LLM-Driven Socio-Economic Estimation for Visegrád Countries

Donghyun Ahn
Max Planck Institute for Security and Privacy
Bochum, Germany
donghyun.ahn@mpi-sp.org

Sumin Lee
Korea Advanced Institute of Science and Technology
Daejeon, Korea
dlsumn03@kaist.ac.kr

Donggyu Lee
Korea Advanced Institute of Science and Technology
Daejeon, Korea
donggyu.lee@kaist.ac.kr

Jungwon Kim
Korea Advanced Institute of Science and Technology
Daejeon, Korea
jungwonkim126@kaist.ac.kr

Sungwon Han
Korea Advanced Institute of Science and Technology
Daejeon, Korea
lion4151@kaist.ac.kr

Seungeon Lee
Korea Advanced Institute of Science and Technology
Daejeon, Korea
archon159@kaist.ac.kr

Younhyung Chae
Seoul National University
Seoul, Korea
yhchae0811@snu.ac.kr

Jihee Kim
Korea Advanced Institute of Science and Technology
Daejeon, Korea
jiheekim@kaist.ac.kr

Meeyoung Cha
Max Planck Institute for Security and Privacy
Bochum, Germany
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
mia.cha@mpi-sp.org

Abstract

Moving beyond traditional surveys, combining heterogeneous data sources with AI-driven inference models brings new opportunities to examine developmental conditions across expansive geographic areas. This research presents GeoSEE, a method that estimates socio-economic indicators such as poverty and population at the subnational level using a large language model. Leveraging a diverse set of information sources and analytical modules, such as satellite imagery and computer vision algorithms, the model efficiently selects the most appropriate components for estimating a given indicator and country. This choice is guided by the language model's prior knowledge, which functions similarly to the insights of a domain expert. The system then computes target indicators through in-context learning, synthesizing information from selected modules into natural language paragraphs. An extensive evaluation across Visegrád countries (i.e., Hungary, Slovakia, Poland, and Czech Republic) with various stages of development and multiple indicators demonstrates that our method accurately predicts estimates in both unsupervised and low-shot contexts. Its reliability in data-scarce environments, combined with cost-effective computation and scalability, makes it a viable tool for large-scale, fine-grained monitoring.

CCS Concepts

• **Applied computing** → **Economics**; • **Security and privacy** → Social aspects of security and privacy.

GoodIT '25, September 03–05, 2025, Antwerp, Belgium

ACM ISBN 979-8-4007-2089-5/2025/09
https://doi.org/10.1145/3748699.3749803

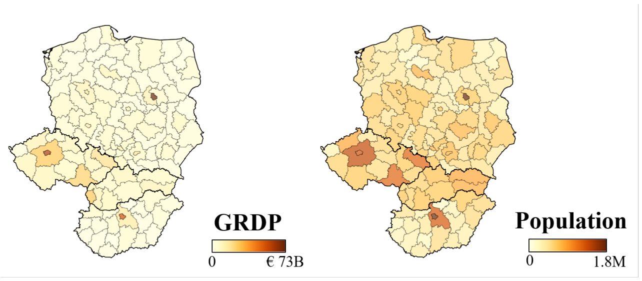

Figure 1: GRDP (billion euros) and population (millions) across subnational regions in the Visegrád countries.

Keywords

Socio-economic estimation, LLM, Satellite imagery

ACM Reference Format:
Donghyun Ahn, Sumin Lee, Donggyu Lee, Jungwon Kim, Sungwon Han, Seungeon Lee, Younhyung Chae, Jihee Kim, and Meeyoung Cha. 2025. LLM-Driven Socio-Economic Estimation for Visegrád Countries. In *International Conference on Information Technology for Social Good (GoodIT '25), September 03–05, 2025, Antwerp, Belgium.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3748699.3749803

1 Introduction

Policy making and evaluation can greatly benefit from timely, granular data on subnational development conditions. Localized data enables governments and international organizations to precisely target resources and track progress toward policies. While certain indicators such as regional GDP or population are available at the subnational level in some countries (Figure 1), many other important indicators—such as labor market dynamics, educational attainment, or health outcomes—are measured only at the national scale and updated infrequently. This lack of spatial and temporal granularity limits the capacity to evaluate local disparities.

In response to these data gaps, recent literature has begun exploring new data sources, such as Wikipedia text [17], street view images [13], mobile phone adoption patterns [16], and high-resolution satellite imagery [2, 3, 7]. These publicly available data sources enable the generation of fine-grained estimates of socio-economic indicators across space and time, even in regions where traditional data collection is limited.

However, most existing models are narrowly specialized, focusing on one or two socio-economic indicators, such as population density, gross domestic product, or Gini coefficient, employing a restricted range of data types [9, 10, 14]. Developing a cross-country model that accommodates *multiple* indicators and leverages diverse non-traditional data sources remains a significant methodological challenge. One key reason is the considerable variability in data availability and contextual factors across different regions and indicators (Figure 1). Moreover, each data type requires tailored methodologies to ensure accurate predictions, demanding specialized expertise and substantial resources [8]. This intensive need for domain knowledge restricts the scalability and multimodality of existing models.

We introduce GeoSEE, a holistic method for estimating a range of socio-economic indicators using a unified pipeline based on a large language model (LLM). The core concept of our approach is an efficient feature selection of heterogeneous data sources and analytical modules to estimate socio-economic indicators [12]. Feature selection involves identifying associations between input data and target labels. Conventionally, this has been done either through data-driven methods that require a substantial amount of ground-truth labels or through expert-guided selection; both of which are resource intensive. To address this, we leverage the extensive textual knowledge and reasoning capabilities of modern LLMs [1, 4], allowing them to act as domain experts and select relevant features from heterogeneous data sources to predict meaningful indicators. As a result, our method only requires natural language descriptions of the target indicator and features, making it applicable even in settings with limited structured training data, such as cross-country or region-level analyses.

We present evaluation results that demonstrate the effectiveness of our approach. Our evaluation compares various existing methods on the Visegrád countries, which represent different stages of economic development. The results highlight the strong predictive performance of GeoSEE, which consistently outperforms other methods. These preliminary findings suggest that the model's outputs are both reliable and broadly applicable in the real world. By providing a scalable framework for timely and detailed social analysis, we hope our approach supports informed and responsive policy decisions.

2 Methodology

2.1 Problem Statement and Overview

Problem definition. Let $\mathcal{D} = \{\mathbf{d}_i\}_{i=1}^{N}$ be a collection of N spatial regions (districts) covering a country. Our model leverages diverse datasets such as high-resolution satellite imagery and regional metadata to estimate socio-economic indicators, even when ground-truth labels are scarce. First we consider a few-shot setting, where a small subset of regions come with labels (i.e., k labeled "shots"), denoted by $\mathcal{D}_l = \{(\mathbf{d}_i, y_i)\}_{i=1}^{k}$, while the remaining regions are unlabeled and represented as $\mathcal{D}_{ul} = \{\mathbf{d}_i\}_{i=k+1}^{N}$, with $k \leq N$. The goal of GeoSEE is to predict the socio-economic label y_i for each region $\mathbf{d}_i$ in the entire dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_{ul}$.

Figure 2 illustrates the flow of our framework; it has two steps. In Step 1, given a list of information modules, the model selects the modules relevant to a specific country and indicator, based on the LLM's prior knowledge, to make predictions for the target geographic regions. The selected modules are then applied to the target regions to extract task-specific information, in the text format using a predefined template (Section 2.2). After extracting text descriptions for each region in the dataset, in Step 2, the model leverages in-context learning by providing generated sample paragraphs of a few other regions as well as its own (Section 2.3). These regions are selected by our strategy designed to provide both detailed comparisons of similar regions and broader insights from the overall distribution of labels, while keeping the input text within the prompt limit.

2.2 Step 1: Task Information Extraction

Module list. GeoSEE has a flexible modular design and can integrate new data sources and functions to address evolving needs. It employs a range of internal information modules to compute socio-economic labels, dynamically selecting and combining relevant modules based on the available data. This selection process is guided by the LLM's prior knowledge of geospatial information and hence supports efficient and context-aware computations. Here are the modules we utilize:

- `get_address`: Retrieves the address of a given region.
- `get_area`: Retrieves the area size of a given region.
- `get_night_light`: Retrieves the nightlight intensity of a given region.
- `get_landcover_ratio`: Includes a set of modules that count the number of pixels that cover each of the target landcover classes (e.g., 'road', 'agricultural') and return the ratio of this count to the total number of pixels in the region's total image set.
- `get_landuse_sum`: Classifies satellite image tiles into land use categories (e.g., 'factory', 'residential') and counts the number of tiles belonging to each class within a given region.
- `get_poi_num`: Retrieves the total count of specified points of interest (e.g., 'hospital', 'clinic') within a given region.
- `get_distance_to_nearest_target`: Includes a set of modules that measure the distance from a given region to each of the target class entities (e.g., 'airport', 'port').

Module selection. For each query, GeoSEE selects pertinent modules via the prompt. This prompt is an instruction for LLM to generate a response to the module selection results, consisting of a module description and a target task description, as shown in Figure 3. *Module description* includes functional specifications along with the input parameters it requires, for example: "`get_area(Loc): Get the area size of a given location's region.`" *Task description* states the indicator and the target country, for example, "`what information is appropriate to infer Poland's GRDP?`"

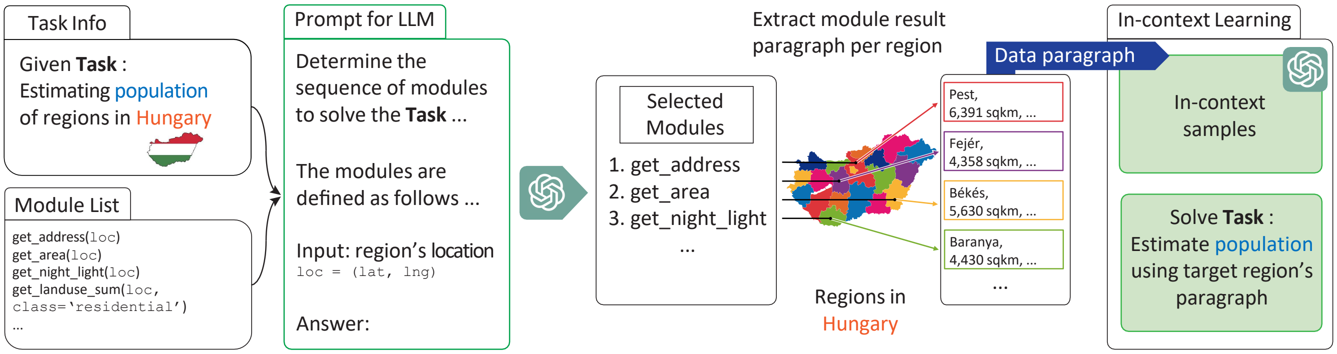

Figure 2: Illustration of GeoSEE. First presented with a task information and a description of the information module list, GeoSEE selects an informative set of modules via the LLM. With these selected modules, GeoSEE extracts information about each region within the country and generates a descriptive paragraph explaining the region. The paragraph, combined with our selection strategy, is then used to estimate the target indicator through in-context learning.

```
Given a modular set, determine the sequence of modules that
can be executed with inputs to solve the question given,
following the format below.

Format for response:
1. MODULE 1
2. MODULE 2
...

The modules are defined as follows:
<Module Description>
Question:
<Task Description>

Input:
- Location of the region - [Loc]
Answer:
```

Figure 3: Prompt for module selection in GeoSEE.

The model processes the given prompt and generates potential candidate modules. LLMs inherently construct diverse logical pathways, each consisting of a unique combination of modules. To ensure reliable selection, we repeat this process over ten iterations, identifying modules that are recommended at least five times. This approach aligns with the principle of self-consistency, which suggests that frequently occurring outcomes are likely to be more effective [18]. The selected modules are applied to each region in a target country, and the retrieved information is serialized into text using a predefined template. The resulting text forms a comprehensive paragraph that represents the key features of the region:

$$\text{Serialize}(f_1, \cdots, f_m, r_1, \cdots, r_m) = \text{“ } f_1 \text{ is } r_1. \ \cdots \ f_m \text{ is } r_m.\text{”} \qquad (1)$$

where $f_1, \cdots, f_m$ are descriptions of the selected modules, and $r_1, \cdots, r_m$ are the results obtained from each module.

2.3 Step 2: Estimation via In-Context Learning

Based on the text-format paragraph descriptions for each region, LLM estimates the region's target indicator via in-context learning. Given the LLM's known limitations in precise numerical reasoning [15], we recast the task as a five-way classification problem rather than having the model directly regress raw socio-economic values, for performance improvement. We ask it to assign each region to one of the following five classes $c \in C$:

$$C = \{\text{Very Low}, \ \text{Low}, \ \text{Medium}, \ \text{High}, \ \text{Very High}\}.$$

Given the region descriptions, we construct an in-context prompt containing ten "shot" examples (i.e., $n_{\text{shot}} = 10$). Each shot is a region description paired with its true value, which we discretize into one of the five classes to form our demonstration set. To ensure that the class assignments of the few-shot samples are well aligned with the ground-truth distribution, we perform class discretization under the following two conditions.

(1) Each of the five classes must be represented by at least one sample in the demonstration set, ensuring class coverage.
(2) To preserve the empirical skew of the true value distribution, we allocate a class to only a few samples when their values are extremely small or large (e.g., assigning only the top region to the "Very High" class when it significantly exceeds the rest).

To satisfy these conditions, we leverage an LLM to automatically propose class splits. Specifically, we provide the LLM with the true values of each few-shot region as input, and prompt it to generate n_{sample} candidate splits that adhere to the above conditions. Then, we quantify the LLM-generated classifications by referencing the sample information: for each class, we compute a discrete *prototype* score based on the average label value of the few-shot examples assigned to that class. Using these prototype scores, we train a scoring function that enables us to reproduce the classification results in a more continuous manner.

Once the in-context prompt is constructed, the LLM classifies each test region into one of the five classes. We then extract one representative region per class, ordered by ascending ground-truth values, resulting in a set of five regions with an inherent ground-truth ranking. We train our scoring function f by maximizing the Spearman rank correlation between the model's predicted scores for these representative regions and their true class ordering. This

Algorithm 1: Regional estimation via in-context learning

```
Input  : Large language model $F$, score estimation function
         $f$, unlabeled dataset $\mathcal{D}_{ul}$, number of shots $n_{\text{shot}}$,
         labeled dataset $\mathcal{D}_l$ with label $Y_l$, class candidate $C$,
         a collection of $n_{\text{sample}}$ sets $C_{\text{shot}}$, each element of
         $C_{\text{shot}}$ containing $n_{\text{shot}}$ label-class pairs (i.e.,
         $C_{\text{shot}} = \{((y_i, c_i)|_{i=1}^{n_{\text{shot}}})|(y_i, c_i) \in Y_l \times C\}$), number
         of iterations $n_{\text{iter}}$, a set of sample class label $\mathcal{D}_l$, a
         set of results from selected modules $\mathcal{R}$
Output : Score estimation result $\mathcal{S}$
$\mathcal{S} \leftarrow \emptyset, i \leftarrow 0$
$\mathcal{P} \leftarrow \{\mathbf{d} : \text{RegionDescription } (\mathcal{D}_l, \mathcal{R}, \mathbf{d}) \text{ for } \mathbf{d} \in \mathcal{D}_{ul}\}$
while $C_{shot} \neq \emptyset$ do
    $c_{\text{sample}} \leftarrow \text{Sample}(C_{\text{shot}}, 1)$
    $prototype \leftarrow \{c : \text{Avg}(\{y_i|(y_i, c_i) \in c_{\text{sample}} \text{ and } c = c_i\})\}$
    while $i < n_{iter}$ do
        $\mathcal{S}_{\text{temp}} \leftarrow \{\mathbf{d} : F(\text{target} = \mathbf{d}, \text{in-context} = \mathcal{P}[\mathbf{d}],$
                      $\text{modules} = \mathcal{R}, \text{class} = C) \in C\}$
        $Loss_{\text{order}} \leftarrow \text{Spearman}(f(\mathbf{d}), prototype(\mathcal{S}_{\text{temp}}[\mathbf{d}]))$
        Train $f$ by maximizing $Loss_{\text{order}}$
        $i \leftarrow i + 1$
    end
end
$\mathcal{S} \leftarrow \text{map}(f, \mathcal{D}_{ul})$
```

encourages f to preserve the relative ordering of socio-economic values across all regions. For each class split, this training process was repeated n_{iter} times. The number of iterations n_{iter} is set to 10.

Algorithm 1 outlines how the model uses in-context learning to infer scores for a given indicator for regions in the target country. The framework is built on *o4-mini* LLM. For the few-shot setting, the number of shots n_{shot} is set to 10. We also set the number of sampling iterations n_{sample} to 10, meaning that we evaluate the model across 10 different class compositions from the same shot.

3 Evaluation

3.1 Data and Implementation

The datasets used in this study encompass daytime and nighttime satellite imagery, along with four socioeconomic indicators from Poland, Czech republic, Slovakia, and Hungary. The daytime data, sourced from WorldView-2/3 and GeoEye and captured between 2015 and 2019, include 3,556,380 images, each with a spatial resolution of 2.4 meter per pixel and a size of 256x256 pixels.

Nighttime data was sourced from the annual global Visible Infrared Imaging Radiometer Suite (VIIRS) nightlights provided by Earth Observation Group (EOG) at a spatial resolution of 500 meter per pixel [6]. We used data from 2022. Four socio-economic indicators were collected to evaluate the model: regional GDP (GRDP), population (POP), highly educated population ratio (HER), and work accidents (WAC). These ground-truth data were obtained from the official websites of each country's national statistical office. NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical classification system developed by Eurostat to ensure consistency in regional statistical analysis across member states. All indicators in our data were standardized to the NUTS-3 level, which corresponds to small administrative regions—such as districts or counties—used for regional statistics within the European Union. The dataset covers 73 NUTS-3 regions in Poland, 8 in Slovakia, 14 in the Czech Republic, and 20 in Hungary. Below are data summary:

- **Regional GDP (GRDP)** for 2019 were collected from Eurostat.
- **Population (POP)** data were categorized into three age groups: pre-productive (0–14 years), productive (15–64 years), and post-productive (65 years and older). Due to limited availability, 2021 census was used for Slovakia, while 2019 data from national statistical offices were used for the remaining countries.
- **Work accidents (WAC)** refers to the number of reported cases involving more than three days of incapacity to work, as well as fatal incidents resulting from work-related injuries. All data refer to 2019 and were obtained from national statistical offices.
- **Highly educated population ratio (HER)** is the percentage of individuals who have completed tertiary education relative to the total population. Due to data availability, 2022 census was used for Hungary, whereas 2019 data from national statistical offices were used for the other countries.

3.2 Performance Comparison

We employ Spearman (ρ) correlation, Pearson (r) correlation, and R^2 to measure agreement between our predictions and ground-truth data. For robustness, we repeated the experiments three times using random seeds and divisions of labeled and unlabeled data.

We evaluated against four baselines with the few-shot setting: (1) READ uses a CNN trained on a human-annotated dataset to summarize embeddings of satellite images within a region into a fixed-sized vector, then trains a regressor on this vector [7]; (2) Tile2Vec is an unsupervised representation learning model on satellite imagery that is fitted on few shot region images and labels to serve as a scorer [11]; (3) SimpleLR uses standard linear regression on all tabular numeric inputs to estimate regional outcomes; and (4) Nightlight trains a linear regression model using two features, the sum of nightlight and the average nightlight [5].

Table 1 presents the results for the 10-shot setting. Our model consistently ranks first or second for all target variables, demonstrating the effectiveness of using label distribution data in in-context learning. Best results are highlighted in bold text. Our results show that across all countries, GeoSEE consistently achieves the highest average performance across the four indicators. While the Nightlight baseline performs well in certain cases—particularly for GRDP and WAC—it struggles to generalize across all target variables. For instance, while Nightlight achieves an R^2 of 0.9687 for GRDP in Hungary and 0.6844 for WAC in Poland, yet its performance on HER in Slovakia and the Czech Republic lags as 0.8258 and 0.5031, respectively. This suggests that nighttime light intensity may serve as a useful proxy for infrastructure, but is less effective in capturing more nuanced socio-economic indicators such as education level.

Estimates on HER consistently exhibit lower correlation values and greater variance for all models, suggesting that this indicator is less "visible" from satellite-observable features. However, our model continues to perform better than baselines for this indicator. For example, in Poland, READ and Tile2Vec produce near-zero

Table 1: Evaluation results for the indicators regional GDP (GRDP), population (POP), work accidents (WAC), and Highly Educated People Ratio (HER) over 3 iterations in the 10-shot setting. Experiments are conducted on the four Visegrád countries: Hungary (HU), Slovakia (SK), Poland (PL), and Czech Republic (CZ). The best performance is highlighted in bold.

Country	Method	GRDP			POP			WAC			HER		
		Spearman	Pearson	R^2	Spearman	Pearson	R^2	Spearman	Pearson	R^2	Spearman	Pearson	R^2
HU	READ	0.1951	0.6109	0.4344	0.0513	0.3482	0.2090	0.0573	0.5992	0.4318	0.1656	0.6649	0.5548
	Tile2Vec	0.1829	0.8696	0.7670	0.0818	0.7474	0.5603	0.0920	0.5814	0.4357	0.2607	0.6142	0.5294
	SimpleLR	0.5093	0.9581	0.9184	0.4062	0.8311	0.7039	0.4062	0.8311	0.7039	0.3892	0.7551	0.5863
	Nightlight	**0.9229**	0.9841	0.9687	**0.9450**	0.9381	0.8831	**0.9336**	0.9685	0.9386	0.6263	0.9306	**0.8660**
	GeoSEE (Ours)	0.8917	**0.9870**	**0.9743**	0.7790	**0.9596**	**0.9210**	0.9317	**0.9708**	**0.9430**	**0.6562**	**0.9306**	0.8659
SK	READ	-0.1270	0.4228	0.2340	0.0794	0.2962	0.1021	0.1667	0.3398	0.1727	0.0714	0.6330	0.5696
	Tile2Vec	0.0952	0.5479	0.3181	0.0317	0.2116	0.0503	0.1190	0.2518	0.1187	0.1746	0.6473	0.4559
	SimpleLR	0.3254	0.6659	0.4662	0.4365	0.4065	**0.2042**	0.4365	0.4065	0.2042	**0.5317**	0.8245	0.6907
	Nightlight	0.7619	**0.9496**	**0.9021**	0.3571	0.3138	0.0991	0.6190	0.7171	0.5143	0.3413	0.9081	0.8258
	GeoSEE (Ours)	**0.8413**	0.9452	0.8950	**0.7302**	**0.4202**	0.1767	**0.8492**	**0.7954**	**0.6435**	0.5000	**0.9401**	**0.8839**
PL	READ	0.0317	0.0412	0.0026	0.0822	0.0713	0.0067	0.0979	0.0447	0.0027	0.0482	0.0538	0.0045
	Tile2Vec	0.0038	0.0379	0.0018	0.0772	0.0687	0.0097	0.0580	0.0094	0.0001	0.1062	0.0693	0.0059
	SimpleLR	0.6542	0.6674	0.4474	0.5946	0.6098	0.4160	0.5946	0.6098	0.4160	0.6422	0.7382	0.5495
	Nightlight	**0.9111**	0.8248	0.6808	0.7600	0.7036	0.5171	**0.8451**	**0.8269**	**0.6844**	0.6745	0.7595	0.5775
	GeoSEE (Ours)	0.9065	**0.8301**	**0.6922**	**0.8602**	**0.8537**	**0.7308**	0.8164	0.8082	0.6563	**0.6897**	**0.8406**	**0.7068**
CZ	READ	0.2226	0.5034	0.3081	0.1279	0.2101	0.0725	0.2622	0.3263	0.2082	0.1479	0.4320	0.2350
	Tile2Vec	0.3288	0.5537	0.3522	0.2020	0.1604	0.0297	0.2791	0.4669	0.2822	-0.0298	0.3677	0.1576
	SimpleLR	0.6024	0.8034	0.6468	0.7722	0.8083	0.6830	0.7722	0.8083	0.6830	-0.1492	0.3501	0.1752
	Nightlight	**0.9212**	**0.9010**	**0.8169**	0.9179	0.9051	0.8238	**0.9106**	**0.8436**	**0.7120**	0.6693	0.6841	0.5031
	GeoSEE (Ours)	0.9201	0.8908	0.7977	**0.9512**	**0.9705**	**0.9418**	0.9044	0.7200	0.5318	**0.7734**	**0.8568**	**0.7597**

Table 2: Ablation results for each indicator (GRDP, POP, HER, and WAC), where the results are averaged across the four Visegrád countries. Each ablation excludes a specific component of GeoSEE (Ablation 1–4). This table summarizes the indicator-level impact by averaging country-level results, under the 10-shot setting.

Model	Spearman (ρ)					Pearson (r)					R^2				
	GRDP	POP	WAC	HER	AVG	GRDP	POP	WAC	HER	AVG	GRDP	POP	WAC	HER	AVG
GeoSEE(Ours)	**0.8899**	**0.8302**	**0.8754**	**0.6548**	**0.8126**	**0.9133**	**0.8010**	**0.8236**	**0.8920**	**0.8575**	**0.8398**	**0.6926**	**0.6936**	**0.8920**	**0.7795**
(Ablation 1) Without geo-spatial modules	0.7711	0.7268	0.7977	0.6179	0.7284	0.9048	0.6738	0.8066	0.8578	0.8108	0.8238	0.5052	0.6701	0.8578	0.7142
(Ablation 2) Randomly selected modules	0.8066	0.7287	0.8192	0.6432	0.7494	0.8981	0.7042	0.8115	0.8753	0.8223	0.8128	0.5641	0.6757	0.8753	0.7320
(Ablation 3) Without classification	0.2556	0.5351	0.0970	0.5434	0.3578	0.5049	0.5752	0.2619	0.6571	0.4998	0.3977	0.4680	0.1108	0.4703	0.3617
(Ablation 4) Without scoring function	0.7944	0.7114	0.7176	0.6216	0.7113	0.7985	0.6684	0.6984	0.8690	0.7586	0.6771	0.5249	0.5463	0.8690	0.6543

R^2 values, while our model achieves 0.7068. This disparity likely reflects the complex, multi-dimensional nature of educational attainment, which may be influenced by institutional, cultural, and policy factors not easily captured by remote sensing data.

3.3 Ablation Study

To test the role of each component in our model, we removed or modified the following components one at a time: (1) Without geo-spatial modules: conducting LLM-based inference using only address information, without external data collected by modules; (2) Randomly selected modules: conducting LLM-based inference using randomly selected modules; (3) Without classification: forcing to directly produce exact number of a indicator as a result of in-context learning; (4) Without scoring function: using only the average value of the classification in section 2.3.

Table 2 shows that any component-wise alteration or exclusion decreases performance. We first evaluate the impact of removing the use of external modules. In Ablation 1, all modules are disabled except the address information, whereas Ablation 2 disables the selection process in section 2.2. The results of Ablation 1 reflect the performance achievable based solely on the LLM's knowledge. In Ablation 2, the performance remains comparable to that of Ablation 1, indicating that without careful module selection, the model cannot surpass what is attainable using prior knowledge alone. A substantial performance drop is observed in Ablation 3, which removed the classification step, suggesting that classification used in the prompt may help prevent critical errors related to LLM's hallucination. Ablation 4, which removes the scoring function, also leads to a performance decline in R^2 highlighting the importance of the scoring mechanism in our framework.

4 Discussion and Conclusion

This current work demonstrated a new method to estimate socio-economic indicators at the subnational level using a large language model. Our model used LLM as a domain expert for the inference task and identified key features from multiple data sources based on its extensive prior knowledge and reasoning abilities. The simplicity of its structure, which only requires natural language descriptions of the desired indicator and features, makes the model adaptable and extensible, allowing computations on any geo-location, even in areas that have limited data. Below we discuss practical issues for deployment and limitations of the proposed method.

Table 3: R^2 performance based on alternative LLMs.

Model	GRDP	POP	WAC	HER
GeoSEE (Ours, o4-mini)	**0.840**	**0.693**	**0.694**	**0.892**
o3-mini	0.824	0.629	0.644	0.872
Deepseek-reasoner	0.821	0.678	0.673	0.772

4.1 Operational Considerations

We discuss operational and deployment considerations. One observation is that nightlight remains a valuable predictor, particularly in resource-limited settings where running LLM-based inference may not be feasible. The target variables in our experiments, such as GRDP and population, exhibit strong correlations with nightlight intensity. Additionally, the nightlight module was consistently selected by GeoSEE across all indicator experiments, contributing directly to final predictions. As a result, predictions based solely on nightlight data often achieve strong performance, at times approaching that of the full model. However, as shown in Table 1, GeoSEE consistently outperforms the nightlight-only baseline across a broad range of target variables, demonstrating its advantages when computational resources allow for LLM-based inference.

Another important aspect we examine is the choice of LLM. The current model used o4-mini for in-context learning. To evaluate its impact, we compare it with two alternative models: *o3-mini*, an earlier reasoning model by the same provider, and *Deepseek-reasoner*, which follows a different training methodology. As shown in Table 3, ours (o4-mini) consistently outperforms the alternatives, confirming it as the best performing option. However, the performance gap introduced by switching the LLM remains relatively small. Compared to the more substantial drops observed in Table 2 due to the removal of key system components, changes in the LLM result in a more graceful degradation. This suggests that while the choice of LLM matters, the overall design of the system has a more significant impact on performance.

4.2 Limitations and Broader Impact

Several factors must be considered before full practical implementation. First, our experiments used a limited set of external data and modules; however, the model is easily expandable, and future studies can incorporate new data sources and modules to build on our findings. Second, the model has been tested for a one-time snapshot per country and indicator. We had demonstrated GeoSEE's effectiveness using empirical validation from 115 administrative regions as reference points, extending predictions across multiple countries and indicators such as regional GDP, population and educational attainment. However, further testing will be necessary to ensure reliable analyses at finer temporal intervals, which are essential for practical applications requiring time-series evaluations.

Our model offers several key advantages for policymakers, including flexible geographic targeting, data integration, and transparent evaluation. First, the model allows for precise geographic targeting by generating estimates at any administratively meaningful level—whether a city district identified for infrastructure investment or a rural area designated for poverty alleviation programs. Second, it supports the integration of diverse data sources and enables policymakers to combine existing administrative records, satellite imagery, and other datasets to enhance coverage, particularly in regions where traditional surveys are limited. Finally, the model ensures transparent evaluation by explicitly selecting which data sources to use, allowing agencies to assess and understand how predictions are generated.

We hope that models like ours will facilitate socio-economic analysis at the subnational level and provide estimates that support decision-making in policy and business. This is an exciting direction, as the method can be applied in data-scarce regions while also have the potential to be fine-tuned with proprietary data in data-rich areas to enhance predictions for regional or specialized use. By reducing reliance on traditional data collection, the proposed method also enables timely and scalable socio-economic monitoring in settings with limited data availability.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv:2303.08774* (2023).

[2] Donghyun Ahn, Jeasurk Yang, Meeyoung Cha, Hyunjoo Yang, Jihee Kim, Sangyoon Park, Sungwon Han, Eunji Lee, Susang Lee, and Sungwon Park. 2023. A human-machine collaborative approach measures economic development using satellite imagery. *Nature Communications* 14, 1 (2023), 6811.

[3] Adrian Albert, Jasleen Kaur, and Marta C Gonzalez. 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In *proceeding of the ACM SIGKDD*. 1357–1366.

[4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv:2305.10403* (2023).

[5] Hasi Bagan and Yoshiki Yamagata. 2015. Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data. *GIScience & Remote Sensing* 52, 6 (2015), 765–780.

[6] Christopher D Elvidge, Mikhail Zhizhin, Tilottama Ghosh, Feng-Chi Hsu, and Jay Taneja. 2021. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. *Remote Sensing* 13, 5 (2021), 922.

[7] Sungwon Han, Donghyun Ahn, Hyunji Cha, Jeasurk Yang, Sungwon Park, and Meeyoung Cha. 2020. Lightweight and Robust Representation of Economic Scales from Satellite Imagery. In *proceeding of the AAAI*.

[8] Andrew Head, Mélanie Manguin, Nhat Tran, and Joshua E Blumenstock. 2017. Can human development be measured with satellite imagery? *ICTD* (2017).

[9] Agustín Indaco. 2020. From twitter to GDP: Estimating economic activity from social media. *Regional Science and Urban Economics* 85 (2020), 103591.

[10] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[11] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *proceeding of the AAAI*, Vol. 33. 3967–3974.

[12] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *proceeding of the NeurIPS* 35 (2022), 3843–3857.

[13] Jin-Hwi Park, Young-Jae Park, Junoh Lee, and Hae-Gon Jeon. 2022. DevianceNet: Learning to Predict Deviance from a Large-Scale Geo-Tagged Dataset. In *proceeding of the AAAI*, Vol. 36. 12043–12052.

[14] Sungwon Park, Sungwon Han, Donghyun Ahn, Jaeyeon Kim, Jeasurk Yang, Susang Lee, Seunghoon Hong, Jihee Kim, Sangyoon Park, Hyunjoo Yang, et al. 2022. Learning economic indicators by aggregating multi-level geospatial information. In *proceeding of the AAAI*, Vol. 36. 12053–12061.

[15] Abulhair Saparov and He He. 2022. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. *CoRR* (2022).

[16] Sanja Šćepanović, Igor Mishkovski, Pan Hui, Jukka K Nurminen, and Antti Ylä-Jääski. 2015. Mobile phone call data as a regional socio-economic proxy indicator. *PloS one* 10, 4 (2015), e0124160.

[17] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2698–2706.

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *proceeding of the ICLR*.