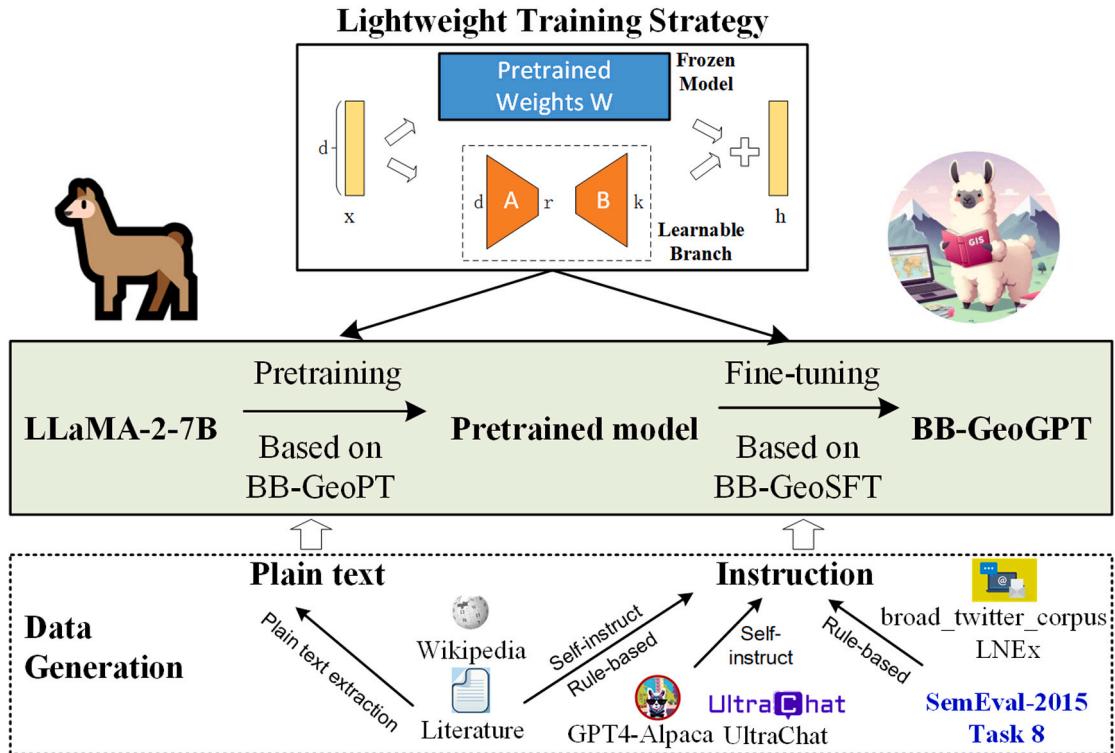
\* Corresponding author at: School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430078, China.  
E-mail address: [yuwh@cug.edu.cn](mailto:yuwh@cug.edu.cn) (W. Yu).

<https://doi.org/10.1016/j.ipm.2024.103808>

Received 19 December 2023; Received in revised form 2 June 2024; Accepted 7 June 2024

Available online 22 June 2024

0306-4573/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



**Fig. 1.** The framework for training a large language model for the geographic information science, termed BB-GeoGPT. Firstly, a comprehensive set of training resources, comprising model pretraining data (BB-GeoPT) and supervised fine-tuning instructions (BB-GeoSFT) is curated. Then, BB-GeoGPT is obtained by first adapting an open-source general-domain LLM, LLaMA-2-7B model (Touvron et al., 2023a) to the pretraining data BB-GeoPT, and then utilizing the instruction tuning data BB-GeoSFT to further supervise fine-tuning the model. A lightweight training strategy is used in both of these stages. Details are elaborated in Sections 3 and 4.

by artificial intelligence (AI) technology. Currently, LLMs exhibit outstanding performance across diverse language tasks within the realm of NLP, demonstrating exceptional capabilities in language comprehension and text generation (Wei et al., 2022). With the continual enhancements of LLMs across general domain tasks, there is a growing interest in the application and deployment of LLMs in specialized domains (Hu et al., 2023a; Li & Ning, 2023; Manvi et al., 2023; Thirunavukarasu et al., 2023). In this paper, we concentrate on developing a large language model for the geographic information science, referred to as BB-GeoGPT. The overall framework of our research is presented in Fig. 1.

In practice, the performance of LLMs in professional tasks is also investigated by researchers from different domains (Chen et al., 2023a; Nori et al., 2023; Wu et al., 2023a). For example, Savelka et al. (2023) evaluated the performance of GPT-4 (OpenAI, 2023) in generating factually accurate, clear, and relevant explanations of terms in legislation, and their results indicate that GPT-4 may yield explanations that appear to be of high quality on the surface but have limitations in terms of factual accuracy. Researchers in the GIS domain also examine whether LLMs can effectively address geospatial tasks (Mooney et al., 2023; Xie et al., 2023). For instance, Mai et al. (2023c) investigated the performance of LLMs on several geospatial tasks including toponym recognition, location description recognition, and US state-level/county-level dementia time series forecasting. Zhang, Wei, He, and Yu (2024a) tested the ability of GPT-3.5-turbo, serving as an agent in the LangChain framework, to comprehend GIS demands and tools and solve geospatial tasks using these tools. However, a similar conclusion in these studies is though LLMs possess a wealth of knowledge in general domains, applying them to professional domains still presents numerous challenges, with the most significant being that general language models often lack a thorough grasp of specific disciplinary expertise, leading to potential misunderstanding errors.

As the success of LLMs hinges significantly on extensive corpus (Snæbjarnarson et al., 2022), one possible explanation for this phenomenon lies in the diverse disciplinary origins of language model training data. Training datasets of LLMs are usually collected from various fields, possibly resulting in a shortfall in the comprehensive collection or processing of data specific to professional domains. This issue may be further pronounced in research domains like GIS, where the conceptual foundation, although partially rooted in information science, diverges substantially from it (Li et al., 2022a). Moreover, the imbalance in data volume across disparate disciplines constitutes a noteworthy potential factor. Evidently, in contrast to corpora data within general domains, data in specialized domains such as medication, agriculture, biology, and geography is more challenging to collect and necessitates more nuanced processing procedures.

Using one of the most outstanding open-source LLMs, the LLaMA-2-7B model (Touvron et al., 2023a), as an example, if we inquire about all the definitions of 'clip', it would provide numerous responses, with the majority being definitions from the general domain,

**Question:** List all the meanings of "clip" across all disciplines.

🦙 **LLaMA-2-7B:**

As a helpful assistant, I can provide you with a comprehensive list of meanings for the term "clip" across various disciplines:

1. Media and Journalism: In media and journalism, a clip refers to a short video or audio recording, often used in news broadcasts or online articles to illustrate a story.
2. Film and Television: In film and television, a clip is a segment of footage that has been edited out of its original context and used as part of another production. This could be a scene from a movie, TV show, or commercial.
3. Music: In music, a clip can refer to ...
4. Computing and Technology: ...
5. Engineering: ...
6. Sports: ...
7. Fashion: ...
8. Education: ...
9. Marketing and Advertising: ...
10. Medical: In medicine, a clip can refer to a small medical device, such as a blood pressure clip, used to measure or monitor a patient's vital signs.

These are just some examples of the many meanings of "clip" across various disciplines. The term can have different meanings depending on the context in which it is used.

### Imbalance in data volume between professional and general disciplines

General domain: "I need to clip the grass in the garden"	➡	Remove "grass"
GIS: "I need to clip this city from the national map"	➡	Retain "city"

### Logical inconsistency among disciplines

**Fig. 2.** The gap between general LLMs and their applications in professional domains, including the imbalance in data volume between professional and general disciplines and accompanying potential logical inconsistencies among different disciplines.

such as media, film, and music (see Fig. 2). However, it lacks a definition specific to the GIS domain, despite 'clip' being one of the most commonly used tools in GIS. This phenomenon illustrates that general language models tend to generate knowledge content more biased towards the data-rich domain. Furthermore, this bias may also exaggerate the influence of logical inconsistencies among different domains. An example is presented in Fig. 2, in the sentence 'I need to clip the grass in the garden', the term 'clip' denotes cleaning away the grass within the garden. In the context of GIS, when stating 'I need to clip this city from the national map' the term 'clip' refers to extracting a specific portion of the map based on certain criteria, rather than a need to discard it.

In practice, an immensely vast dataset may have the potential to mitigate this constraint. For instance, benefiting from extensive and diverse training data sources from different fields, the top-performing LLMs proposed by commercial companies, such as ChatGPT and GPT-4, usually surpass the performance of open-source models in professional domains (Katz et al., 2023; Waisberg et al., 2023). However, challenges arise as these leading models typically have substantial parameter sizes (e.g., GPT-3 with 175 billion parameters), while many of them are solely accessible through APIs, which poses obstacles to in-depth domain-specific research and deployment. Hence, a viable approach to bridge the gap between general language models and professional tasks is training domain-specific LLMs grounded in domain knowledge (Nori et al., 2023). This avenue of exploration has been pursued by researchers across various fields, including law (Song, 2023; Yue et al., 2023), transportation (DUOMO, 2023), education (Latif et al., 2023; Lee et al., 2023), agriculture (Lu et al., 2023), internet of things (Dou et al., 2023), medicine (Li et al., 2023; Thirunavukarasu et al., 2023), and art & humanities (Liu et al., 2023).

In this paper, our focus is on the GIS domain, intending to bridge the gap between general LLMs and specialized GIS knowledge. Since the training corpus forms the foundation for training LLMs, we first constructed a set of corpora within the GIS domain, including model pretraining data corpus (BB-GeoPT), supervised fine-tuning data corpus (BB-GeoSFT), and evaluation data corpus (BB-GeoEval). More specifically, BB-GeoPT comprises a substantial collection of raw texts sourced from GIS-related literature and Wikipedia. BB-GeoSFT is a compilation of self-instruct, well-organized instructions based on professional literature, which includes 6 diverse geospatial knowledge and task data with prompts, such as geospatial question answering (Chen et al., 2013; Mai et al., 2020a, 2020b, 2021), geospatial text summarization, geospatial text generation, and spatial relation extraction (Qiu et al., 2022). Then, BB-GeoEval is gathered from both subjective and objective perspectives, aiming to offer a comprehensive assessment of the language model's abilities in understanding and applying domain-specific knowledge within GIS.

With the training data prepared, we introduce BB-GeoGPT, a language model enriched with GIS-specific knowledge. We termed our proposed model BB(BaBy)-GeoGPT, because we believe that our research as a start will prompt the development of LLMs in GIS, and there will be more powerful and stronger GIS large language models and multimodal foundation models in the future. Specifically, BB-GeoGPT is obtained by first adapting an open-source general-domain LLM, LLaMA-2-7B model (Touvron et al., 2023a), to the pretraining data, and then utilizing the instruction tuning data to further supervise fine-tuning the model. Empowered by enriched geospatial knowledge, our BB-GeoGPT outperforms the original model in understanding geospatial concepts, which can ultimately benefit researchers within GIS.

To sum up, our contributions can be summarized as follows:

1. We introduce BB-GeoGPT, a foundational language model in GIS capable of understanding geospatial concepts and responding to geospatial queries using appropriate prompts. This model can offer practical assistance to GIS researchers.

2. Our data collection strategy and the constructed data corpora can serve as a foundation for LLM research in the GIS domain.
3. Experiments conducted on the evaluation data validate the effectiveness of our proposed BB-GeoGPT and the collected datasets. These findings affirm the viability of our research idea and provide a practical approach for future researchers.

## 2. Related work

Large language models (LLMs) have achieved astounding performance on different textual tasks and showed strong abilities in language understanding, thinking, reasoning, and planning (Chang et al., 2023; Huang & Chang, 2022; Huang et al., 2022). The rapid advancement of LLMs has presented researchers with numerous opportunities, leading to the exploration of intriguing ideas and applications (Caines et al., 2023; Singh et al., 2023; Wang et al., 2023b). Currently, researchers are not only focused on traditional tasks in natural language processing (NLP) but are also interested in integrating LLMs with applications in other domains, such as recommender systems (Gao et al., 2023), image generation (Zhang et al., 2023c), location description extraction (Hu et al., 2023a; Mai et al., 2022b), travel plan generation (Roberts et al., 2023), socioeconomic indicator prediction (Manvi et al., 2023), and agents (Andreas, 2022; Huang et al., 2023a; Richards, 2023). In general domain scenarios, due to the extensive knowledge and powerful logical reasoning and thinking abilities possessed by LLMs, they can perform exceptionally well than traditional methods (Min et al., 2023). However, when it comes to tasks in many professional domains, the performance of generic LLMs may not satisfy the demands of experts. A phenomenon presented is LLMs may output an answer that looks good but lacks professionalism or factual accuracy when facing a professional question (Savelka et al., 2023). This has significantly fueled researchers' eagerness to investigate the application of LLMs in professional fields (Yue et al., 2023).

However, two challenges needed to be addressed before training a domain-specific LLM: the high demand for computing resources and the scarcity of large-volume professional training data (Hadi et al., 2023). Given that LLMs typically have a large number of parameters and are trained on extensive datasets, the demand for computing resources is substantial. To address this challenge, various effective lightweight training strategies for large models have been proposed, including LoRA (Hu et al., 2021b; Zhang, Zhang, Deng, Zhang, Lin, Huang, Zhang, & Yu, 2024b) and Q-LoRA (Dettmers et al., 2023). These strategies enable LLMs to undergo further training or fine-tuning based on specific datasets, even with limited computing resources. Then, as for the collection of training data, researchers also provide some synthetic-data-generation strategies, such as generating datasets based on LLMs (Ding et al., 2023a; Wang et al., 2022b). For example, Wang et al. (2022b) proposed a Self-Instruct framework, designed to generate instruction datasets to enhance the instruction-following capabilities of LLMs. This is achieved through a bootstrapping mechanism, utilizing content generated by an LLM with abundant knowledge (e.g., vanilla GPT-3). Taori et al. (2023a) applied the self-instruct framework, using an LLM (OpenAI's Text-Davinci-003) to generate 52 K instructions for model fine-tuning. After fine-tuning an open-source LLM (LLaMA (Touvron et al., 2023b)) on the generated instructions, the fine-tuned model demonstrated comparable performance with Text-Davinci-003. Many subsequent works also validate the effectiveness of self-generated datasets based on LLMs (Chiang et al., 2023; Ding et al., 2023a; Xu et al., 2023).

With these prior contributions, researchers from various fields have endeavored to tailor LLMs in their respective domains to offer more specialized services. For example, Yue et al. (2023) proposed DISC-LawLLM, an intelligent legal system utilizing a fine-tuned law-specific LLM to provide a broad range of legal services. Qi et al. (2023) introduced FoodGPT, a LLM for the field of food testing. Singhal et al. (2023) proposed a medical-specific LLM towards expert-level medical question answering. LLAVA-Med was developed as a multimodal foundation model including an LLM for various medical tasks such as medical visual question answering (Li et al., 2023) by fine-tuning on the PubMed corpus. Deng et al. (2023) proposed a foundational language model for geoscience knowledge understanding and utilization. These endeavors suggest that compared with generic LLMs, training a domain-specific LLM can offer better assistance to both expert and non-expert users when dealing with various tasks in the professional domain. However, we have yet to see a special LLM trained for tasks in the GIS domain despite professional geospatial and GIS knowledge having unique characteristics compared to information within general domains. Therefore, we aim to address this gap by proposing BB-GeoGPT and presenting relevant collected datasets, to provide professional GIS language services for users.

## 3. BB-GeoGPT data collection and processing

In this paper, we have constructed three corpora for different training stages of BB-GeoGPT: pretraining corpus (BB-GeoPT, Section 3.1), supervised fine-tuning corpus (BB-GeoSFT, Section 3.2), and evaluation corpus (BB-GeoEval, Section 3.3). Generally, the pretraining corpus comprises a large volume of unlabeled raw text aimed at injecting professional knowledge into language models (LLMs). The fine-tuning corpus consists of high-quality professional instructions designed to enhance the instruction-following ability of LLMs in professional tasks. Lastly, the evaluation corpus is employed to assess the effectiveness of the obtained LLMs in various professional tasks. The subsequent sections will provide detailed insights into the collection and processing of these three types of datasets.

### 3.1. Pretraining corpus: BB-GeoPT

Our collected text corpus for model pretraining consists of two parts, including GIS-related papers published in high-quality journals and GIS-related pages from Wikipedia. The statistic of the collection of text corpus is shown in Table 1.

**Table 1**  
Statistics of pretraining data.

Data source	Document	Size (Megabyte)
GIS-related papers	2,499	173
GIS-related Wikipedia pages	24,408	115

**Table 2**  
Statistics of instruction tuning data.

Dataset	Data generation method	Task	Count
GPT4-Alpaca	Self-instruct dataset	General domain	52,000
	Self-instruct dataset	Geospatial knowledge question answering and word semantics	20,145
	Rule-based dataset	Text summarization	2,172
	Open-source professional dataset	Text generation Named entity recognition Spatial relation extraction Multiple GIS-related topics	2,172 7,500 553 3,334
BB-GeoSFT			

### 3.1.1. From GIS literature

In BB-GeoPT, we collected 2,499 GIS-related papers from 12 journals in the GIS domain. Since the documents we obtained are in PDF format, the first step is to convert them into plain text. In this paper, we employ the method outlined in pdfminer ([Jeong, 2011](#)) to accomplish this task. PDF is a layout-based format that specifies the fonts and positions of individual characters, and different files may have different layouts and coded formats. As a result, the conversion process may produce unexpected unusual characters or garbled text in the resulting plain text ([Bast & Korzen, 2017](#)). During the pretraining phase, the model is trained in an unsupervised autoregressive manner, where the objective is to predict the next word in a sequence given the previous words. Including unusual characters or garbled text would render the prediction training ineffective. Therefore, we remove any unusual characters from the converted texts to enhance their quality.

### 3.1.2. From Wikipedia

Besides professional literature, we also utilize information from Wikipedia as a significant resource to train BB-GeoGPT. Specifically, the root node of the Wikipedia category we used is ‘Geographical technology’ ([https://en.wikipedia.org/wiki/Category:Geographical\\_technology](https://en.wikipedia.org/wiki/Category:Geographical_technology)), from which we iteratively extract all the pages within this category and its subcategories. In the end, we collected a total of 24,408 pages to comprise BB-GeoPT.

## 3.2. Instruction tuning corpus: BB-GeoSFT

### 3.2.1. Dataset overview

Our instruction tuning dataset comprises two parts: general instruction data and professional instruction data tailored to the GIS domain. The details are listed in [Table 2](#).

### 3.2.2. General instruction tuning data

In this paper, we collect general instruction data to enrich the diversity of our training set and further align the selected base model (LLaMA-2-7B) with human-like language understanding ([Wang et al., 2023a](#)). These instructions help mitigate the risk of diminishing the foundational capability of the language model during the supervised fine-tuning process ([Deng et al., 2023; Yue et al., 2023](#)). For our dataset, we gathered general instructions from the open-source GPT4-Alpaca dataset ([Taori et al., 2023a](#)), which has been validated for its effectiveness in aligning LLMs with human-level language understanding.

### 3.2.3. Professional instruction tuning data

In this section, the strategy to collect professional instruction tuning data is described. Compared to general instructions, professional instructions are used to align LLMs with GIS experts. Our data collection methods include three approaches: self-instruct, rule-based, and gathering from open-source professional datasets.

**Self-instruct dataset.** Self-instruct was introduced by [Wang et al. \(2022b\)](#) and the method is proposed to overcome the constraint that fine-tuning language models heavily relies on limited and less diverse human-written instruction data. The fundamental concept behind self-instruct involves leveraging the extensive knowledge, advanced language understanding, and text generation capabilities

inherent in LLMs to generate high-quality instructions on diverse topics. In this process, an LLM is initially provided with human-written topics and then prompted to generate relevant questions or tasks based on these topics. Subsequently, the LLM is tasked with responding to these questions or tasks. In this way, instruction data can be constructed based on these questions or tasks and corresponding responses to them. In practice, the whole process can be automatic since these human-written topics can even be generated by LLMs (Ding et al., 2023a).

However, as discussed in Section 1, LLMs demonstrate greater robustness in general domains, while potentially yielding unsatisfactory results in specialized fields like GIS. Given our objective of enhancing the instruction-following capability of general LLMs in the GIS domain, relying solely on self-instruct based on LLMs is not suitable for our research. Specifically, we cannot depend on the limited GIS knowledge within general LLMs to generate diverse professional topics or tasks related to GIS, let alone respond to corresponding professional questions. Therefore, in this paper, general LLMs are not tasked with generating professional instructions from their knowledge base but rather with comprehending, summarizing, and reformatting the knowledge provided by professional literature. In practice, the tasks of comprehending, summarizing, and reformatting texts are also areas where LLMs excel.

Specifically, we meticulously crafted a prompt (see Fig. 3a) to guide an LLM to understand and summarize text sourced from professional literature and Wikipedia pages collected in BB-GeoPT. Subsequently, the LLM is tasked with generating questions or tasks and their corresponding answers based on the provided text and their summarization. These questions and answers then serve as the instruction data. Importantly, all the topics and knowledge within the questions and answers are highly pertinent to GIS and possess a high degree of professionalism, given that the source text is derived from a professional GIS corpus.

Since we do not require LLMs to generate knowledge but rather to understand and summarize provided text, there is no need for a very large LLM (e.g., GPT-4), as regular LLMs already demonstrate robust language understanding abilities. Therefore, we utilize an excellent open-source LLM, the LLaMA-2-7B-chat model, for generating professional instructions. An example demonstrating the effectiveness of our approach is presented in Fig. 3b and 3c. Following the generation of instructions, a manual check was conducted to eliminate any generated instructions with poor quality. In practice, the number of instructions of poor quality was found to be minimal (less than 1%), further validating the effectiveness of our approach. Ultimately, we obtained a total of 20,145 professional instructions.

**Rule-based dataset.** The instructions generated by self-instruct primarily focus on geospatial knowledge question answering and geospatial word semantics. In this section, we employ a rule-based method to generate instructions, aiming to enhance the ability of general LLMs in text generation and summarization in the GIS domain. Specifically, we manually selected 1,087 representative papers from GIS-related journals. We then employed the method provided by GROBID (Lopez, 2008) to convert these PDF files into XML files, allowing us to extract the title, abstract, and keywords of each paper. Subsequently, we employed a rule-based method to convert the title, abstract, and keywords in each paper into four types of instructions. These four types of instructions include summarizing the title from the abstract, summarizing the keywords from the abstract, writing the abstract based on the title, and writing the abstract based on the keywords. An example of summarizing the title from the abstract is presented in Fig. 4a. This process resulted in a total of 4,344 instructions being obtained.

#### Open-source professional dataset

- **Name entity recognition.** Named entity recognition aims for the automated identification of individuals, organizations, spatial locations, and other entities from natural language texts (Li et al., 2022b). This task involves the identification and extraction of geographical location names from the text, aiming to enhance the comprehension and utilization of spatial information within textual data (Hu et al., 2022). Accurately identifying named entities from natural text is crucial for various applications within social media analysis (Liu, Chen, & Xia, 2022a) and relevant spatial analysis applications, such as disaster response (Kedzie, McKeown, & Diaz, 2015; Neubig et al., 2011). To enhance the capabilities of the language model on this task, we have incorporated instruction data related to named entity recognition into our dataset. Specifically, our collected instruction data is based on annotated tweet data originating from Broad Twitter Corpus (Derczynski, Bontcheva, & Roberts, 2016) and LNEx (Al-Olimat et al., 2017). The Broad Twitter Corpus (Derczynski et al., 2016) is collected from different regions, temporal periods, and different types of Twitter users. The named entity annotations in this dataset are made by a collaboration between NLP experts and crowd workers. This collaborative approach ensures that the data benefits from crowd recall while upholding high-quality standards, and the dataset has also been utilized by many subsequent works (Hu et al., 2021a). Since Broad Twitter Corpus (Derczynski et al., 2016) contains all original text and the annotated corpus, we conducted preprocessing on the data, ultimately yielding 3,000 instruction data entries. Then, LNEx (Al-Olimat et al., 2017) focuses more on place name extraction, and the contributors annotated three distinct Twitter streams related to flooding events in different locations including the 2015 Chennai flood, the 2016 Louisiana flood, and the 2016 Houston flood. After preprocessing, we obtained a total of 4,500 instruction data entries from LNEx (Al-Olimat et al., 2017). The specific format of this type of instruction data is presented in Fig. 4b.
- **Spatial relation extraction.** Extracting spatial relations from natural language text is a pivotal research focus within the field of spatial analysis and NLP, drawing significant attention from researchers from various fields (Hu et al., 2023b; Mani et al., 2010). Numerous methodologies have been proposed by researchers in attempts to address this issue (D'Souza & Ng, 2015). In our collected instruction dataset, we have endeavored to compile a set of spatial relation extraction instructions from open-source datasets. Specifically, our collection predominantly stems from the SemEval-2015 Task 8 dataset (Pustejovsky et al., 2015), which is dedicated to identifying and classifying items from an inventory of spatial concepts. This task encompasses the extraction of locations, entities participating in spatial relations, topological relations, direction relations, orientation,

Assuming you are the teacher of Geographic Information Science Major in University and you should come up with Q&A pairs.  
Here are some reference books and documents for you.  
Extract conceptual knowledge about professional foundations in the article.  
Generate professional questions in the field of geographic information science according to the answers in the article.  
The generated questions and answers cannot contain such referential words as "chapter", "book", "section", "box", "article" or "figure".  
Questions and answers need to be matched one by one.  
Wrap after completing a set of questions and answers.  
After generating question and answer pairs, please output them in the following format:  
"  
[  
Q: The generated questions are here.  
A: The generated answer is placed here.  
]  
"  
The generated question and answer pairs must start with Q and end with A.

(a) Prompt used for self-instruct instruction generation

are often used to store the attribute data associated with these two data structures. Overview This chapter first reviews the characteristics of the two main data models that are used to represent spatial data and some common file formats. The chapter then describes some of the ways that geographic information systems store, query, and relate attribute data to these data structures through the use of databases. Students should gain a good understanding of these concepts to effectively structure, edit, and analyze spatial data. GIS Data Models Data models define how real-world spatial features are represented in a GIS (Bolstad and Smith, 1992). Geospatial data are typically represented by two main data models (also called data structures): vector and raster. The representation of real-world spatial information in vector and raster format is shown in Figure 5-1. The vector representation consists of points, lines, and polygons created using beginning and ending nodes and intervening vertices, each with detailed x,y coordinate information. The raster representation stores the spatial information in a user-defined grid where every cell (picture element or pixel) has a unique geographic location and attribute value. Vector data models can be further subdivided into georelational and object-based data models. Vector Data Vector maps represent the most common form of maps that you typically use. For example, almost all road atlases, GPS navigation

Q: What are the two main data models used to represent spatial data in a GIS?

A: The two main data models used to represent spatial data in a GIS are vector and raster.

Q: How do vector data models represent real-world spatial features?

A: Vector data models represent real-world spatial features by using points, lines, and polygons created using beginning and ending nodes and intervening vertices, each with detailed x,y coordinate information.

(b) Provided professional literature

(c) Generated instructions

Fig. 3. The process of generating instruction tuning data based on GIS professional literature by self-instruct method.

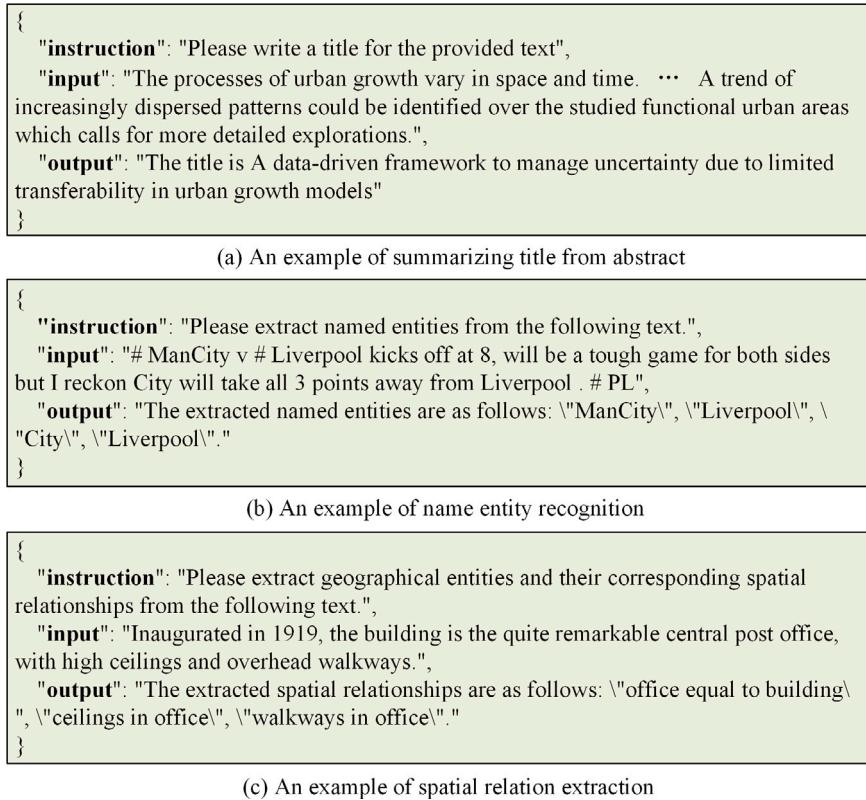
and other related content. This dataset has been widely adopted in spatial relation extraction research (Wang, Li, & Zhu, 2022a). This dataset comprises original text along with annotated spatial relation information. Based on these annotations, we conducted preprocessing on the data, ultimately yielding 553 instruction data entries. The specific format of the instruction data is presented in Fig. 4c.

- **Multiple GIS-related topics.** We also extract professional data from the existing open-source self-instruct dataset UltraChat Ding et al. (2023a) to enrich the diversity of our dataset. Specifically, UltraChat comprises 1.5 million high-quality dialogues, spanning a diverse array of topics and instructions. Certainly, it also includes a substantial amount of instructional data related to GIS topics. In this way, we use a rule-based method to extract corresponding instructions to supplement our dataset. Specifically, we first collected a set of keywords in the GIS domain, such as “geographic information science” and “spatial analysis”, and then we retrieved data containing these keywords from the instructional dataset. Finally, we obtained a total of 3,334 instructions as our training data, which originates from various GIS-related scenes and topics.

### 3.3. Evaluation corpus: BB-GeoEval

To evaluate the effectiveness of our trained language model in addressing geospatial inquiries and its proficiency in comprehending and applying GIS knowledge, we gather data from various question-answer platforms, conduct crawling on multiple open-source test websites, and subsequently establish a benchmark, denominated BB-GeoEval. Specifically, BB-GeoEval is constructed to assess the ability of language models from both the objective perspective and the subjective perspective.

First, for the objective and quantitative assessment of language models’ GIS knowledge and reasoning proficiency, we constructed an objective evaluation dataset. This dataset consists of multiple-choice questions, each allowing for the possibility of a singular correct answer. The objective evaluation dataset is sourced from the Studocu website (<https://www.studocu.com/>), a platform where students can find and share study materials such as lecture notes, summaries, and past exams. After collecting the dataset, all questions are manually inspected and corrected for any errors to ensure their effectiveness in evaluating models. The diversity of the objective evaluation dataset is presented in Table 3.



**Fig. 4.** Formats of rule-based instructions and instructions generated from open-sourced professional datasets. (a) An example of summarizing a title from an abstract, (b) an example of name entity recognition, and (c) an example of spatial relation extraction.

**Table 3**  
Statistics of evaluation data.

Category	Spatial Analysis	Geodatabase	Cartography	Remote Sensing	Surveying	Total
Objective Questions	163	174	130	83	50	600
Subjective Questions	44	38	24	29	15	150

Subsequently, a subjective evaluation was devised to delve into the model's comprehension of GIS knowledge and answering abilities. This evaluation adopts a question-answering paradigm, simulating the process akin to subjective examination questions. The subjective evaluation dataset is collected from multiple sources, including the Baidu Document Library (<https://wenku.baidu.com/>) and the National Postgraduate Entrance Exam questions on GIS in China. All questions undergo manual inspection and verification.

In summary, the objective task comprises 600 multiple-choice questions, while the subjective task involves 150 questions. To further scrutinize the efficacy of our trained model across diverse domains, all questions are categorized into five relevant types: spatial analysis, geodatabase, cartography, remote sensing, and surveying. All questions have been meticulously processed and organized, accompanied by corresponding answers provided by professionals. The case examples of subjective and objective questions are illustrated in Fig. 5.

#### 4. The training process of BB-GeoGPT

In this section, we elaborate on the training process of BB-GeoGPT. Training a professional LLM from scratch demands significant computing resources and a large data corpus. A more feasible approach is to develop BB-GeoGPT based on a foundational general-domain pretrained LLM. In this paper, we chose the open-source LLaMA-2-7B model as the foundational model, which is an open-source LLM with about 7 billion parameters. LLaMA-2-7B is chosen as the foundational model primarily for its outstanding performance across various NLP tasks, including inference, encoding, and knowledge testing, surpassing many other open-source language models (Touvron et al., 2023a; Zhao et al., 2023a). Moreover, the feasibility and effectiveness of training a professional LLM based on the LLaMA-2-7B model has been validated by related research (Huang et al., 2023b; Taori et al., 2023b). Specifically, the training process is designed to involve two subprocedures: pretraining and supervised fine-tuning, and the framework is presented

**Question:** \_\_\_\_\_ deals with spatial properties that do not change under certain transformations.  
**Options:** Node Shapefile Topology Model  
**Answer:** Topology  
**Category:** Spatial Analysis

(a) An objective question example

**Question:** What is TIN?  
**Answer:** Triangulated Irregular Network (TIN) is an approximate representation of a surface obtained from irregularly spaced sample points and breakline features. It includes the topological relationships between points and the triangles adjacent to them. The shape and size of the triangles depend on the density and location of the irregularly distributed sample points, which helps avoid data redundancy in flat terrains and allows for the representation of digital elevation features based on terrain characteristic points.  
**Category:** Geodatabase

(b) A subjective question example

**Fig. 5.** The illustration for objective and subjective questions: (a) an objective question example and (b) a subjective question example.

in Fig. 1. Considering the limitations of computing resources, a lightweight training strategy is used during both the two phases (Hu et al., 2021b). Therefore, in this section, this strategy is first described, followed by the details of two phases to train the BB-GeoGPT.

#### 4.1. Lightweight training strategy

In the field of NLP, a commonly used paradigm for applying language models for specific tasks or domains is pretraining a foundational language model on general domain data and then fine-tuning it on professional data. This approach eliminates the need for researchers to train separate language models for individual tasks, providing significant advantages. However, since the emergence of LLMs, fine-tuning all their parameters has become less feasible due to the prohibitively expensive nature of retraining LLMs with large parameter sizes (e.g., GPT-3, 175B). In this way, Parameter-Efficient Fine-Tuning (PEFT) methods for large-scale pre-trained language models are needed. These methods aim to fine-tune models by optimizing a small portion of the model parameters while keeping the rest fixed, thereby reducing computation and storage costs. There are many PEFT methods, such as LoRA (Hu et al., 2021b), soft prompt tuning (Lester, Al-Rfou, & Constant, 2021), prefix-tuning (Li & Liang, 2021), and adapter (Houlsby et al., 2019). These methods adopt different strategies and have been shown to be effective for different tasks (Ding et al., 2023b). In the task of fine-tuning a foundational general LLM to a professional LLM, LoRA (Hu et al., 2021b) is one of the most commonly used PEFT methods, and has been validated by numerous studies, such as training professional LLMs in medicine (Gema et al., 2023), law (Song, 2023), and geoscience (Deng et al., 2023). Therefore, we also adopted LoRA for fine-tuning the selected foundational LLaMA-2-7B model.

Specifically, LoRA (Hu et al., 2021b) adopted a strategy of freezing the pretrained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer (Vaswani et al., 2017) architecture in LLMs. An illustration is presented at the top of Fig. 1. With LoRA, the model weights  $\mathbf{W} \in \mathbb{R}^{d \times k}$  are frozen, where  $d$  and  $k$  are the dimensions of input and output feature, respectively. Then, only additional, specially designed modules  $\mathbf{A} \in \mathbb{R}^{d \times r}$  and  $\mathbf{B} \in \mathbb{R}^{r \times k}$  with small sizes of parameters are trained based on domain knowledge, where the rank  $r \ll \min(d, k)$ . For example, given input feature  $\mathbf{x}$ , the output hidden feature  $\mathbf{h}$  can be obtained as follows.

$$\mathbf{h} = \mathbf{x}\mathbf{W} + \mathbf{x}\mathbf{AB} \quad (1)$$

In this way, since  $\mathbf{W}$  is frozen and does not receive gradient updates and only  $\mathbf{A}$  and  $\mathbf{B}$  contain trainable parameters during LoRA training, this approach can significantly conserve training resources. Moreover, the structure of extra branches can be customized to accommodate various computing resource settings. The effectiveness of LoRA has been validated by relevant works (Wang, 2023), indicating that LoRA performs on par or even better than fine-tuning in terms of model quality. Therefore, we also adopt LoRA in our research.

#### 4.2. BB-GeoGPT pretraining

First, further pretraining the foundational general model on a professional GIS corpus can enhance its fundamental understanding and generation capabilities in professional domains (Cui, Yang, & Yao, 2023). Therefore, in the pretraining phase, we use domain-specific text corpus BB-GeoPT to further pretrain the base LLaMA-2-7B model. In this phase, the model is trained in an unsupervised autoregressive manner, where the objective is to predict the next word in a word sequence given the previous words. LoRA is used in this phase.

The hyperparameters setting of this pretraining process are as follows: a global batch size of 32, a learning rate of 2e-4, 1 epochs training stage, and a maximum target length of 512 tokens. The pretraining process was carried out on 4×A100 GPUs. As for the

**Table 4**

Evaluation results based on the objective tasks.

Model	Spatial Analysis	Geodatabase	Cartography	Remote Sensing	Surveying	Average
LLaMA-2-7B	0.387	0.402	0.446	0.470	0.580	<b>0.432</b>
Alpaca-7B	0.380	0.362	0.446	0.494	0.460	<b>0.412</b>
Vicuna-7B	0.534	0.500	0.546	0.663	0.600	<b>0.550</b>
K2	0.387	0.339	0.385	0.349	0.420	<b>0.370</b>
GPT-3.5-turbo	0.816	0.707	0.800	0.843	0.760	<b>0.780</b>
BB-GeoGPT	0.583	0.529	0.638	0.711	0.720	<b>0.608</b>

LoRA setup, we set LoRA rank  $r$  as 8. We set the target modules as  $k_{proj}$ ,  $q_{proj}$ ,  $v_{proj}$ , and MLP layers, based on our experimental observation.  $k_{proj}$ ,  $q_{proj}$ ,  $v_{proj}$ , and MLP layers are part of the Transformer structure (i.e., the  $W$  in Fig. 1). During the pretraining stage, we use bf16 as the floating point format to load the model and save computing resources.

#### 4.3. BB-GeoGPT supervised fine-tuning

With the pretraining stage, the trained model can process more GIS professional knowledge than its original version. During the pretraining phase, the objective is to predict the next word in a word sequence given the previous words. However, users often expect language models to "follow their instructions helpfully and safely", a requirement not directly addressed by the pretraining objective. To bridge this gap, instruction tuning is proposed as an effective technique to enhance the capabilities and controllability of large language models (Ouyang et al., 2022). Instruction tuning involves further training LLMs using (*instruction*, *input*, *output*) pairs (see Fig. 4), where *instruction* and *input* denote the human instruction for the model, and *output* denotes the desired response that follows the instruction. This process helps align language models with the desired response characteristics and domain knowledge (Zhang et al., 2023). It should be noted that instruction tuning differs from few-shot learning (Brown et al., 2020; Perez, Kiela, & Cho, 2021), which aims to quickly adapt models to some specific downstream tasks by providing them with a few examples related to those tasks. Therefore, few-shot learning does not align with our goal, which is to better align the pretrained language model with professional instructions and GIS knowledge.

Therefore, to further enhance the instruction-following ability, the pretrained model is further fine-tuned based on our collected general instructions and BB-GeoSFT dataset. In this phase, LoRA is also utilized, and the settings are also similar to that in the pretraining stage with only the learning rate changed as 1e-4. After the fine-tuning process, our BB-GeoGPT can provide professional services for GIS-related users, and the evaluation section will validate its effectiveness.

### 5. BB-GeoGPT evaluation

In this section, we introduce the evaluation methods and present corresponding results for our trained model, BB-GeoGPT, along with related baselines. The evaluation is based on the dataset BB-GeoEval. In this paper, three general-purpose open-sourced LLMs including the original LLaMA-2-7B model (Touvron et al., 2023a), Alpaca-7B (Taori et al., 2023a), and Vicuna-7B (Chiang et al., 2023) are selected as the baselines. Additionally, K2 (Deng et al., 2023), a professional foundational language model for geoscience knowledge understanding and utilization, is also selected as a baseline. Moreover, we also compare our BB-GeoGPT with one of the most prominent commercial close-sourced LLM, GPT-3.5-turbo which served as the commercial LLM reference.

#### 5.1. Objective task

First, the evaluation on the objective task is illustrated. As described in Section 3.3, there are a total of 600 multiple-choice questions, each allowing for the possibility of a singular correct answer. Therefore, a prompt is constructed for all the baselines to acquire the corresponding answer for each question, which is described as: "Which option is the best answer to this question: {question}. Options: {options}". The obtained answers were then used to calculate Accuracy scores based on the ground truth. Since the objective questions have been categorized into relevant geospatial domains, corresponding scores in each domain are also calculated for a comprehensive evaluation. The results are detailed in Table 4. Our model demonstrated significantly higher accuracy scores than all three open-source LLMs, including LLaMA-2-7B, Alpaca-7B, and Vicuna-7B, as well as the professional LLM K2 in geoscience domain, across all five domains. This validates the effectiveness of the collected training data and strategies.

While there remains a performance gap between GPT-3.5-turbo and BB-GeoGPT, our model's accuracy scores in certain domains (e.g., remote sensing and surveying-related questions) closely approach those of GPT-3.5-turbo. We find this phenomenon exciting, considering that GPT-3.5-turbo is one of the most outstanding commercial close-sourced large language models (LLMs), boasting a vast parameter count (175B) and trained on a large-scale dataset. Additionally, it incorporates numerous optimization strategies, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). In other words, even though our model yields slightly inferior results compared to GPT-3.5-turbo, the comparison also validates the effectiveness of our model.

Finally, the results indicate a relatively larger gap between our model and GPT-3.5-turbo in spatial analysis, geodatabase, and cartography-related questions. This suggests a need for more training data in these areas, and we plan to enhance our model in these specific directions.

Please act as an impartial judge and evaluate the quality of the responses provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and an assistant's answer. Your job is to evaluate the assistant's answer in terms of Accuracy, Completeness, and Clarity and assign scores.

Accuracy means the content and semantics of the pending scored answer should be consistent with the reference answer. Completeness means compared to the reference answer, the pending scored answer does not miss any details in the reference answer. Do not let the length of the pending scored answer influence your judgment. Clarity means compared to the reference answer, the juridical logic analysis of the pending scored answer is rigorous and clear, and the sentences are well-organized.

Begin your evaluation by comparing the assistant's answer with the reference answer. Identify and correct any mistakes. Avoid any potential biases. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: Accuracy: 1 to 5; Completeness: 1 to 5; Clarity: 1 to 5. All the scores should be between 1 and 5, and higher value means better response.

[User Question]

{question}

[The Start of Reference Answer]

{answer\_ref}

[The End of Reference Answer]

[The Start of Assistant's Answer]

{answer}

[The End of Assistant's Answer]

**Fig. 6.** The prompt used for GPT-4 as a referee.

**Table 5**  
Evaluation results based on subjective task by GPT-4.

Index	Category	LLaMA-2-7B	Alpaca-7B	Vicuna-7B	K2	GPT-3.5-turbo	BB-GeoGPT
Accuracy	Spatial Analysis	3.023	3.750	3.182	3.795	4.705	4.000
	Geodatabase	3.184	3.816	3.796	3.789	4.579	4.158
	Cartography	3.208	3.917	3.152	3.542	4.625	4.250
	Remote Sensing	3.448	3.793	2.730	3.897	4.793	4.000
	Surveying	4.067	3.867	3.000	4.000	4.733	4.333
	Average	<b>3.280</b>	<b>3.813</b>	<b>3.220</b>	<b>3.793</b>	<b>4.680</b>	<b>4.113</b>
Completeness	Spatial Analysis	2.455	2.977	2.964	3.045	4.567	3.068
	Geodatabase	2.500	3.053	3.367	3.184	4.289	3.211
	Cartography	2.458	3.125	2.636	2.750	4.500	3.292
	Remote Sensing	2.759	3.069	2.432	3.069	4.670	3.034
	Surveying	3.467	3.267	2.769	3.400	4.467	3.467
	Average	<b>2.627</b>	<b>3.067</b>	<b>2.885</b>	<b>3.073</b>	<b>4.500</b>	<b>3.173</b>
Clarity	Spatial Analysis	4.455	4.591	4.400	4.568	5.000	5.000
	Geodatabase	4.447	4.895	4.673	4.842	5.000	4.974
	Cartography	4.583	4.792	4.303	4.667	5.000	4.958
	Remote Sensing	4.483	4.828	3.622	4.931	5.000	4.931
	Surveying	4.400	4.667	3.577	4.533	5.000	5.000
	Average	<b>4.473</b>	<b>4.753</b>	<b>4.200</b>	<b>4.720</b>	<b>5.000</b>	<b>4.973</b>

## 5.2. Subjective task

In this section, the evaluation on the subjective task is illustrated. In contrast to objective questions, where each question has a distinct and objectively correct answer, assessing subjective questions is often more intricate. It necessitates assessors to hold substantial domain expertise and make efforts to minimize the impact of potential biases. In this paper, each subjective question has been provided with an answer from professionals as a reference. Additionally, we have employed two evaluation methods to ensure objective and impartial assessment results. One method involves using GPT-4 as a referee, while the other includes inviting multiple professionals specializing in GIS for assessment.

### 5.2.1. GPT-4 as a referee

GPT-4 is currently one of the most outstanding commercial large language models, possessing extensive domain knowledge and logical reasoning capabilities. Several studies have successfully utilized it for the assessment of subjective answers in professional

**Table 6**

Evaluation results based on subjective task by professionals. Win denotes the ratio BB-GeoGPT outperforms the corresponding model.

Category	Evaluation	LLaMA-2-7B	Alpaca-7B	Vicuna-7B	K2	GPT-3.5-turbo
Spatial Analysis	Win	0.583	0.386	0.545	0.455	0.295
	Loss	0.347	0.364	0.205	0.227	0.477
	Tie	0.070	0.250	0.250	0.318	0.227
Geodatabase	Win	0.645	0.474	0.368	0.474	0.316
	Loss	0.298	0.237	0.289	0.263	0.447
	Tie	0.057	0.289	0.342	0.263	0.237
Cartography	Win	0.590	0.500	0.542	0.542	0.167
	Loss	0.351	0.167	0.292	0.125	0.333
	Tie	0.059	0.333	0.167	0.333	0.500
Remote Sensing	Win	0.503	0.276	0.379	0.310	0.034
	Loss	0.382	0.172	0.207	0.207	0.552
	Tie	0.115	0.552	0.414	0.483	0.414
Surveying	Win	0.439	0.600	0.467	0.400	0.400
	Loss	0.422	0.333	0.067	0.200	0.467
	Tie	0.139	0.067	0.467	0.400	0.133
Average	Win	<b>0.570</b>	<b>0.427</b>	<b>0.460</b>	<b>0.440</b>	<b>0.240</b>
	Loss	<b>0.349</b>	<b>0.260</b>	<b>0.227</b>	<b>0.213</b>	<b>0.460</b>
	Tie	<b>0.081</b>	<b>0.313</b>	<b>0.313</b>	<b>0.347</b>	<b>0.300</b>

domains, and it has demonstrated commendable performance (Deng et al., 2023; Yue et al., 2023; Zheng et al., 2023). GPT-4 has recently been widely used for the LLM performance evaluation. For example, Li et al. (2023) used language-only GPT-4 to compute the relative scores to evaluate the multimodal chat instruction-following abilities of LLaVA-Med. Therefore, we also employ the GPT-4 model as an evaluator to assess the quality of subjective answers. Specifically, three indexes are utilized to evaluate the quality of the model response, including answer accuracy, completeness, and clarity. Accuracy means the content and semantics of the pending scored answer should be consistent with the reference answer. Completeness means that compared to the reference answer, the pending scored answer does not miss any details in the reference answer. Clarity means that compared to the reference answer, the juridical logic analysis of the pending scored answer is rigorous and clear, and the sentences are well-organized. Then, a prompt is designed to guide GPT-4 for an objective evaluation, which is presented in Fig. 6. Specifically, GPT-4 is tasked with objectively scoring model responses based on a reference answer across the three mentioned criteria. Scores range from 1 to 5, with higher scores indicating better response quality. Scores are calculated for each of the five domains to provide a comprehensive evaluation.

All the results are detailed in Table 5. In comparison to LLaMA-2-7B, BB-GeoGPT exhibits improvements across all three indexes in the five domains, affirming the efficacy of our model. Simultaneously, when compared to Alpaca-7B and Vicuna-7B, our model excels across nearly all metrics, with only slight drawbacks in 2 out of the  $18 \times 2$  metrics. Additionally, our model also demonstrated superior performance compared to K2, with only a slight drawback on one metric. This further underscores the superior performance of our model over general language models with similar parameter sizes (e.g., 7 billion). When compared to GPT-3.5-turbo, the primary discrepancy lies in the Completeness scores. One potential contributing factor to this outcome is that GPT-3.5-turbo tends to generate relatively longer responses, while GPT-4 may assign higher values to longer answers (Zheng et al., 2023), even though we have reminded it not to consider the answer length in the prompt. Certainly, our responses may have some omissions compared to the reference answers, leading to lower Completeness scores. Therefore, two practical solutions to further enhance our results include collecting more training data to enrich the knowledge database and, more importantly, improving the quality (e.g., length) of instruction data, which significantly influences the output length of the model.

### 5.2.2. Assessment based on professionals

Then, we also invite 12 professionals specializing in GIS for assessment. These professionals are provided with questions and generated answers by different models. For a fair comparison, the sequence of answers for each question by different models is randomly disordered, and potential identification information in model answers is removed. Professionals are asked to judge which answer, between two outputs (answers from BB-GeoGPT and one baseline), is better for an individual question, and there are three choices: A is better, B is better, and tie. Statistical results are listed in Table 6. The results indicate that our model achieves more satisfactory outcomes than LLaMA-2-7B, Alpaca-7B, Vicuna-7B, and K2 in the majority of subjective tasks, underscoring the effectiveness of our research for professional users. However, akin to the evaluation results with GPT-4, a gap persists between our model and GPT-3.5-turbo. Therefore, we will try additional strategies based on the evaluation results to further enhance the optimization of the model in the future.

### 5.3. The task of toponym extraction

For a geo-foundation model, it is essential to evaluate its effectiveness on well-established tasks, such as toponym extraction (Mai, 2024; Xie et al., 2023). Toponym extraction is a crucial natural language processing task that aims to recognize location names

**Table 7**  
Evaluation results on the task of toponym extraction.

Method	Harvey2017			Ju2016
	precision	recall	F-score	accuracy
TopoCluster	0.953	0.214	0.350	0.248
Mordecai	0.933	0.197	0.326	0.507
BB-GeoGPT	0.563	0.817	0.667	0.914

<b>Case1</b>
Text (label): Harvey Mayor : Message from <b>Port Aransas</b> : <b>Trailer park</b> a ' 100 - percent ' loss , search and rescue starts today
TopoCluster: Null
Mordecai: <b>Port Aransas</b>
BB-GeoGPT: Harvey, <b>Port Aransas</b> , <b>Trailer park</b>
<b>Case2</b>
Text (label): Gerardos in <b>Houston</b> on <b>Patton Street</b> . Can we say price gouging during HurricaneHarvey . Shame .. Shame !
TopoCluster: <b>Houston</b>
Mordecai: <b>Houston</b>
BB-GeoGPT: Gerardos, <b>Houston</b> , and <b>Patton Street</b> .

Fig. 7. The toponym extraction results of two cases obtained by the baselines and BB-GeoGPT.

(toponyms) from text, such as social media messages (Hu et al., 2023a; Karimzadeh et al., 2019; Wang, Hu, & Joseph, 2020). This task has garnered significant attention from researchers due to its valuable applications, especially in emergency situations. For example, social media messages, such as tweets, are often used by people during natural disasters, and accurately recognizing and geolocating these locations is critical for providing assistance to those in need (Hu et al., 2023a; Inkpen et al., 2017). In this section, we apply BB-GeoGPT to this task to investigate its performance.

In this experiment, we used two commonly used datasets for evaluating BB-GeoGPT: Harvey2017 (Mai et al., 2022b; Wang et al., 2020) and Ju2016 (Ju et al., 2016). The Harvey2017 dataset consists of 1,000 samples obtained and annotated from real tweets related to 'Harvey2017', 'HoustonFlood', and 'Hurricane Harvey'. The Ju2016 dataset contains sentences automatically extracted from web pages, with 5,500 samples and corresponding toponym annotations. A difference between the Harvey2017 and Ju2016 datasets is that Harvey2017 annotates all the toponyms in the text, whereas Ju2016 annotates parts of the toponyms contained in the text. To evaluate the performance of BB-GeoGPT on these two datasets, we follow the same evaluation metrics as in Wang et al. (2020). Specifically, precision, recall, and F-score are used for Harvey2017 (Eq. (2), (3), (4)), and the metric of accuracy is used for Ju2016, since Ju2016 only contains parts of the toponym labels in the samples (Eq. (5)).

$$\text{precision} = \frac{tp}{tp + fp} \quad (2)$$

$$\text{recall} = \frac{tp}{tp + fn} \quad (3)$$

$$F\text{-score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

$$\text{accuracy} = \frac{|\text{Annotated} \cap \text{Recognized}|}{|\text{Annotated}|} \quad (5)$$

where  $tp$  denotes the correctly identified toponyms,  $fp$  denotes incorrectly identified toponyms,  $tp + fp$  denotes all the recognized words by the models (i.e., *Recognized*), and  $tp + fn$  denotes all the annotated toponyms (i.e., *Annotated*).

Two representative geo-specific models in the task of toponym extraction, TopoCluster (DeLozier, Baldridge, & London, 2015) and Mordecai (Halterman, 2017), are used as the baselines. All the results are presented in Table 7. The results indicate that BB-GeoGPT achieves lower precision values but higher recall (accuracy) values than the baselines. To further analyze the results, the toponym extraction results of two cases obtained by the baselines and BB-GeoGPT are presented in Fig. 7. In case 1, TopoCluster obtained a 'Null' result, and Mordecai only recognized part of the toponyms. In case 2, both of them recognized only part of the toponyms. A common characteristic of TopoCluster and Mordecai is that most of their identified toponyms are correct, resulting in relatively high precision. However, they may miss some toponyms, leading to relatively low recall values. In contrast, BB-GeoGPT can recognize most of the potential toponyms, resulting in relatively high recall and accuracy values. However, BB-GeoGPT may also incorrectly identify some non-toponyms (see two cases), leading to relatively low precision. Therefore, considering the characteristics of BB-GeoGPT and the baselines, users can use them in different scenarios. For example, when recall is prioritized, BB-GeoGPT can be used to ensure most of the toponyms are identified.

**Table 8**  
Evaluation results on the task of temporal reasoning.

Method	TEMPREASON-L1	
	EM	MAE
<sup>a</sup> T5-L-NQ	26.1	20.6
<sup>a</sup> FLAN-T5-large	8.1	7.7
BB-GeoGPT	13.7	3.2

<sup>a</sup> Indicates that the results were cited from Tan et al. (2023).

#### 5.4. The task of temporal reasoning

Besides evaluating the performance of BB-GeoGPT in toponym extraction from the spatial aspect, its performance on temporal reasoning is also important, since the concept of time is crucial in the field of geographic information. In this section, BB-GeoGPT is applied to the task of temporal reasoning (Cai et al., 2023; Øhrstrøm & Hasle, 2007). Specifically, we used a benchmark dataset TEMPREASON-L1 (Tan et al., 2023), which is specifically designed to test the temporal reasoning capability of large language models, to evaluate the performance of BB-GeoGPT in temporal reasoning. TEMPREASON-L1 is based on a template of ‘What is the year  $x$  years before  $y$ ?’, where ‘ $x$ ’ and ‘ $y$ ’ are randomly chosen numbers. The dataset comprises 6,000 samples covering the time period from 1900 to 2040. Two representative large models in the task of temporal reasoning, the T5-L-NQ model (Raffel et al., 2020) and the FLAN-T5-large model (Wei et al., 2021), are used as baselines. Both models are fine-tuned on multiple datasets for different NLP tasks. Two metrics are used to evaluate the performance of different methods. The first metric is Exact Match (EM), a standard metric for question answering that measures the ratio of outputs exactly matching corresponding answers. Since the predicted results are numeric values (i.e., year), we also use Mean Absolute Error (MAE) as an evaluation metric following (Tan et al., 2023). All the results are presented in Table 8. The results show that BB-GeoGPT achieves better (higher) EM values than FLAN-T5-large but worse (lower) values than T5-L-NQ. Additionally, BB-GeoGPT obtains better MAE values than both baselines. These results suggest that BB-GeoGPT demonstrates comparable performance in temporal reasoning compared to existing methods.

## 6. Discussion

### 6.1. Further performance improvement

From the results presented in Section 5, BB-GeoGPT appears to be more robust in solving professional GIS problems from both objective and subjective perspectives compared to general-domain LLMs with similar parameter sizes (e.g., LLaMA-2-7B, Alpaca-7B, and Vicuna-7B), validating the effectiveness of the training strategy and the collected dataset. However, there are still two limitations.

- First, BB-GeoGPT still lags behind commercial models like GPT-3.5-turbo in certain aspects, indicating room for further enhancement in model training and data collection. To improve the performance of BB-GeoGPT, we plan to take two approaches. Firstly, we plan to train BB-GeoGPT on top of open-source general LLMs with larger model parameters, such as LLaMA-2-13B and LLaMA-2-70B (Touvron et al., 2023a). Larger language models generally exhibit more robust performance (Kaplan et al., 2020). One of the reasons GPT-3.5-turbo outperforms BB-GeoGPT is its 175B parameters, whereas BB-GeoGPT is trained on a model with 7B parameters. Secondly, collecting training data, especially instruction tuning data, of higher quality is another essential aspect. In the subjective task evaluation (Table 5 in Section 5.2), BB-GeoGPT shows a slight drawback in Accuracy and Clarity but a larger drawback in Completeness compared to GPT-3.5-turbo. This result indicates a need for further refinement in training, particularly in generating longer, more comprehensive responses. High-quality instruction tuning dataset is critical for this purpose (Ouyang et al., 2022). In this paper, most of the instructions for BB-GeoSFT are generated based on the LLaMA-2-7B-chat model through self-instruction (see Section 3.2.3). While using instructions generated by the LLaMA-2-7B-chat model is cost-effective and useful, it should be noted that LLaMA-2-7B-chat may not perform as well as more robust language models (such as GPT-4) or manual annotations by professionals in generating instruction data (Peng et al., 2023). Since many studies highlight instruction tuning as a critical process in training a language model (Zhang et al., 2023), collecting higher-quality instructions is a focus for future work.
- Secondly, in the evaluation of the subjective task (Table 5), BB-GeoGPT outperforms language models with similar sizes in most indexes, except for three values: the Completeness of Geodatabase compared to Vicuna-7B, and the Completeness of Remote sensing compared to Alpaca-7B and K2. The lower performance in Geodatabase and Remote sensing reflects the insufficient training data in these two domains. Therefore, collecting more training corpus in these fields is also a focus to improve BB-GeoGPT’s performance in solving professional problems in the geodatabase and remote sensing domains.

### 6.2. The challenge of deployment

While large language models offer excellent performance, their deployment poses challenges, including compute- and memory-intensive requirements, as well as some model limitations. The first challenge makes it difficult to deploy them on mobile and edge devices. To address this, we plan to deploy BB-GeoGPT using two approaches.

**Table 9**

Evaluation results based on the objective tasks by using 8-bit quantization on BB-GeoGPT.

Model	Spatial analysis	Geodatabase	Cartography	Remote sensing	Surveying	Average
BB-GeoGPT	0.583	0.529	0.638	0.711	0.720	<b>0.608</b>
BB-GeoGPT-8 bit	0.509	0.523	0.654	0.639	0.640	<b>0.573</b>

- The first approach is to use quantization on BB-GeoGPT, which can reduce computational and memory requirements and accelerate inference by representing model parameters with lower precision (Bai et al., 2022). This technique enables more efficient deployment of large models on various hardware platforms, including devices with low-performance CPUs and GPUs. However, a common issue with quantization is the potential performance degradation of models (Tao et al., 2022). In this section, we tested 8-bit quantization on BB-GeoGPT, which involves converting floating-point model parameters to 8-bit integers. During the inference process, BB-GeoGPT requires approximately 15 GB of GPU memory, while BB-GeoGPT-8 bit requires only 10 GB, demonstrating the effectiveness of quantization in reducing computational and memory requirements. However, the evaluation results listed in Table 9 show that the performance of BB-GeoGPT-8 bit is lower than BB-GeoGPT on most indexes, highlighting a challenge that needs to be addressed. Recent research has focused on accuracy-preserving quantization, which aims to enable quantization for LLMs while maintaining model performance (Xiao et al., 2023). Therefore, we will also consider deploying our BB-GeoGPT with this type of quantization.
- The second approach is to apply distillation learning to BB-GeoGPT. Distillation learning is a powerful technique used to compress large models (teacher models) into smaller ones (student models). The goal is to enable the student model to achieve similar performance to the teacher model but with lower computational resources and memory requirements (Liu et al., 2022b). Research has shown the effectiveness of distillation learning on large language models (LLMs) (Gu et al., 2023; Liang et al., 2023). We plan to train a student model of BB-GeoGPT using distillation learning to achieve effective deployment.

The second deployment difficulty is the limitations of language models themselves. Although the evaluation (Section 5) shows BB-GeoGPT to be effective in solving professional GIS problems, it still has limitations, such as common concerns of hallucination, toxicity, and stereotypes found in other language models (Deng et al., 2023; Xu, Jain, & Kankanhalli, 2024). Since accuracy is crucial in geographic information, it is important to train the model to follow user instructions helpfully and safely, and to output responses that avoid hallucination. Additionally, since BB-GeoGPT is trained on top of LLaMA-2-7B, it inherits some of its limitations, such as limited support for non-English languages. We plan to collect more instruction data (e.g., about safety control) and further fine-tune the model to enhance its robustness in these aspects before deploying it for users.

### 6.3. The development of large multimodal models

In this paper, we focus on the text modality and propose the BB-GeoGPT with GIS-specific knowledge capable of answering professional geospatial questions in textual form. However, in the GIS discipline, there are also many other important data resources with different modalities, such as remote sensing images and other geospatial big data (e.g., DEM and vector data). Therefore, an assistant model that can understand and extract geographic information from these data sources, and ultimately answer professional questions based on this information, could significantly benefit GIS users. In the field of computer science, there has been significant interest in developing Large Multimodal Models (LMMs) that can process and understand information from multiple modalities such as text, images, and voices (Wu et al., 2023b; Zhang et al., 2024c). These developments can provide valuable insights for advancing LMMs in the field of GIS. Generally, training LMMs requires intensive computational resources and more types of training data compared to training language models. As such, we plan to advance our work on LMMs in GIS in two stages of gradual progress.:

- The first step is to train a GIS LMM considering only two modalities: text containing geographic information and remote sensing images. This type of model can also be referred to as a large Vision Language Model (VLM) (Chen et al., 2023b). In the field of LMMs, there are many contributions that integrate modalities of text and vision. For example, LLaVA (Large Language and Vision Assistant) is an LMM that connects a vision encoder with a language model for general-purpose visual and language understanding (Liu et al., 2024). MiniGPT-4 aligns a visual encoder with Vicuna, which can write stories and poems inspired by given images (Zhu et al., 2023). Leveraging these previous contributions can be beneficial for our initial attempts, potentially reducing the time required for trial-and-error, as training and evaluating LMMs can be resource-intensive in terms of manpower and computing power.
- Building on the first step, the second step aims to train a GIS Large Multimodal Model (LMM) capable of processing spatial data from various modalities, such as DEMs and vector data. Designing appropriate spatial data encoders (Mai, Li, & Lao, 2023) tailored to the characteristics of these data types is crucial at this stage. To address this challenge, we plan to draw inspiration from current deep learning models used for encoding spatial data, such as points (Mac Aodha, Cole, & Perona, 2019; Mai et al., 2022a, 2023b, 2023d), polylines (Yu & Chen, 2022a, 2022b), polygons (Hu et al., 2023c; Mai et al., 2023e), and DEM (Zhang, Yu, & Zhu, 2022). Moreover, as spatial vector data (e.g., points, polylines, and areas) can be structured as graphs, utilizing large graph models could be a promising method for encoding them (Klemmer, Safir, & Neill, 2023). Another challenge is how to collect aligned labeled data to train such an LMM. In the first step, existing text-remote sensing image pairs (such as data in Hu et al., 2023d) can be utilized to align the vision encoder with the language model (e.g., BB-GeoGPT). However, automatically annotating a large volume of spatial data with different modalities remains a valuable problem to be addressed.

#### 6.4. The transferability of our approach to other specialized fields

Training Large Language Models (LLMs) in professional fields plays a significant role in their development, which has garnered increased attention in recent years. In this paper, we introduce BB-GeoGPT and elaborate on the process and strategy of collecting and processing data, as well as how to train BB-GeoGPT based on these datasets. While the focus of this paper is on the GIS domain, the primary principles used to construct training data can be applied to other specialized fields. Specifically, BB-GeoGPT undergoes two training phases: pretraining and supervised fine-tuning, and the training datasets are constructed accordingly for these two phases.

During the pretraining phase, the model is trained with an autoregressive language modeling objective. This phase is primarily designed to enhance the fundamental understanding and generation capabilities of models in professional domains. One key principle to follow in this stage is to collect a large volume of professional corpus that covers the knowledge users want the target model to process. Professional books, papers, and Wikipedia pages are recommended as pretraining data sources, as they contain abundant professional knowledge. Regarding the optimal pretraining data size, there is an evolving understanding of the relationship between model parameters and the amount of data that should be used. Kaplan et al. (2020) proposed precise power-law scalings to describe the relationship between model performance and training time, context length, dataset size, model size, and compute budget. For instance, with a 10-fold increase in the computational budget, it is advisable to expand the model's size by 5.5 times and increase the number of training tokens by 1.8 times. Hoffmann et al. (2022) suggested that the amount of training data should increase at the same rate as the size of the model. However, Touvron et al. (2023b) demonstrated that smaller models trained with large datasets can match or even outperform their larger counterparts. Additionally, it has been suggested that existing scaling laws may not always apply in situations where smaller models are trained for longer periods (Thaddeé, 2023). Zhang et al. (2024d) proposed TinyLlama, a model with 1.1B parameters but trained on approximately 3 trillion tokens. This dataset size is much larger than what is suggested by the scaling law (Hoffmann et al., 2022), yet TinyLlama outperforms models with similar parameter sizes. Therefore, if users aim to deploy a relatively small language model (e.g., 1B parameters) on mobile devices, it is suggested to refer to Zhang et al. (2024d) to collect a relatively large pretraining dataset.

Secondly, the supervised fine-tuning phase aims to enhance the chat and instruction-following abilities of language models. Unlike the pretraining phase, supervised fine-tuning requires high-quality instruction data, which is generally smaller in size than the pretraining data. To generate high-quality instruction data, users in other specialized fields can refer to the self-instruct strategy used in this paper. They can use a robust commercial LLM (e.g., GPT-4) or an open-source LLM (e.g., LLaMA-2-7B-chat, as used in this paper) to generate instructions based on given plain text (such as the corpus collected from the pretraining phase). Since the fine-tuning process is crucial, a key principle is trying to control the style of the output in the instruction data to align with professional responses in corresponding fields.

### 7. Conclusion

In this paper, we introduce BB-GeoGPT, a large language model with GIS-specific knowledge capable of answering professional geospatial questions. We have curated professional training data, including model pretraining data BB-GeoPT and supervised fine-tuning data BB-GeoSFT, and we outline the methods used to construct these datasets. Simultaneously, we create the evaluation dataset BB-GeoEval from both objective and subjective perspectives, providing a comprehensive means to assess the professional understanding ability of large language models in GIS and related domains. On the collected evaluation datasets, BB-GeoGPT outperforms the original LLaMA-2-7B model and two general-purpose open-sourced LLMs (Alpaca-7B and Vicuna-7B), confirming the effectiveness of our research. Overall, our data collection strategy and the amassed data can serve as a foundation for advancing research on Large Language Models (LLMs) in the GIS domain, fostering further development.

However, there are also limitations. Firstly, despite our extensive collection of training data, a performance gap still exists between our model and one of the most outstanding commercial close-sourced LLMs, GPT-3.5-turbo. Given that GIS encompasses knowledge from multiple subjects, such as geography, computer science, and geology, incorporating professional knowledge from these relevant subjects may enhance the performance of the models. Additionally, subjective evaluations suggest a relatively larger gap between our model and GPT-3.5-turbo in terms of completeness. Our ongoing efforts will prioritize enhancing the quality and length of model responses. A practical avenue for improvement involves refining the quality of instruction tuning data. As this paper utilizes the open-source LLaMA-2-7B-chat model to generate self-instructed instructions, employing a more robust and long-text language model, such as GPT-4, may lead to improved performance.

### CRediT authorship contribution statement

**Yifan Zhang:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Zhiyun Wang:** Validation, Software, Methodology, Investigation, Data curation. **Zhengting He:** Validation, Data curation. **Jingxuan Li:** Validation, Data curation. **Gengchen Mai:** Writing – review & editing. **Jianfeng Lin:** Supervision, Project administration. **Cheng Wei:** Validation, Data curation. **Wenhai Yu:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

BB-GeoEval is available with a DOI at <https://github.com/AGI-GIS/BB-GeoGPT>.

## Acknowledgments

The project was supported by the National Natural Science Foundation of China (42371446 and 42071442), Natural Science Foundation of Hubei Province, China (2024AFD412), and by the Fundamental Research Funds for the Central Universities, China, China University of Geosciences (Wuhan) (No. CUG170640). This research was also supported by Meituan.

## References

- Al-Olimat, H. S., et al. (2017). Location name extraction from targeted text streams using gazetteer-based statistical language models. arXiv preprint arXiv: 1708.03105.
- Andreas, J. (2022). Language models as agent models. arXiv preprint arXiv:2212.01681.
- Bai, H., et al. (2022). Towards efficient post-training quantization of pre-trained language models. *Advances in Neural Information Processing Systems*, 35, 1405–1418.
- Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE joint conference on digital libraries* (pp. 1–10). IEEE.
- Brown, T., et al. (2020). Language models are few-shot learners. In *Advances in neural information processing systems*: vol. 33, (pp. 1877–1901).
- Cai, L., et al. (2023). HyperQuaternionE: A hyperbolic embedding model for qualitative spatial and temporal reasoning. *GeoInformatica*, 27(2), 159–197.
- Caines, A., et al. (2023). On the application of large language models for language teaching and assessment technology. arXiv preprint arXiv:2307.08393.
- Chang, Y., et al. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.
- Chen, W., et al. (2013). A synergistic framework for geographic question answering. In *2013 IEEE seventh international conference on semantic computing* (pp. 94–99). IEEE.
- Chen, Z., et al. (2023a). Exploring the potential of large language models (llms) in learning on graphs. arXiv preprint arXiv:2307.03393.
- Chen, F.-L., et al. (2023b). Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1), 38–56.
- Chiang, W.-L., et al. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality.
- Cui, Y., Yang, Z., & Yao, X. (2023). Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177.
- DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. Vol. 29, In *Proceedings of the AAAI conference on artificial intelligence*. (1).
- Deng, C., et al. (2023). Learning A foundation language model for geoscience knowledge understanding and utilization. arXiv preprint arXiv:2306.05064.
- Derczynski, L., Bontcheva, K., & Roberts, I. (2016). Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers* (pp. 1169–1179).
- Dettmers, T., et al. (2023). Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.
- Ding, N., et al. (2023a). Enhancing chat language models by scaling high-quality instructional conversations. arXiv preprint arXiv:2305.14233.
- Ding, N., et al. (2023b). Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5, 220–235.
- Dou, F., et al. (2023). Towards artificial general intelligence (agi) in the internet of things (iot): Opportunities and challenges. arXiv preprint arXiv:2309.07438.
- D'Souza, J., & Ng, V. (2015). Sieve-based spatial relation extraction with expanding parse trees. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 758–768).
- DUOMO (2023). TransGPT. <https://github.com/DUOMO/TransGPT>.
- Gao, Y., et al. (2023). Chat-rec: Towards interactive and explainable llms-augmented recommender system. arXiv preprint arXiv:2303.14524.
- Gema, A., et al. (2023). Parameter-efficient fine-tuning of llama for the clinical domain. arXiv preprint arXiv:2307.03042.
- Gu, Y., et al. (2023). Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*.
- Hadi, M. U., et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. TechRxiv.
- Halterman, A. (2017). Mordecai: Full text geoparsing and event geocoding. *Journal of Open Source Software*, 2(9), 91.
- Hoffmann, J., et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
- Houlsby, N., et al. (2019). Parameter-efficient transfer learning for NLP.
- Hu, X., et al. (2021a). GazPNE2: A general and annotation-free place name extractor for microblogs fusing gazetteers and transformer models.
- Hu, E. J., et al. (2021b). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
- Hu, X., et al. (2022). GazPNE: annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. *International Journal of Geographical Information Science*, 36(2), 310–337.
- Hu, Y., et al. (2023a). 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.
- Hu, X., et al. (2023b). Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56(5), 1–37.
- Hu, Y., et al. (2023c). PolyBuilding: Polygon transformer for building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199, 15–27.
- Hu, Y., et al. (2023d). Rsgpt: A remote sensing vision language model and benchmark. arXiv preprint arXiv:2307.15266.
- Huang, J., & Chang, K. C.-C. (2022). Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.
- Huang, W., et al. (2022). Inner monologue: Embodied reasoning through planning with language models. arXiv preprint arXiv:2207.05608.
- Huang, Q., et al. (2023a). Benchmarking large language models as AI research agents. arXiv preprint arXiv:2310.03302.
- Huang, Q., et al. (2023b). Lawyer LLaMA technical report. arXiv preprint arXiv:2305.15062.
- Inkpen, D., et al. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49, 237–253.
- Jeong, J. (2011). pdfminer. <https://github.com/jaepil/pdfminer3k>.
- Ju, Y., et al. (2016). Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In *Knowledge engineering and knowledge management: 20th international conference, EKAW 2016, bologna, Italy, November 19–23, 2016, proceedings*. Vol. 20 (pp. 353–367). Springer.
- Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- Karimzadeh, M., et al. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1), 118–136.
- Katz, D. M., et al. (2023). Gpt-4 passes the bar exam. Available at SSRN 4389233.
- Kedzie, C., McKeown, K., & Diaz, F. (2015). Predicting salient updates for disaster summarization. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers)* (pp. 1608–1617).
- Klemmer, K., Safir, N. S., & Neill, D. B. (2023). Positional encoder graph neural networks for geographic data. In *International conference on artificial intelligence and statistics* (pp. 1379–1389). PMLR.
- Latif, E., et al. (2023). Artificial general intelligence (AGI) for education. arXiv preprint arXiv:2304.12479.
- Lee, G.-G., et al. (2023). Multimodality of AI for education: Towards artificial general intelligence. arXiv preprint arXiv:2312.06037.

- Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning.
- Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation.
- Li, Z., & Ning, H. (2023). Autonomous GIS: the next-generation AI-powered GIS. arXiv preprint arXiv:2305.06453.
- Li, X., et al. (2022a). Information geography: The information revolution reshapes geography. *Science China Earth Sciences*, 65(2), 379–382.
- Li, J., et al. (2022b). Unified named entity recognition as word-word relation classification. Vol. 36, In *Proceedings of the AAAI conference on artificial intelligence* (10), (pp. 10965–10973).
- Li, C., et al. (2023). Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the thirty-seventh conference on neural information processing systems*.
- Liang, C., et al. (2023). Less is more: Task-aware layer-wise distillation for language model compression. In *International conference on machine learning* (pp. 20852–20867). PMLR.
- Liu, X., Chen, H., & Xia, W. (2022a). Overview of named entity recognition. *Journal of Contemporary Educational Research*, 6(5), 65–68.
- Liu, C., et al. (2022b). Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1001–1011).
- Liu, Z., et al. (2023). Transformation vs tradition: Artificial general intelligence (AGI) for arts and humanities. arXiv preprint arXiv:2310.19626.
- Liu, H., et al. (2024). Visual instruction tuning. In *Advances in neural information processing systems: vol. 36*.
- Lopez, P. (2008). GROBID. <https://github.com/kermitt2/grobid>.
- Lu, G., et al. (2023). AGI for agriculture. arXiv preprint arXiv:2304.06136.
- Mac Aodha, O., Cole, E., & Perona, P. (2019). Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9596–9606).
- Mai, G. (2024). Geo-foundation model. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*.
- Mai, G., Li, Z., & Lao, N. (2023). Spatial representation learning in GeoAI. In *Handbook of geospatial artificial intelligence* (pp. 99–120). CRC Press.
- Mai, G., et al. (2020a). Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model. In *Geospatial technologies for local and regional development: proceedings of the 22nd AGILE conference on geographic information science* 22 (pp. 21–39). Springer.
- Mai, G., et al. (2020b). SE-KGE: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24(3), 623–655.
- Mai, G., et al. (2021). Geographic question answering: Challenges, uniqueness, classification, and future directions. *AGILE: GIScience Series*, 2, 8.
- Mai, G., et al. (2022a). A review of location encoding for GeoAI: methods and applications. *International Journal of Geographical Information Science*, 36(4), 639–673.
- Mai, G., et al. (2022b). Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th ACM SIGSPAtAL international conference on advances in geographic information systems* (pp. 1–4).
- Mai, G., et al. (2023b). CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. arXiv preprint arXiv:2305.01118.
- Mai, G., et al. (2023c). On the opportunities and challenges of foundation models for geospatial artificial intelligence. arXiv preprint arXiv:2304.06798.
- Mai, G., et al. (2023d). Sphere2Vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 439–462.
- Mai, G., et al. (2023e). Towards general-purpose representation learning of polygonal geometries. *GeoInformatica*, 27(2), 289–340.
- Mani, I., et al. (2010). SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44, 263–280.
- Manvi, R., et al. (2023). Geollm: Extracting geospatial knowledge from large language models. arXiv preprint arXiv:2310.06213.
- Min, B., et al. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), 1–40.
- Mooney, P., et al. (2023). Towards understanding the geospatial skills of ChatGPT: Taking a geographic information systems (GIS) exam. In *Proceedings of the 6th ACM SIGSPAtAL international workshop on AI for geographic knowledge discovery* (pp. 85–94).
- Neubig, G., et al. (2011). Safety information mining—What can NLP do in a disaster. In *Proceedings of 5th international joint conference on natural language processing* (pp. 965–973).
- Nori, H., et al. (2023). Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
- Øhrstrøm, P., & Hasle, P. (2007). *Vol. 57, Temporal logic: From ancient ideas to artificial intelligence*. Springer Science & Business Media.
- OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.
- Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Peng, B., et al. (2023). Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.
- Perez, E., Kiela, D., & Cho, K. (2021). True few-shot learning with language models. In *Advances in neural information processing systems: vol. 34*, (pp. 11054–11070).
- Pustejovsky, J., et al. (2015). Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th international workshop on semantic evaluation* (pp. 884–894). ACL.
- Qi, Z., et al. (2023). Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. arXiv preprint arXiv:2308.10173.
- Qiu, Q., et al. (2022). Spatially oriented convolutional neural network for spatial relation extraction from natural language texts. *Transactions in GIS*, 26(2), 839–866.
- Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1), 5485–5551.
- Richards, T. (2023). Auto-gpt: An autonomous gpt-4 experiment.
- Roberts, J., et al. (2023). GPT4GEO: How a language model sees the world's geography. arXiv preprint arXiv:2306.00020.
- Savelka, J., et al. (2023). Explaining legal concepts with augmented large language models (GPT-4). arXiv preprint arXiv:2306.09525.
- Singh, I., et al. (2023). Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE international conference on robotics and automation* (pp. 11523–11530). IEEE.
- Singhal, K., et al. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.
- Snæbjarnarson, V., et al. (2022). A warm start and a clean crawled corpus—A recipe for good language models. arXiv preprint arXiv:2201.05601.
- Song, P. (2023). Lawgpt. <https://github.com/pengxiao-song/LaWGPT>.
- Tan, Q., Ng, H. T., & Bing, L. (2023). Towards benchmarking and improving the temporal reasoning capability of large language models. arXiv preprint arXiv:2306.08952.
- Tao, C., et al. (2022). Compression of generative pre-trained language models via quantization. arXiv preprint arXiv:2203.10705.
- Taori, R., et al. (2023a). Vol. 3, *Alpaca: A strong, replicable instruction-following model* (6), (p. 7). Stanford Center for Research on Foundation Models, <https://crfm.stanford.edu/2023/03/13/alpaca.html>.
- Taori, R., et al. (2023b). Stanford alpaca: An instruction-following llama model.
- Thaddée, Y. T. (2023). Chinchilla's death. <https://espadrine.github.io/blog/posts/chinchilla-sdeath.html>.
- Thirunavukarasu, A. J., et al. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940.
- Touvron, H., et al. (2023a). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- Touvron, H., et al. (2023b). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- Vaswani, A., et al. (2017). Attention is all you need. In *Advances in neural information processing systems: vol. 30*.
- Waisberg, E., et al. (2023). GPT-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science* (1971-), 1–4.

- Wang, E. J. (2023). alpaca-lora. <https://github.com/tloen/alpaca-lora>.
- Wang, J., Hu, Y., & Joseph, K. (2020). NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Transactions in GIS*, 24(3), 719–735.
- Wang, F., Li, P., & Zhu, Q. (2022a). A hybrid model of classification and generation for spatial relation extraction. In *Proceedings of the 29th international conference on computational linguistics* (pp. 1915–1924).
- Wang, S., et al. (2021). Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint [arXiv:2112.12731](https://arxiv.org/abs/2112.12731).
- Wang, Y., et al. (2022b). Self-instruct: Aligning language model with self generated instructions. arXiv preprint [arXiv:2212.10560](https://arxiv.org/abs/2212.10560).
- Wang, Y., et al. (2023a). Aligning large language models with human: A survey. arXiv preprint [arXiv:2307.12966](https://arxiv.org/abs/2307.12966).
- Wang, Z., et al. (2023b). Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. arXiv preprint [arXiv:2302.01560](https://arxiv.org/abs/2302.01560).
- Wei, J., et al. (2021). Finetuned language models are zero-shot learners. arXiv preprint [arXiv:2109.01652](https://arxiv.org/abs/2109.01652).
- Wei, J., et al. (2022). Emergent abilities of large language models. arXiv preprint [arXiv:2206.07682](https://arxiv.org/abs/2206.07682).
- Wu, S., et al. (2023a). Bloomberggpt: A large language model for finance. arXiv preprint [arXiv:2303.17564](https://arxiv.org/abs/2303.17564).
- Wu, J., et al. (2023b). Multimodal large language models: A survey. In *2023 IEEE international conference on big data (bigData)* (pp. 2247–2256). IEEE.
- Xiao, G., et al. (2023). Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning* (pp. 38087–38099). PMLR.
- Xie, Y., et al. (2023). Geo-foundation models: Reality, gaps and opportunities (vision paper). In *Proceedings of the 31st ACM SIGSPATIAL international conference on advances in geographic information systems*.
- Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint [arXiv:2401.11817](https://arxiv.org/abs/2401.11817).
- Xu, C., et al. (2023). Baize: An open-source chat model with parameter-efficient tuning on self-chat data. arXiv preprint [arXiv:2304.01196](https://arxiv.org/abs/2304.01196).
- Yu, W., & Chen, Y. (2022a). Data-driven polyline simplification using a stacked autoencoder-based deep neural network. *Transactions in GIS*, 26(5), 2302–2325.
- Yu, W., & Chen, Y. (2022b). Filling gaps of cartographic polylines by using an encoder-decoder model. *International Journal of Geographical Information Science*, 36(11), 2296–2321.
- Yue, S., et al. (2023). Disc-lawllm: Fine-tuning large language models for intelligent legal services. arXiv preprint [arXiv:2309.11325](https://arxiv.org/abs/2309.11325).
- Zhang, Y., Wei, C., He, Z., & Yu, W. (2024a). GeoGPT: An assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, [ISSN: 1569-8432] 131, Article 103976. <http://dx.doi.org/10.1016/j.jag.2024.103976>, <https://www.sciencedirect.com/science/article/pii/S1569843224003303>.
- Zhang, Y., Yu, W., & Zhu, D. (2022). Terrain feature-aware deep learning network for digital elevation model superresolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 189, 143–162.
- Zhang, W., Zhang, Y., Deng, Y., Zhang, W., Lin, J., Huang, B., et al. (2024b). Ta-Adapter: Enhancing few-shot CLIP with task-aware encoders. *Pattern Recognition*, [ISSN: 0031-3203] 153, Article 110559. <http://dx.doi.org/10.1016/j.patcog.2024.110559>, <https://www.sciencedirect.com/science/article/pii/S0031320324003108>.
- Zhang, S., et al. (2023). Instruction tuning for large language models: A survey. arXiv preprint [arXiv:2308.10792](https://arxiv.org/abs/2308.10792).
- Zhang, R., et al. (2023c). Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15211–15222).
- Zhang, D., et al. (2024c). Mmlms: Recent advances in multimodal large language models. arXiv preprint [arXiv:2401.13601](https://arxiv.org/abs/2401.13601).
- Zhang, P., et al. (2024d). Tinyllama: An open-source small language model. arXiv preprint [arXiv:2401.02385](https://arxiv.org/abs/2401.02385).
- Zhao, H., et al. (2023a). Ophtha-llama2: A large language model for ophthalmology. arXiv preprint [arXiv:2312.04906](https://arxiv.org/abs/2312.04906).
- Zhao, W. X., et al. (2023b). A survey of large language models. arXiv preprint [arXiv:2303.18223](https://arxiv.org/abs/2303.18223).
- Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint [arXiv:2306.05685](https://arxiv.org/abs/2306.05685).
- Zhu, D., et al. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint [arXiv:2304.10592](https://arxiv.org/abs/2304.10592).