GeoGR²: Zero-Shot Geospatial Inference via Geostatistically-Guided Iterative Refinement with LLMs

Jinfan Tang[1]*, Kunming Wu[1]*, Xieruifeng Gong[2]*, Yuya He[1]
HuJie[1], Wu Junhui[3], Yuankai Wu[1]†

[1]College of Computer Science, Sichuan University
[2]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University
[3]College of Life Science, Sichuan University

Abstract

Standard large language model prompting treats geospatial inference as independent, instance-wise prediction, ignoring the fundamental spatial dependencies that govern geographic reality. Consequently, even advanced models struggle with spatial consistency and exhibit severe biases toward populous regions. To bridge this gap, we propose **GeoGR²** (**Geo**spatial **G**raph **R**efine **R**easoning), a framework that formalizes zero-shot geospatial prediction as an iterative message-passing process on a dynamically constructed graph. Unlike static retrieval methods, GeoGR² instantiates three dynamic operators via collaborating operators: (1) a *Topology Operator* that constructs graph topology to enforce the Spatial Markov property; (2) a *Feature Operator* that enriches nodes with task-relevant semantic covariates; and (3) an *Update Operator* that performs natural language message passing to iteratively minimize spatial discrepancy. Theoretically, we frame this refinement as a contraction mapping that approximates the fixed point of a global consistency equation. Empirically, we validate GeoGR² on diverse physical and socioeconomic tasks. Results demonstrate that by explicitly embedding geostatistical inductive biases, GeoGR² significantly outperforms standard prompting baselines, while effectively mitigating systematic geographic bias. Our framework leverages large language models' intrinsic capacity for understanding spatial correlations through explicit topological scaffolding, without resorting to general graph reasoning paradigms.

1 Introduction

Large Language Models (LLMs) have emerged as powerful reasoning engines across domains (Li et al., 2022; Boiko et al., 2023), with growing applications in geospatial intelligence (Manvi et al., 2024b; Li et al., 2024). However, a fundamental mismatch persists: **LLMs process instances independently (i.i.d.) (Figure 1-right), whereas geospatial data is structured by *Tobler's First Law* (Tobler, 1970).** Unlike classical geostatistics (e.g., Kriging (Cressie, 1993)) that explicitly models spatial autocorrelation, standard prompting ignores latent geographic connectivity. Consequently, even advanced models exhibit spatial inconsistency, multi-hop hallucination, and bias toward populous regions due to skewed training distributions (Manvi et al., 2024a).

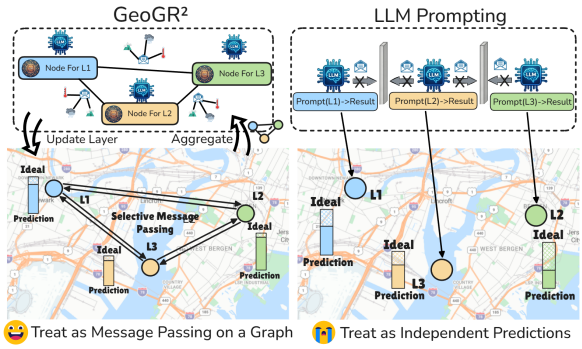


Figure 1: Conceptual Comparison Between GeoGR²(**Left**) and LLM Direct Prediction(**Right**) in geospatial prediction.

While Graph Neural Networks (GNNs) (Kipf, 2016; Wu et al., 2020) effectively encode spatial structure via message passing, they require fixed topologies and heavy supervised training (Wu et al., 2021), lacking the semantic flexibility of foundation models. Conversely, current LLM-based geospatial methods rely on costly fine-tuning (Deng et al., 2024), which is static and fails to adapt to novel spatial distributions at inference.

Recent work indicates LLMs can grasp spatial dependencies when given explicit topological context (Shang and Huang, 2025). This motivates our hypothesis: LLMs' latent spatial understanding can be operationalized for geostatistics by *dynamically*

*Equal contribution.
†Corresponding author: wuyk0@scu.edu.cn

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 6560–6580
October 24-29, 2026 ©2026 Association for Computational Linguistics

constructing interaction graphs during reasoning, transforming standard i.i.d. prompting into a spatially coherent inference paradigm.

To this end, we propose **GeoGR^2** (**Geo**spatial **G**raph **R**efine **R**easoning), a training-free framework that embeds geostatistical principles directly into the LLM inference loop. Rather than a general graph engine, GeoGR^2 uses graph structures as a scaffold to operationalize Tobler's Law. We formulate zero-shot prediction as message passing on a dynamic graph (Figure 1-left), approximating GNN workflows via three natural-language operators: (1) **Topology Operator** identifies the target's "Markov Blanket"; (2) **Feature Operator** injects semantic covariates; and (3) **Update Operator** iteratively aggregates neighborhood signals to minimize spatial discrepancy. This loop enables "spatial analogy" reasoning, refining predictions through local consensus. Evaluated on physical (Temperature, Precipitation) and socioeconomic (GDP, Mortality) tasks, GeoGR^2 achieves SOTA zero-shot accuracy while substantially reducing geographic bias. Our contributions are:

- **Conceptual Formulation:** We formalize zero-shot geospatial LLM inference as iterative spatial refinement, bridging classical geostatistics and prompting paradigms.
- **GeoGR^2 Framework:** A novel training-free architecture that dynamically constructs interaction graphs and performs natural-language neighborhood aggregation, enforcing spatial consistency without fine-tuning.
- **Experimental Validation:** Extensive experiments show consistent superiority over prompting and ML baselines. Theoretically, we show the refinement process acts as a contraction mapping, ensuring convergence to accurate and fair spatial representations.

2 Related Work

2.1 LLMs for Geospatial Intelligence

Geospatial LLMs have evolved from basic prompting or RAG-based coordinate memorization (Manvi et al., 2024b; Zhang et al., 2023) to domain-adaptive pretraining and instruction tuning (Li et al., 2024; Deng et al., 2024). While performant, these supervised paradigms demand extensive labeled data and remain sensitive to distribution shifts. Notably, Gurnee and Tegmark (2023) demonstrate that LLMs implicitly encode linear spatiotemporal representations. Instead of costly parameter updates, we *explicitly* activate these latent structures via inference-time graph construction. Although recent evidence confirms LLMs can grasp spatial and relational patterns through explicit topological prompting (Shang and Huang, 2025), this capability remains underexploited in geospatial prediction—a gap our approach directly bridges.

2.2 Agentic Frameworks and Self-Refinement

To address the limitations of single-turn prompting, agentic frameworks such as Chain-of-Thought (Wei et al., 2022) and ReAct (Yao and et al., 2023) facilitate task decomposition, while self-refinement paradigms like Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023) incorporate iterative self-critique. Despite their efficacy in logic and coding domains, these generic strategies lack *domain-specific inductive biases*, typically relying on internal consistency or scalar feedback rather than structured external constraints.

3 Problem Formulation

We frame the problem of geospatial inference as a zero-shot regression task over a set of distributed locations. Let $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ denote a set of geographic locations, where each ℓ_t is defined by its coordinates (latitude, longitude). Associated with each location is a ground-truth scalar variable $y_t \in \mathbb{R}$.

Task Definition. Our objective is to estimate y_t using a pre-trained Large Language Model (LLM) $f_\theta(\cdot)$, without any parameter updates or supervised examples (zero-shot). The estimation is formalized as:

$$\hat{y}_t = f_\theta(\texttt{prompt}(\ell_t, \mathcal{T})), \tag{1}$$

where $\mathcal{T}$ represents the task description.

Evaluation Metrics. Following Manvi et al. (2024a), we assess performance based on two axes: *predictive accuracy* and *fairness*.

1) Ranking Accuracy (ρ): Since LLMs often output values on arbitrary scales unaligned with real-world units, we measure the structural alignment using Spearman's rank correlation coefficient ρ between the predicted vector $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$.

2) Geographic Bias (Bias_y*):* To quantify systematic inequity, we adopt the composite bias metric proposed by Manvi et al. (2024a). This metric penalizes models that exhibit high correlation with a demographic anchor distribution (e.g., population density) while accounting for prediction variance and refusal rates:

$$\text{Bias}_y(\hat{\mathbf{y}}) = \rho(\hat{\mathbf{y}}, \mathbf{d}) \cdot \text{MAD}(\hat{\mathbf{y}}) \cdot a^2, \quad (2)$$

where $\mathbf{d} \in \mathbb{R}^n$ is the anchor distribution vector (e.g., population density per location), $\text{MAD}(\cdot)$ denotes the Mean Absolute Deviation from the mean, and $a \in [0, 1]$ is the answer rate. Intuitively, a fair model should yield predictions uncorrelated with the anchor variable $\mathbf{d}$ rather than hallucinating based on population priors.

4 Method

As illustrated in Figure 2, GeoGR^2 departs from independent prompting by formalizing inference as an iterative message-passing process driven by three collaborating operators: (1) a **Topology Operator** that dynamically defines local connectivity; (2) a **Feature Operator** that enriches node representations; and (3) an **Update Operator** that performs iterative state refinement.

4.1 Spatial Dependency Formulation via Graph Structures

Our framework is grounded in Probabilistic Graphical Models (PGMs) (Koller and Friedman, 2009). We postulate that the joint distribution of geospatial variables follows a Markov Random Field (MRF) structure (Besag, 1974) regarding an underlying latent graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, the graph formalism serves exclusively as a mathematical abstraction for spatial autocorrelation; it is not our objective to advance general-purpose LLM graph reasoning.

To bridge the gap between this abstract formulation and our implementation, we formally define the physical interpretations of the mathematical components:

- **Nodes ($\mathcal{V}$):** The set of discrete geographic locations $\mathcal{L}$.
- **Graph Topology ($\mathbf{W}$):** The binary adjacency matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ constructed by the **Topology Operator** (ψ). $W_{ij} = 1$ implies location j is selected as a semantic or physical neighbor of i.
- **Graph Laplacian ($\mathbf{L}$):** Defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ (where $\mathbf{D}$ is the degree matrix).
- **Refinement Step (λ):** The scalar λ controls the rate of information diffusion during message passing, corresponding to the *trust level* the Update Operator places in neighborhood consensus.

Based on the definitions above, we provide three theoretical pillars that justify the design of our framework.

Assumption 4.1 (Spatial Markov Property)**.** *Following (Besag, 1974), for any location ℓ_t, given its set of neighbors $\mathcal{N}_t$ (the Markov blanket), the variable y_t is conditionally independent of all other variables $y_{\mathcal{V} \setminus (\mathcal{N}_t \cup \{\ell_t\})}$. Formally:*

$$P(y_t \mid \mathbf{y}_{\mathcal{V} \setminus \{t\}}) = P(y_t \mid \mathbf{y}_{\mathcal{N}_t}) \quad (3)$$

Remark 1. *This assumption justifies the design of our **Topology Operator** (ψ). It implies that instead of reasoning over the computationally intractable global context, it suffices to construct a high-quality local neighborhood $\mathcal{N}_t$ to theoretically achieve optimal predictive performance.*

Proposition 4.1 (Benefit of Spatial Aggregation)**.** *As established in (Stein, 1956), in multi-dimensional estimation tasks ($|\mathcal{V}| \geq 3$), independent estimation of node-wise variables is* inadmissible *under quadratic loss. Specifically, there exists a spatially aggregated estimator $\hat{\mathbf{y}}^{agg}$ (acting as a linear smoother over the neighbor graph) that strictly dominates independent estimation in terms of total Mean Squared Error (MSE):*

$$\mathbb{E}[\|\hat{\mathbf{y}}^{agg} - \mathbf{y}\|^2] < \mathbb{E}[\|\hat{\mathbf{y}}^{ind} - \mathbf{y}\|^2], \quad (4)$$

provided the aggregation strength is within a non-trivial bound determined by the graph Laplacian. (proof in Appendix A)

Remark 2. *This proposition formally explains* why *our message passing mechanism is superior to standard independent prompting. By conditioning each prediction on its spatial neighbors, GeoGR^2 performs structured variance reduction, mitigating the hallucination and bias inherent in isolated LLM predictions.*

Lemma 4.1 (Convergence Condition)**.** *Let the refinement process be viewed as an operator $T : \mathbb{R}^{|\mathcal{V}|} \to \mathbb{R}^{|\mathcal{V}|}$ acting on a complete metric space. If the Update Operator is designed to act as a contraction mapping with Lipschitz constant $L < 1$,*

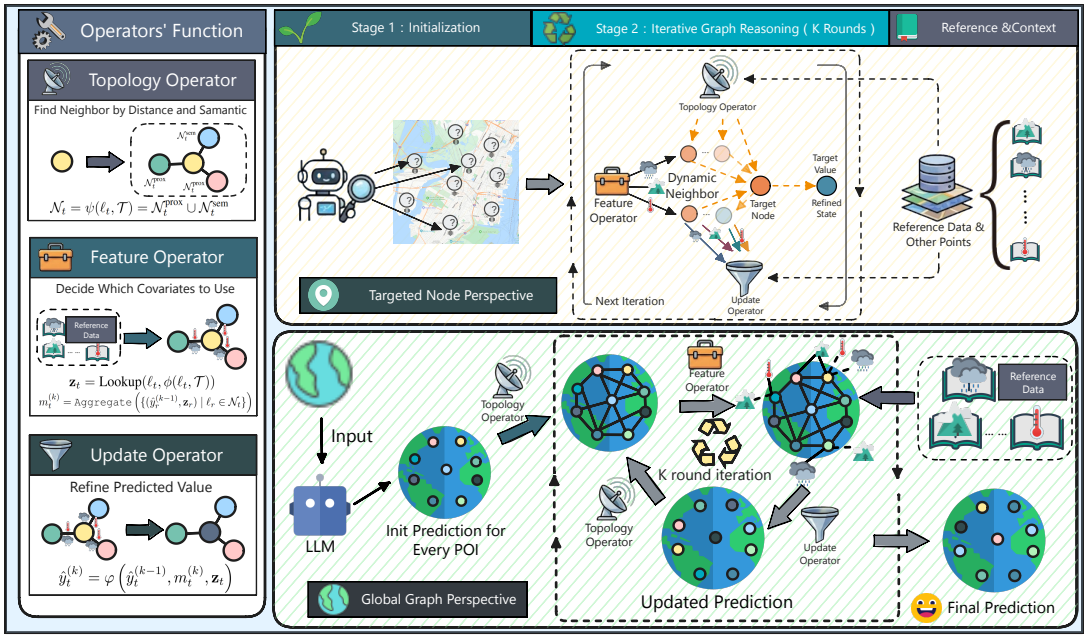


Figure 2: Overview of GeoGR^2. We formulate geospatial prediction as an iterative graph reasoning process where operators dynamically construct spatial topology, select semantic features, and perform natural language message passing to refine predictions.

then by the Banach Fixed Point Theorem (Kreyszig, 1991), the iterative sequence $\hat{\mathbf{y}}^{(k+1)} = T(\hat{\mathbf{y}}^{(k)})$ *converges to a unique fixed point* $\mathbf{y}^*$*:*

$$\lim_{k\to\infty} \hat{\mathbf{y}}^{(k)} = \mathbf{y}^*. \tag{5}$$

Remark 3. *This lemma validates the stability of our* ***Update Operator*** *(φ).*

4.2 Initial State Estimation

The process initializes with the LLM, π, which generates a preliminary state estimate $\hat{y}_t^{(0)}$ for each location ℓ_t. Leveraging the GeoLLM framework (Manvi et al., 2024b), π utilizes OpenStreetMap data to ground the zero-shot prediction in basic geographic context:

$$\hat{y}_t^{(0)} = \pi(\texttt{prompt}(\ell_t, \mathcal{T})), \quad \forall \ell_t \in \mathcal{L}, \tag{6}$$

where $\mathcal{T}$ denotes the task context. While informative, these initial estimates $\hat{y}_t^{(0)}$ are often spatially incoherent, lacking the constraints imposed by neighboring correlations.

4.3 Dynamic Topology Construction

To enforce the Spatial Markov Property (Theorem 4.1), we must explicitly model the interactions between locations. The Topology Operator, ψ, acts as a dynamic topology generator. For a target node ℓ_t, it identifies a neighborhood set $\mathcal{N}_t \subset \mathcal{L} \setminus \{\ell_t\}$ that exerts the strongest spatial influence.

Unlike fixed distance-based kernels in traditional geostatistics, ψ constructs the neighborhood $\mathcal{N}_t$ semantically and spatially:

$$\mathcal{N}_t = \psi(\ell_t, \mathcal{T}) = \mathcal{N}_t^{\text{prox}} \cup \mathcal{N}_t^{\text{sem}}, \tag{7}$$

where $\mathcal{N}_t^{\text{prox}}$ comprises the k-nearest neighbors ensuring local continuity, and $\mathcal{N}_t^{\text{sem}}$ includes additional reference points selected via LLM reasoning to capture global context or analogous regions.

4.4 Semantic Feature Augmentation

Geographic variables are rarely independent; they are conditioned on latent covariates (e.g., vegetation index influences temperature). The Feature Operator, ϕ, functions as a feature extractor that maps a location ℓ_t to a task-specific subset of auxiliary variables $\mathcal{S}_t \subseteq \mathcal{V}$.

$$\mathbf{z}_t = \text{Lookup}(\ell_t, \phi(\ell_t, \mathcal{T})), \tag{8}$$

where $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{S}_t|}$ represents the values of selected covariates at ℓ_t.

4.5 Natural Language Message Passing

The core of GeoGR^2 is the Update Operator, φ, which mimics the *Aggregate-and-Update* mechanism of GNNs, but operates within the semantic

space of an LLM rather than vector space. At iteration k, for each target ℓ_t, the operator aggregates information from its neighborhood $\mathcal{N}_t$. The "message" $m_t^{(k)}$ effectively summarizes the spatial context:

$$m_t^{(k)} = \texttt{Aggregate}\left(\{(\hat{y}_r^{(k-1)}, \mathbf{z}_r) \mid \ell_r \in \mathcal{N}_t\}\right). \tag{9}$$

The Update Operator then performs a non-linear update to the prediction state, conditioned on the current estimate, the aggregated spatial message, and the local features:

$$\hat{y}_t^{(k)} = \varphi\left(\hat{y}_t^{(k-1)}, m_t^{(k)}, \mathbf{z}_t\right). \tag{10}$$

This process allows the LLM to perform "reasoning by spatial analogy" adjusting $\hat{y}_t$ to align with the observed values of its neighbors. Driven by the convergence principle (Proposition 4.1), the refinement continues for K rounds or until spatial consistency is achieved.

4.6 Algorithm and Iterative Refinement

The overall procedure of GeoGR^2 is summarized in Algorithm 1 (Appendix C). The process begins with an independent initialization state, $\hat{\mathbf{y}}^{(0)}$, where the base LLM generates zero-shot predictions for each location.

Instead of relying on traditional numerical solvers, our framework iteratively updates the global prediction state $\hat{\mathbf{y}}^{(k)}$ via natural language reasoning. In each iteration k, the Topology Operator ψ and Feature Operator ϕ provide the necessary spatial and semantic context for each location. Subsequently, the LLM-based Update Operator φ guides each location ℓ_t to reconcile its prediction with the gathered neighborhood consensus Ω_t. This refinement mechanism drives a sequential state transition: $\hat{\mathbf{y}}^{(0)} \to \hat{\mathbf{y}}^{(1)} \to \cdots \to \hat{\mathbf{y}}^{(K)}$. By iteratively aligning individual outputs with local context, the system effectively reduces spatial inconsistencies, yielding a final set of predictions $\{\hat{y}_t^{(K)}\}$ that is spatially coherent and grounded in geospatial features.

5 Experiment

We evaluate GeoGR^2 following Manvi et al. (2024b), requiring models to generate 10-point scores per task. Performance is measured via Spearman correlation and scoring bias (Eq. (2)), using population density as anchor distribution $\mathbf{d}$ to account for regional imbalances.

We assess four geospatial tasks spanning social and physical geographic data: **Infant Mortality** (healthcare access, social equity; (Manvi et al., 2024b), 1950 cities), **GDP** (economic disparities; (Kummu et al., 2025), 1950 cities), and climate variables **Temperature** and **Precipitation** (climate/hydrological patterns; WorldClim* (Fick and Hijmans, 2017), 2000 global locations each).

GeoGR^2 builds on LLMs with GeoLLM (Manvi et al., 2024b) as the *LLM-only baseline* using local context without global topology. We treat GeoLLM outputs as initial predictions $\mathbf{y}_0$ and integrate GeoGR^2 as a graph-based refinement module to quantify gains from global reasoning. Evaluated LLMs include: **GPT-3.5-Turbo** (Ouyang et al., 2022) (175B generalist, cutoff Sep 2021), **GPT-4o-mini** (OpenAI et al., 2024) (compact MoE, 10-50B), **DeepSeek-V3** (DeepSeek-AI et al., 2025) (671B MoE), **GeoGPT** (LLaMA-3.1-70B fine-tuned on 800k geoscience samples), and **Gemini-3-flash** (DeepMind and Research, 2025) (multimodal lightweight, cutoff Dec 2024). Additionally, we include KNN (Bishop, 2006) and XGBoost (Chen and Guestrin, 2016) as traditional baselines under a strict few-shot setting (20 randomly sampled training points) to maintain parity with our training-free approach.

5.1 Comparison with Generic Agentic Reasoning

We implement a ReAct agent (Yao and et al., 2023) equipped with geographic retrieval, neighbor lookup, and estimation tools, capped at 20 rounds per location. Table 1 shows severe *convergence instability*: unstructured tool calls exhaust the budget across all tasks, yielding lower Spearman correlation (e.g., 0.599 vs. **0.747** on Infant Mortality) and higher geographic bias (0.379 vs. **-0.006**). Token costs reach 11,306–15,739 per location (**7–10**× GeoGR^2). These results highlight a core limitation of generic agents: lacking spatial inductive biases, they incur redundant retrieval without enforcing neighborhood consistency. GeoGR^2 resolves this by embedding spatial dependencies directly into the inference topology, guaranteeing single-pass execution with lower cost and higher fidelity.

5.2 Main Results

Table 2 summarizes the performance gains achieved by GeoGR^2 across various LLM backbones. Our framework consistently improves

*https://www.worldclim.org/

Task	Method	ρ	Bias	Tokens/Loc.
Infant Mort.	ReAct	0.599	0.379	15,739
	GeoGR^2	**0.747**	**-0.006**	**1,638**
GDP	ReAct	0.501	0.283	12,450
	GeoGR^2	**0.653**	**0.077**	**1,744**
Temperature	ReAct	0.535	0.311	11,306
	GeoGR^2	**0.622**	**0.024**	**1,591**
Precipitation	ReAct	0.457	0.140	12,579
	GeoGR^2	**0.694**	**0.039**	**1,503**

Table 1: Efficiency and Performance: GeoGR^2 vs. ReAct Agent (GPT-3.5-Turbo).

Spearman correlation while significantly reducing scoring bias (e.g., a **96.8%** reduction for GPT-3.5). The following key insights emerge:

Comparison with Traditional Baselines. Traditional machine learning models, KNN and XGBoost, serve as competitive baselines for capturing local spatial patterns. While vanilla LLMs (e.g., GPT-3.5, GPT-4o-mini) often underperform these numerical models in tasks like GDP and Infant Mortality, the integration of GeoGR^2 enables even lightweight or general-purpose LLMs to significantly surpass these traditional methods. This highlights the framework's ability to bridge the gap between structured spatial regression and linguistic reasoning.

Scaling and Retrofitting Effects. GeoGR^2 demonstrates a robust "retrofitting" effect across different model scales. Weaker models like GPT-3.5-Turbo exhibit the most dramatic relative improvements, while state-of-the-art models like DeepSeek-V3 and Gemini-3-flash achieve new performance ceilings. Notably, for GeoGPT—which is already fine-tuned on 800k geoscience samples—GeoGR^2 further refines its output calibration, suggesting that our retrieval-reasoning mechanism is complementary to domain-specific pretraining.

Mitigating Geographic Bias. A critical contribution of GeoGR^2 is the drastic reduction in absolute bias. Even high-performing models like Gemini-3-flash show severe geographic disparities at baseline (e.g., -0.456 on infant mortality). By utilizing population density as an anchor distribution, GeoGR^2 successfully mitigates data-driven biases, ensuring that spatial reasoning remains equitable across diverse regions regardless of their representation in the original training corpus.

GeoGR^2 significantly reduces bias in nearly all settings, even achieving near-zero values in some cases (e.g., **-0.006** for GPT-3.5-Turbo on infant mortality, **-0.011** for GPT-4o-mini). For cutting-edge lightweight models like Gemini-3-flash that already possess strong inherent spatial reasoning, GeoGR^2 still yields notable fairness optimization—reducing its infant mortality bias by 86.6% with a marginal accuracy boost—with only negligible performance fluctuations on individual tasks that do not compromise overall model performance. While domain-adapted models like GeoGPT begin with strong priors and relatively fair outputs, general-purpose models—particularly those with smaller capacities or outdated training data—stand to gain the most from GeoGR^2. This pattern highlights a key strength of our approach: **GeoGR^2 offers consistent and interpretable benefits across diverse model types**, enabling both accuracy and equity improvements without requiring changes to the underlying model architecture or retraining procedures.

Figure 3 provides a qualitative comparison of spatial precipitation predictions using GPT-3.5-Turbo. From the comparison, we observe that GeoGR^2 improves the spatial alignment of predicted precipitation distributions. Specifically, the prediction map on the right more closely resembles the ground-truth middle panel, particularly in regions with historically high bias in LLM outputs such as **Africa**, **South America** and **Southeast Asia**. In the baseline GeoLLM output (left), many regions in those regions are consistently over- or under-estimated. These spatial visualizations further support our claim that GeoGR^2 can effectively correct for structural inequities in LLM outputs, especially for regions that are underrepresented in training data.

Comparison with Strong Spatial Baselines. To validate that GeoGR^2's gains stem from LLM-based geographical reasoning rather than mere spatial smoothing, we compare it against 7 established spatial methods trained with $N = 100$ labeled samples: Kriging, Gaussian Process (GP), Random Forest with coordinates (RF+coord), Kernel Ridge Regression (KRR), Graph Convolutional Network (GCN), Geographically Weighted Regression (GWR), and Graph Smooth.

As shown in Table 3, our zero-shot approach achieves comparable or superior performance to supervised baselines using 100 labeled points. Notably, on the Infant Mortality task (which exhibits

Model	Infant Mortality		GDP		Temperature		Precipitation	
	Spearman	Bias	Spearman	Bias	Spearman	Bias	Spearman	Bias
GeoGPT	0.798	-0.175	0.515	0.274	0.753	0.136	0.797	0.185
+GeoGR^2	0.816 (+2.26%)	-0.031 (-82.3%)	0.523 (+1.55%)	0.146 (-46.7%)	0.771 (+2.39%)	0.108 (-20.6%)	0.802 (+0.63%)	0.157 (-15.1%)
GPT-3.5-Turbo	0.445	-0.188	0.512	0.618	0.462	0.074	0.594	0.031
+GeoGR^2	0.747 (+67.87%)	**-0.006** (-96.8%)	**0.653** (+27.54%)	0.077 (-87.5%)	0.622 (+34.63%)	**0.024** (-67.6%)	0.694 (+16.84%)	0.039 (+25.8%)
GPT-4o-mini	0.704	-0.142	0.565	0.295	0.310	0.203	0.564	0.103
+GeoGR^2	0.762 (+8.24%)	-0.011 (-92.3%)	0.585 (+3.54%)	0.087 (-70.5%)	0.517 (+66.77%)	0.085 (-58.1%)	0.634 (+12.41%)	**0.011** (-89.3%)
DeepSeek-V3	0.820	-0.078	0.620	0.251	0.478	0.232	0.793	0.116
+GeoGR^2	0.834 (+1.71%)	-0.026 (-66.7%)	0.644 (+3.87%)	**0.063** (-74.9%)	0.563 (+17.78%)	0.097 (-58.2%)	**0.813** (+2.52%)	0.061 (-47.4%)
Gemini-3-flash	0.870	-0.456	0.522	0.220	**0.806**	0.147	0.769	0.097
+GeoGR^2	**0.879** (+1.03%)	-0.061 (-86.6%)	0.535 (+2.49%)	0.172 (-21.8%)	0.805 (-0.12%)	0.143 (-2.72%)	0.785 (+2.08%)	0.115 (+18.65%)
KNN	0.352	/	0.347	/	0.431	/	0.399	/
XGBoost	0.415	/	0.460	/	0.497	/	0.510	/

Table 2: Performance Gains with GeoGR^2. Baseline Spearman correlation (higher is better) and Bias (lower absolute value is better) are shown, with percentage increase in Spearman and percentage decrease in Bias for GeoGR^2-enhanced models.

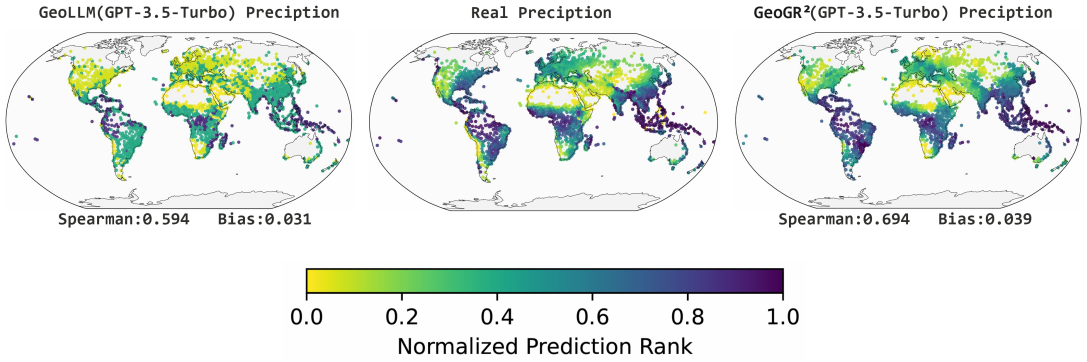


Figure 3: Spatial prediction rank maps for Precipitation prediction using GPT-3.5-Turbo. **Middle**: Ground-truth normalized Precipitation ranks. **Left**: Predictions from the base GeoLLM. **Right**: Predictions from GeoGR^2-enhanced model.

Table 3: GeoGR^2 (Zero-Shot, $N = 0$) vs. Strong Spatial Baselines ($N = 100$). Spearman correlation reported.

Method	Infant Mort.	GDP	Temp.	Precip.
Kriging	0.663	0.681	0.823	0.698
GP (RBF)	0.752	0.695	0.879	0.749
RF+coord	0.799	0.652	0.885	0.694
KRR	0.774	0.637	0.847	0.731
GCN	0.807	**0.707**	**0.910**	0.745
GWR	0.587	0.538	0.696	0.422
GraphSmooth	0.212	0.287	0.280	0.299
Gemini-3-flash	0.870	0.522	0.806	0.769
+ GeoGR^2	**0.879**	0.535	0.805	**0.785**

weaker spatial agglomeration), GeoGR^2 significantly outperforms all supervised spatial interpolators (0.879 vs. GCN's 0.807). This confirms that the performance gain is driven by the LLM's semantic reasoning over geographical contexts (e.g., POI semantics, climate covariates), which pure numerical spatial smoothers cannot capture.

5.3 Ablation Study

Topology Construction Strategy. We further investigate the effectiveness of our topology construction strategy, which comprises two components: (i) automatically selecting the ten nearest neighboring points (GeoGR^2 w/oN_t^{sem}), and (ii) selecting supplementary points dynamically by the LLM (GeoGR^2 w/o N_t^{prox}). We evaluate these variants using GPT-3.5-Turbo as the base model, on the same datasets as in the main experiments. As reported in Table 4, the full GeoGR^2 model achieves the highest Spearman correlation on all tasks except for temperature, where the GeoGR^2 w/o N_t^{sem} variant slightly outperforms it (0.668 vs. 0.644). This suggests that supplementary points chosen by LLM contribute marginally to performance in this specific case, likely due to the strong spatial autocorrelation of temperature data. Importantly, the performance drop is minor, demonstrating the robustness of the full model. In terms of fairness, removing the nearest-neighbor selection (GeoGR^2 w/o N_t^{prox}) only slightly reduces bias. For exam-

ple, in the infant mortality task, the absolute bias decreases by just 0.02, indicating that the inclusion of the nearest points plays a limited role in mitigating bias. We further observe that GeoGR^2 w/o N_t^{sem} consistently achieves higher Spearman correlations than GeoGR^2 w/o N_t^{prox} across tasks. This result reinforces *Tobler's First Law of Geography*, which posits that "everything is related to everything else, but near things are more related than distant things."

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR^2 w/o N_t^{prox}	Spearman	0.645	0.577	0.640	0.659
	Bias	0.034	0.198	0.038	0.051
GeoGR^2 w/o N_t^{sem}	Spearman	0.727	0.650	**0.668**	0.653
	Bias	0.004	0.064	0.031	0.027
GeoGR^2	Spearman	**0.747**	**0.653**	0.644	**0.694**
	Bias	-0.006	0.077	0.024	0.039

Table 4: Ablation Study on Point Selection Strategy

Feature Operator. To evaluate the impact of covariates selected by the Feature Operator, we conduct an ablation study comparing the full GeoGR^2 model with a variant that excludes these external covariates. As shown in Table 5, incorporating bioclimatic variables consistently improves performance across most tasks. For instance, in the infant mortality task, the Spearman correlation increases from 0.705 to 0.747, alongside a reduction in absolute bias. Similar gains are observed for GDP and precipitation tasks. Interestingly, for the temperature task, excluding bioclimatic covariates results in a marginal Spearman increase, but the improvement is negligible and likely reflects feature redundancy rather than added value—since the target itself is temperature, additional temperature-related inputs may contribute limited new information.

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR^2 w/o external variables	Spearman	0.705	0.650	**0.649**	0.691
	Bias	-0.019	0.027	0.022	0.041
GeoGR^2	Spearman	**0.747**	**0.653**	0.644	**0.694**
	Bias	-0.006	0.077	0.024	0.039

Table 5: Ablation Study on External Variables

Effect of Iterative Refinement Rounds. We evaluate how refinement rounds impact GeoGR^2 using DeepSeek-V3 across four tasks. Spearman correlation improves initially, typically peaking at rounds 2–3. Beyond this, additional iterations yield marginal gains or slight degradation, likely due to capacity saturation or overfitting to local structures. Conversely, geographic bias steadily decreases with further refinement, revealing an accuracy–fairness trade-off: later rounds prioritize equity over fidelity by reducing regional output disparities. For GDP and infant mortality, Spearman slightly declines post-peak while bias drops monotonically. We hypothesize that extended refinement induces a conservative prediction regime, narrowing the output distribution and mitigating regional variance. This aligns with fairness-oriented optimization, trading minor precision losses for more equitable geographic treatment.

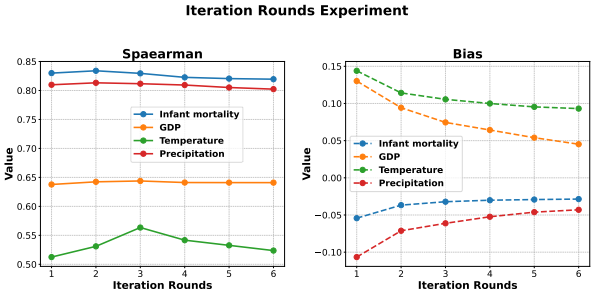


Figure 4: Effect of iterative refinement rounds on model performance.

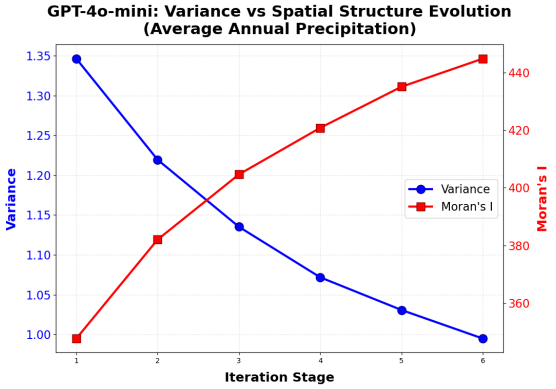


Figure 5: Spatial heterogeneity and convergence metrics across refinement rounds (Precipitation task, GPT-4o-mini).

Spatial Structure Preservation. Iterative refinement reduces prediction variance while **increasing** Moran's I (Figure 5), contradicting oversmoothing where spatial heterogeneity collapses. This indicates that GeoGR^2 amplifies true spatial signals: estimates converge toward *spatially analogous* neighbors per Tobler's First Law, not a homogeneous average. The precipitation task exemplifies this—coastal vs. continental rainfall gradients sharpen as initial noise suppresses.

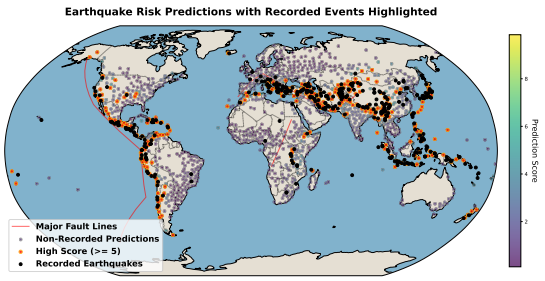


Figure 6: High Earthquake Risk Given by GeoGR²

6 Application: Disaster Prediction

GeoGR² applies to diverse scenarios including urban planning, disaster prediction, and community surveys. We compiled global seismic records (2000–2025) covering **374 cities with earthquake events**, supplemented by 1,091 randomly selected cities, yielding 1,465 cities evaluated via **GeoGR²** (Gemini-3-flash) for seismic risk. Figure 6 shows the spatial distribution: red-shaded areas denote tectonic plate boundaries (historically seismically active), black markers indicate cities with recorded earthquakes, and red markers highlight model-identified high-risk cities. As shown, **GeoGR²** accurately identifies cities facing elevated seismic risk.

7 Conclusion

We presented GeoGR^2, a training-free framework that operationalizes LLMs' latent capacity for spatial correlation understanding through geostatistically-motivated iterative refinement. By dynamically constructing spatial neighborhoods and performing natural language message passing, GeoGR^2 enforces spatial consistency without altering model parameters. Extensive experiments across physical and socioeconomic tasks demonstrate state-of-the-art zero-shot accuracy alongside drastic reductions in geographic bias. Future work will extend this paradigm to spatiotemporal forecasting and multi-scale spatial hierarchies.

Beyond geospatial applications, this work offers broader insights for the LLM community: it demonstrates how non-linguistic structural priors can be effectively injected into LLM inference via natural-language message passing and structured prompting, providing a new paradigm for structured reasoning without parameter updates.

Limitations

Iteration Stopping without Ground Truth. Operating in a zero-shot regime, our framework lacks access to ground-truth labels during inference, precluding objective selection of the optimal iteration count K. The K values reported in our experiments (e.g., $K = 2$ or $K = 3$) are determined through post-hoc analysis on the full dataset, which is unavailable at deployment time. Future work should explore unsupervised stopping criteria, such as monitoring spatial consistency convergence or stabilization of prediction deltas across rounds, to determine termination without labeled data.

Temporal Dimension Extension. The present work is confined to spatial prediction tasks. Incorporating temporal dynamics would significantly broaden applicability, enabling spatiotemporal forecasting for critical domains such as climate trend projection, urban growth modeling, and epidemic spread prediction. Subsequent research should focus on extending the operator design to accommodate time-evolving neighborhood structures and sequential covariate dependencies.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62406206, and by the Fundamental Research Funds for the Central Universities.

References

Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. 2023. Label-free node classification on graphs with large language models (llms). *arXiv preprint arXiv:2310.04668*.

N. Cressie. 1993. *Statistics for Spatial Data*. Wiley, New York.

Google DeepMind and Google Research. 2025. Gemini 3 flash. https://deepmind.google/models/gemini/flash//. Accessed: 2026-01-26.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2024. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 161–170.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2023. Talk like a graph: Encoding graphs for large language models. *arXiv preprint arXiv:2310.04560*.

S. E. Fick and R. J. Hijmans. 2017. WorldClim 2: New 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315.

W. Gurnee and M. Tegmark. 2023. Language Models Represent Space and Time. *Preprint*, arXiv:2310.02207.

TN Kipf. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Erwin Kreyszig. 1991. *Introductory functional analysis with applications*. John Wiley & Sons.

M. Kummu, M. Kosonen, and S. Masoumzadeh Sayyar. 2025. Downscaled gridded global dataset for gross domestic product (GDP) per capita PPP over 1990–2022. *Scientific Data*, 12(1):178.

P. Langley. 2000. Crafting papers on machine learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA. Morgan Kaufmann.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, and 7 others. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5351–5362.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, and Y. Yang. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Proceedings of the Sixty-First Annual Meeting of the Association for Computational Linguistics*, pages 7697–7711, Stroudsburg, Pa. Association for Computational Linguistics.

Rohin Manvi, Samar Khanna, Marshall Burke, David B Lobell, and Stefano Ermon. 2024a. Large language models are geographically biased. In *International Conference on Machine Learning*, pages 34654–34669. PMLR.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024b. GeoLLM: Extracting geospatial knowledge from large language models. In *The Twelfth International Conference on Learning Representations*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Wenbo Shang and Xin Huang. 2025. A survey of large language models on generative graph analytics: Query, learning, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):1–20.

N. Shinn, F. Cassano, B. Labash, A. Gopinath, and V. Ramanujan. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems*, pages 12956–12970, La Jolla, Calif. Neural Information Processing Systems Foundation.

Charles Stein. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics*, volume 3, pages 197–207. University of California Press.

Waldo R Tobler. 1970. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. 2023. Can language models solve graph problems in natural language? *Advances in Neural Information Processing Systems*, 36:30840–30861.

J. Wei, D. Xiong, P. Wang, and M. Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, pages 878–892, La Jolla, Calif. Neural Information Processing Systems Foundation.

Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. 2021. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4478–4485.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

S. Yao and et al. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Jiawei Zhang. 2023. Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. *arXiv preprint arXiv:2304.11116*.

Yifan Zhang, Cheng Wei, Shangyou Wu, Zhengting He, and Wenhao Yu. 2023. Geogpt: Understanding and processing geospatial tasks through an autonomous gpt. *arXiv preprint arXiv:2307.07930*.

A Proof of Proposition 4.1

Proof. Let the ground truth be $\mathbf{y} \in \mathbb{R}^n$ ($n = |\mathcal{V}|$) and the independent zero-shot predictions be $\hat{\mathbf{y}}^{\text{ind}} = \mathbf{y} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The risk (expected total squared error) of the independent estimator is $R(\hat{\mathbf{y}}^{\text{ind}}) = n\sigma^2$.

The GeoGR^2 refinement acts as a linear smoother $\mathbf{S} = \mathbf{I} - \lambda\mathbf{L}$, where $\mathbf{L}$ is the symmetric normalized Laplacian and $\lambda > 0$ is the step size. The new estimator is $\hat{\mathbf{y}}^{\text{agg}} = \mathbf{S}\hat{\mathbf{y}}^{\text{ind}}$. The exact risk of this estimator is the sum of squared bias and variance:

$$\begin{aligned} R(\hat{\mathbf{y}}^{\text{agg}}) = \|(\mathbf{S} - \mathbf{I})\mathbf{y}\|^2 \\ + \sigma^2 \operatorname{tr}(\mathbf{S}\mathbf{S}^T). \end{aligned} \tag{11}$$

The first term is the squared bias, and the second term is the total variance.

Substituting $\mathbf{S} - \mathbf{I} = -\lambda\mathbf{L}$ and using the symmetry of $\mathbf{L}$, we obtain

$$\begin{aligned} \Delta R &= R(\hat{\mathbf{y}}^{\text{agg}}) - R(\hat{\mathbf{y}}^{\text{ind}}) \\ &= \lambda^2\|\mathbf{L}\mathbf{y}\|^2 \\ &\quad + \sigma^2 \operatorname{tr}\big((\mathbf{I} - \lambda\mathbf{L})^2\big) - n\sigma^2. \end{aligned} \tag{12}$$

Expanding the square and using $\operatorname{tr}(\mathbf{I}) = n$ gives

$$\begin{aligned} \Delta R = \lambda^2\|\mathbf{L}\mathbf{y}\|^2 - 2\lambda\sigma^2 \operatorname{tr}(\mathbf{L}) \\ + \lambda^2\sigma^2 \operatorname{tr}(\mathbf{L}^2). \end{aligned} \tag{13}$$

We group terms by powers of λ:

$$\begin{aligned} \Delta R = \lambda^2\big(\|\mathbf{L}\mathbf{y}\|^2 + \sigma^2 \operatorname{tr}(\mathbf{L}^2)\big) \\ - 2\lambda\sigma^2 \operatorname{tr}(\mathbf{L}). \end{aligned} \tag{14}$$

The first term is the quadratic penalty, while the subtracted term represents the linear gain.

Since $\operatorname{tr}(\mathbf{L}) > 0$ for any valid graph structure, the linear reduction term dominates the quadratic penalty term for small λ. Specifically, $\Delta R < 0$ holds strictly whenever:

$$0 < \lambda < \frac{2\sigma^2 \operatorname{tr}(\mathbf{L})}{\|\mathbf{L}\mathbf{y}\|^2 + \sigma^2 \operatorname{tr}(\mathbf{L}^2)}. \tag{15}$$

Thus, there always exists a non-trivial range of λ where GeoGR^2 strictly dominates the independent estimator. □

B On the Distinction from General LLM Graph Reasoning

We clarify that GeoGR^2 is fundamentally a **geospatial prediction framework**, not a contribution

to general-purpose LLM graph reasoning. Standard LLM graph reasoning benchmarks (e.g., node classification on citation networks, link prediction on knowledge graphs) assume a *given, static graph topology* and evaluate the LLM's ability to reason over explicit relational structures (Zhang, 2023; Fatemi et al., 2023; Wang et al., 2023; Chen et al., 2023; Shang and Huang, 2025). In contrast, GeoGR^2 addresses a **zero-shot regression task over latent spatial graphs**, where (1) the graph topology is unknown and must be dynamically discovered from geographic coordinates, (2) node features are natural language descriptions of locations, and (3) the objective is to estimate continuous geospatial variables rather than classify discrete graph properties. Due to this mismatch in task formulation, input modality, and evaluation protocol, direct comparison with general LLM graph reasoning methods is neither feasible nor informative. The graph formalism in GeoGR^2 serves exclusively as a mathematical abstraction for spatial autocorrelation, consistent with classical geostatistical practice (Cressie, 1993; Besag, 1974).

C Algorithm Details

Algorithm 1 GeoGR^2: Iterative Spatial Refinement via Dynamic Neighborhood Aggregation

Require: Locations $\mathcal{L}$, Task $\mathcal{T}$, Iterations K
Ensure: Refined Predictions $\{\hat{y}_t^{(K)}\}$
Initialization: $\hat{y}_t^{(0)} \leftarrow \pi(\ell_t, \mathcal{T})$ for all $\ell_t \in \mathcal{L}$
for $k = 1$ to K **do**
  // Parallel Update Step
  for all $\ell_t \in \mathcal{L}$ **do**
    Topology: Construct neighborhood $\mathcal{N}_t \leftarrow \psi(\ell_t, \mathcal{T})$
    Features: Retrieve covariates $\mathbf{z}_t, \{\mathbf{z}_r\}_{r \in \mathcal{N}_t}$ via ϕ
    Message Passing:
      Gather neighbor states $\Omega_t = \{\hat{y}_r^{(k-1)} \mid \ell_r \in \mathcal{N}_t\}$
      Update $\hat{y}_t^{(k)} \leftarrow \varphi(\hat{y}_t^{(k-1)}, \Omega_t, \mathbf{z}_t, \{\mathbf{z}_r\})$
  end for
end for
return $\{\hat{y}_t^{(K)} \mid \ell_t \in \mathcal{L}\}$

Note: The neighborhood construction function $\psi(\ell_t, \mathcal{T})$ selects 15 reference points through a hybrid strategy: (1) 10 nearest neighbors computed using Haversine distance; (2) up to 5 additional points selected by an LLM from candidates within a 5°×5° geographic window based on prediction diversity and spatial pattern revelatory power.

D Application

In this study, we evaluate the earthquake risk prediction capability of **GeoGR²** against a baseline **GeoLLM** (both implemented on Gemini-3-flash). Using global seismic records from 2000 to 2025, we identified **374 cities** with documented earthquake events. To examine model robustness across varying data scales, we constructed three datasets: a **1:1 balanced set** (374 recorded cities + 374 randomly selected cities without records), a **1:2 extended set** (374 recorded + 748 non-recorded), and a **full set of 1,465 cities** (374 recorded + 1,091 randomly sampled globally). Each city was assigned an earthquake risk score (0–10) via an initial **GeoLLM** prediction based on its name, coordinates, and task description. **GeoGR²** then refined these scores through topology construction, covariate feature selection, and iterative value updating. A city was classified as high-risk if its score ranked in the top 50% of all evaluated cities; prediction was considered a true positive if the city had a recorded earthquake.

Table 6: Earthquake Prediction Given by GeoLLM(Gemini-3-flash)

Metrics	Data Configuration		
	1:1 Ratio (748 points)	1:2 Ratio (1122 points)	Full Dataset (1465 points)
Precision	0.7469	0.5575	0.4198
Recall	0.8128	0.8556	0.8743
Accuracy	0.7687	0.7255	0.6594
F1-score	0.7785	0.6751	0.5672
AUC-ROC	0.8217	0.8184	0.8160

Table 7: Earthquake Prediction Given by GeoGR²(Gemini-3-flash)

Metrics	Data Configuration		
	1:1 Ratio (748 points)	1:2 Ratio (1122 points)	Full Dataset (1465 points)
Precision	0.7781	0.6011	0.4531
Recall	0.7696	0.8895	0.9171
Accuracy	0.7726	0.7626	0.6962
F1-score	0.7738	0.7174	0.6065
AUC-ROC	0.8335	0.8439	0.8480

As detailed in Tables 6 and 7, **GeoGR²** consistently outperforms **GeoLLM** across all configurations. Notably, on the full dataset, **GeoGR²** improves precision from 0.4198 to **0.4531**, recall from 0.8743 to **0.9171**, accuracy from 0.6594 to **0.6962**, F1-score from 0.5672 to **0.6065**, and AUC-ROC from 0.8160 to **0.8480**. These gains highlight **GeoGR²**'s enhanced capacity to integrate structural and covariate information, leading to more reliable and calibrated risk assessments—especially as data scale increases.

E Prompts

Figure 7, Figure 8, Figure 11 and Figure 10 visually represent the prompt design for the initial, topology construction, covariate selection, and update prediction, illustrating how topology structure and covariate message are aggregated to optimize geospatial prediction within the GeoGR² framework.

F Token Count

We measured the token consumption per data point per round, and the results are summarized in Table 8.

Table 8: Token consumption per point per round

Method	Average	Min	Max
GeoLLM	311	184	489
GeoGR²	1,868	1,215	2,597

Prompt for Prediction

You will be given data about a specific location randomly sampled from all human-populated locations on Earth.
You give your **rating** keeping in mind that **it is relative to all other human-populated** locations on Earth (from all continents, countries, etc.).
You provide **ONLY** your answer in the exact format "My answer is X.X." where 'X.X' represents your rating for the given topic.

Coordinates: (48.06208, 86.45292)

Address: "加依勒玛乡, 哈巴河县, 阿勒泰地区, 新疆维吾尔自治区 شىنجاڭ ,رايونى ئالتۇنئوم ئۇيغۇر, 中国"

Nearby Places:
"
2.7 km West: 阿克齐镇
3.0 km North-East: 萨尔塔木乡
3.2 km West: 哈巴河县 قبا
4.3 km North: 加依勒玛乡
10.0 km West: 库勒拜镇
18.7 km East: 克依克拜村
26.0 km East: 哈夏库都克
28.5 km North-East: 吐孜拜
28.9 km North-East: 阔勒送能村
29.0 km South-West: 齐巴尔镇
"

Annual GDP per capita(On a Scale from 0.0 to 9.9):

COMPLETION: My answer is 5.5.

RATING: 5.5

Figure 7: The prompt for the initial prediction.

Prompt for Topology Operator

As a **geospatial selection expert**, select up to 5 reference points from the 5x5 degree area around the target:

Target Point:
- **Coordinates**: (X, Y)
- **Prediction**: X.XX
- **Task: Annual GDP per capita**

Candidate Points (Total 6 points in 5x5 degree area):
1. Coordinates: (X1, Y1) Prediction: X.X | BIO1=X.X, BIO12=X.X, BIO15=X.X, BIO7=X.X, BIO4=X.X
2. Coordinates: (X2, Y2) Prediction: X.X | BIO1=X.X, BIO12=X.X, BIO15=X.X, BIO7=X.X, BIO4=X.X
.....

Selection criteria:
1. **Select points** that are **MOST REVEALING for spatial patterns**
2. **Prioritize** points with **significant prediction differences**
3. Ensure **diversity** (avoid clustering)

IMPORTANT OUTPUT RULES:
1. YOU MUST output EXACTLY one line starting with "Selected order numbers:"
2. Format: Selected order numbers: [comma-separated numbers]
   Example: Selected order numbers: 3,7,12,15,22
3. DO NOT include any additional text, explanations, or formatting symbols

Figure 8: The prompt for the topology construction.

Prompt for Feature Operator

As a geospatial climate expert, select the MOST RELEVANT bioclimatic variables for refining predictions at the target location: Target Location:
- **Coordinates**: (X.XXXX, X.XXXX)
- **Task: Annual GDP per capita**

Available Bioclimatic Variables:
- BIO1: Annual Mean Temperature
- BIO2: Mean Diurnal Range (Mean of monthly (max temp – min temp))
- BIO3: Isothermality (BIO2/BIO7) (100)
- BIO4: Temperature Seasonality (standard deviation 100)
- BIO5: Max Temperature of Warmest Month
- BIO6: Min Temperature of Coldest Month
- BIO7: Temperature Annual Range (BIO5-BIO6)
- BIO8: Mean Temperature of Wettest Quarter
- BIO9: Mean Temperature of Driest Quarter
- BIO10: Mean Temperature of Warmest Quarter
- BIO11: Mean Temperature of Coldest Quarter
- BIO12: Annual Precipitation
- BIO13: Precipitation of Wettest Month
- BIO14: Precipitation of Driest Month
- BIO15: Precipitation Seasonality (Coefficient of Variation)
- BIO16: Precipitation of Wettest Quarter
- BIO17: Precipitation of Driest Quarter
- BIO18: Precipitation of Warmest Quarter
- BIO19: Precipitation of Coldest Quarter

Selection Guidelines:
1. Consider the target's **geographic context** (e.g., coastal, mountainous, tropical)
2. Prioritize variables that **explain spatial patterns** for Annual GDP per capita
3. Select 1-5 variables that provide complementary and useful information for the task
4. Justify each selection based on climatic relevance

STRICT OUTPUT RULES:
1. MUST output exactly one line starting with "Selected Variables:"
2. Format: Selected Variables: [comma-separated BIO codes]
3. NO additional text, explanations, or formatting symbols

Figure 9: The prompt for the covariates selection.

Prompt for Update Operator

Some information you need to know:
"BIO1": "Annual Mean Temperature",
"BIO2": "Mean Diurnal Range (Mean of monthly (max temp - min temp))",
"BIO3": "Isothermality (BIO2/BIO7) (100)",
.....
"BIO18": "Precipitation of Warmest Quarter",
"BIO19": "Precipitation of Coldest Quarter"
As a geospatial prediction expert, refine the initial rating based on **spatial patterns and bioclimatic context**:

Task: Annual GDP per capita
Target Location: (X, Y)
Initial Prediction: X.XX | BIO1=X.X, BIO12=X.X, BIO15=X.X, BIO7=X.X, BIO4=X.X

Nearby Points (with bioclimatic variables):
- (X1, Y1): Annual GDP per capita prediction=X.XX | BIO1=X.X, BIO12=X.X, BIO15=X.X, BIO7=X.X, BIO4=X.X
- (X2, Y2): Annual GDP per capita prediction=X.XX | BIO1=X.X, BIO12=X.X,
.....
Instructions:
1. Analyze **regional bioclimatic patterns**
2. Compare target and reference points' **environmental context**
3. Propose scientifically justified adjustment

STRICT OUTPUT RULES:
1. MUST output exactly one line starting with "Final Prediction:"
2. Format: Final Prediction: [X.XX] (X.XX must be float with 2 decimals between 0.00 to 10.00)
3. NO additional text, explanations, or formatting symbols
4. You don't have to refine the prediction if you are not confident in your answer

Figure 10: The prompt for the update prediction.

Prompt for ReAct

You are a geospatial prediction expert using the ReAct (Reasoning + Action) framework.

CONTEXT INFORMATION
{context}

AVAILABLE ENVIRONMENTAL VARIABLES
The following bioclimatic variables are available for analysis (variables that would leak ground truth have been excluded):
- BIO1: Annual Mean Temperature
...
...
- BIO19: Precipitation of Coldest Quarter

YOUR TASK
Refine the prediction for the target location by:
1. Analyzing spatial patterns from the 10 nearest reference points
2. Comparing environmental conditions (BIO variables) between target and references
3. Reasoning about how these factors should influence the final prediction

REACT FRAMEWORK INSTRUCTIONS
In each iteration, you MUST follow this format:

Step 1 - THOUGHT: Analyze the current information and reason about what needs to be done
Step 2 - ACTION: Choose one of the following actions:
- THINK
- CONTINUE
- QUERY_BIO: Query specific BIO variable values for target or reference points
- FINISH

Step 3 - PARAMETERS (only when ACTION is QUERY_BIO):
- BIO_VAR: The BIO variable to query (e.g., BIO1, BIO12)
- POINT_TYPE: Either 'target' or 'reference'
- POINT_IDX: Reference point number (1-10), only required when POINT_TYPE is 'reference'

Step 4 - FINAL PREDICTION (only when ACTION is FINISH): a number with exactly 2 decimal places

BEGIN REASONING
Now begin your analysis for Iteration {iteration}. Remember to think step-by-step and provide clear reasoning.

Figure 11: The prompt for ReAct, one of the baseline methods.

G Use of AI assistant

We acknowledge the use of large language models (GPT-5.2 and Claude) for language polishing, grammar correction, and stylistic refinement during the preparation of this manuscript. All scientific content, research findings, data analysis, and conclusions were conceived, verified, and approved solely by the authors. The AI tools were employed exclusively for enhancing readability and linguistic clarity, without altering the substantive scientific meaning or intellectual contributions of the work.

H Supplementary Experiment

H.1 Absolute Numerical Prediction.

While our main experiments use 1-10 relative scoring to mitigate hallucination in weaker LLMs, stronger models can directly output calibrated physical units. As shown in Table 9, Gemini-3-flash + GeoGR^2 achieves an R^2 of 0.787 and MAE of 6.11 on Infant Mortality, vastly outperforming supervised baselines ($N = 100$), demonstrating the framework's practical utility for real-world deployment.

Method	**Spearman**	**RMSE**	**MAE**	R^2
Kriging	0.663	17.00	12.39	0.438
GP (RBF)	0.752	14.14	9.14	0.633
GCN	0.807	16.15	10.03	0.522
Gemini-3-Flash	0.894	11.72	6.93	0.749
+ GeoGR^2	**0.926**	**10.52**	**6.11**	**0.787**

Table 9: Absolute Numerical Prediction on Infant Mortality ($N = 100$ for baselines).

I More Visualizations

We provide visualizations of all models in the Figure 12, Figure 13 and Figure 14, and we find that GeoGR² consistently enhances the overall performance of the models. In some tasks, such as those involving GPT-3.5-Turbo, GeoGR² significantly improves the model's effectiveness. However, we observe that large language models still exhibit tendencies to systematically overestimate or underestimate certain regions; for instance, all models tend to overestimate China's GDP, and although GeoGR² mitigates this to some extent, the overall bias remains upward.

J More Ablation Studies on Point Selection Strategy

In the main text, we provided an ablation study for GPT-3.5-turbo. We also included results from applying the Point Selection Strategy to three additional models: DeepSeek V3, GPT-4o-mini, and GeoGPT. The findings in Table 10, Table 11, and Table 12 show that the impact of the Point Selection Strategy on these models is consistent with its effect on GPT-3.5-turbo, with comparable overall performance across all models.

Table 10: Ablation Study on Point Selection Strategy for DeepSeek-V3

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o N_t^{prox}	Spearman	0.822	0.632	0.504	0.809
	Bias	-0.057	0.192	0.106	0.097
GeoGR² w/o N_t^{sem}	Spearman	0.830	0.641	0.560	0.815
	Bias	-0.021	0.057	0.089	0.044
GeoGR²	Spearman	**0.834**	**0.644**	**0.563**	**0.813**
	Bias	-0.026	0.063	0.097	0.061

K More Ablation Studies on External Variable

We use the 19 bioclimatic variables (bio1–bio19) from WorldClim v2.1[†] as spatially explicit covariates. Derived from long-term monthly temperature and precipitation records, these indices capture both mean climatic conditions and seasonal extremes. Table 13 provides concise ecological definitions.

[†] https://www.worldclim.org/

Table 11: Ablation Study on Topology Construction Strategy for GPT-4o-mini

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o N_t^{prox}	Spearman	0.730	0.581	0.504	0.631
	Bias	-0.052	0.192	0.102	0.097
GeoGR² w/o N_t^{sem}	Spearman	0.744	0.577	0.510	0.626
	Bias	-0.009	0.057	0.047	0.044
GeoGR²	Spearman	**0.762**	**0.585**	**0.517**	**0.634**
	Bias	-0.011	0.087	0.085	0.011

Table 12: Ablation Study on Topology Construction Strategy for GeoGPT

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o N_t^{prox}	Spearman	0.801	0.517	0.758	0.799
	Bias	-0.022	0.127	0.112	0.160
GeoGR² w/o N_t^{sem}	Spearman	0.811	0.525	0.769	0.807
	Bias	-0.011	0.088	0.099	0.144
GeoGR²	Spearman	**0.816**	0.523	**0.771**	0.802
	Bias	-0.031	0.146	0.108	0.157

In the main text, we provided an ablation study for GPT-3.5-turbo. We also included results from applying the Variable Selection Strategy to three additional models: DeepSeek V3, GPT-4o-mini, and GeoGPT. The findings in Table 14, 15, and 16 show that the impact of the Variable Selection Strategy on these models is consistent with its effect on GPT-3.5-turbo, with comparable overall performance across all models.

L More Ablation Studies on Iteration Rounds Experiment

we also provide iteration rounds experiment results of three additional models: GeoGPT, GPT-4o-mini, and GPT-3.5-Turbo.As shown in Figure 15, 16 and 17, the Spearman values of all models reached their maximum after several iterations and then began to decline, while the bias scores continued to decrease with further iterations.

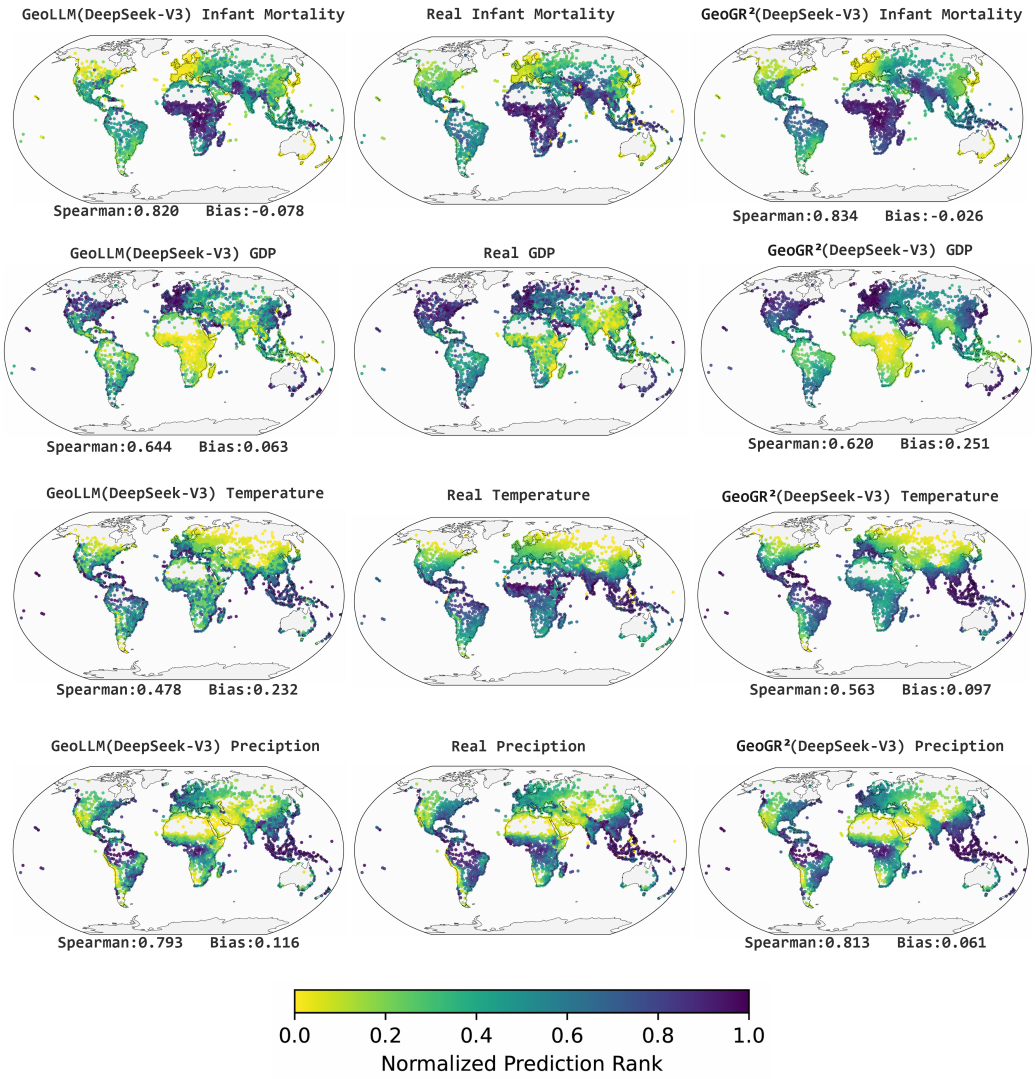


Figure 12: Spatial prediction rank maps using DeepSeek-V3. **Left**: Predictions from the base GeoLLM. **Middle**: Ground-truth. **Right**:Predictions from GeoGR²-enhanced model.

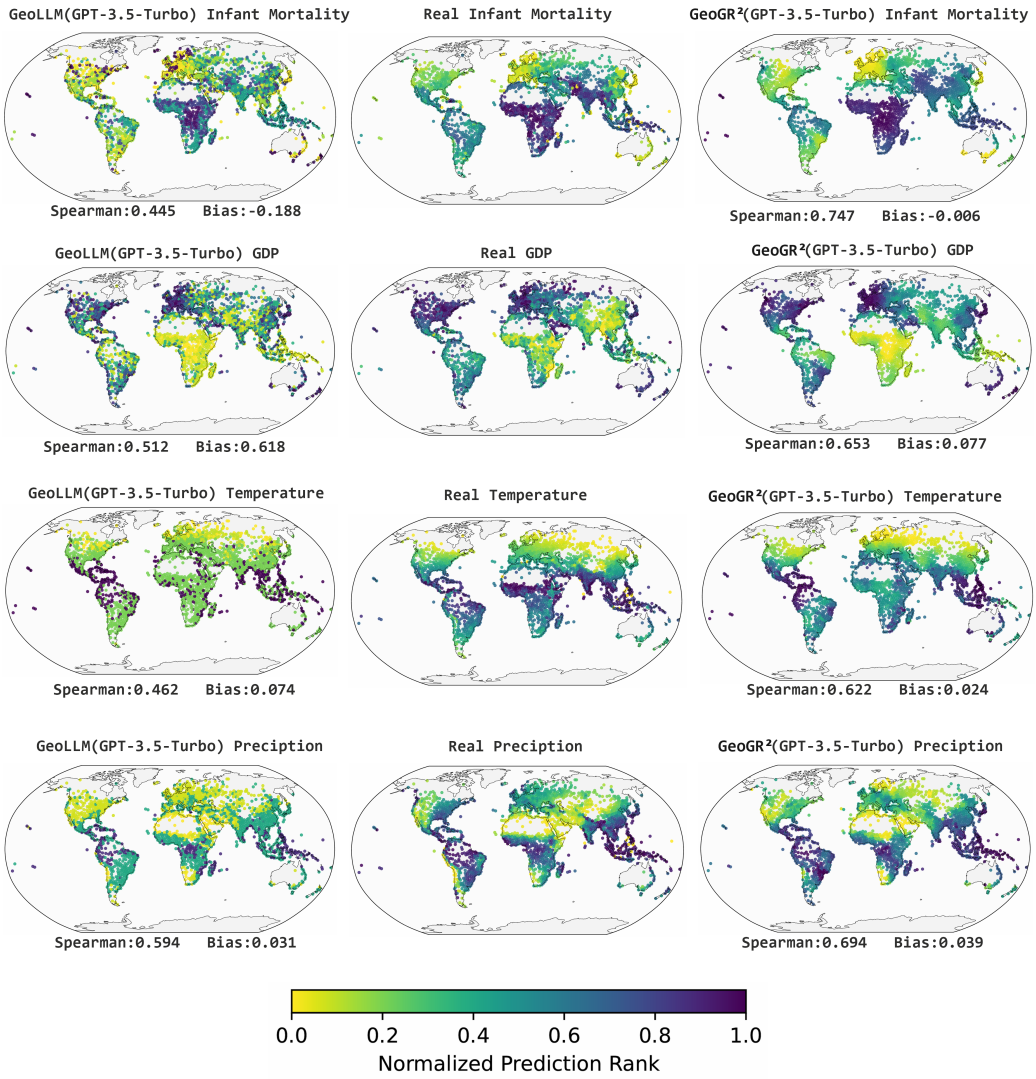


Figure 13: Spatial prediction rank maps using GPT-3.5-Turbo. **Left**: Predictions from the base GeoLLM. **Middle**: Ground-truth. **Right**: Predictions from GeoGR²-enhanced model.

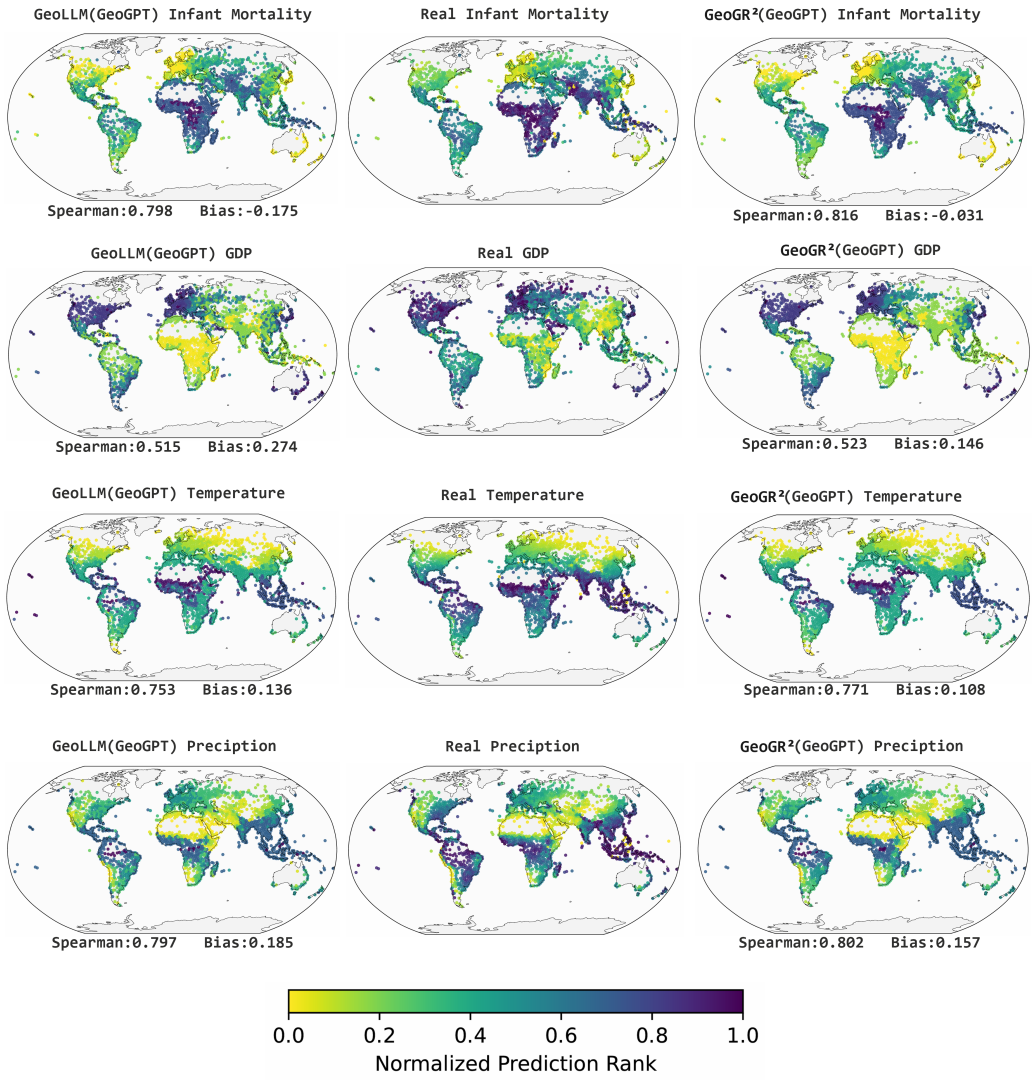


Figure 14: Spatial prediction rank maps using GeoGPT. **Left**: Predictions from the base GeoLLM. **Middle**: Ground-truth. **Right**: Predictions from GeoGR²-enhanced model.

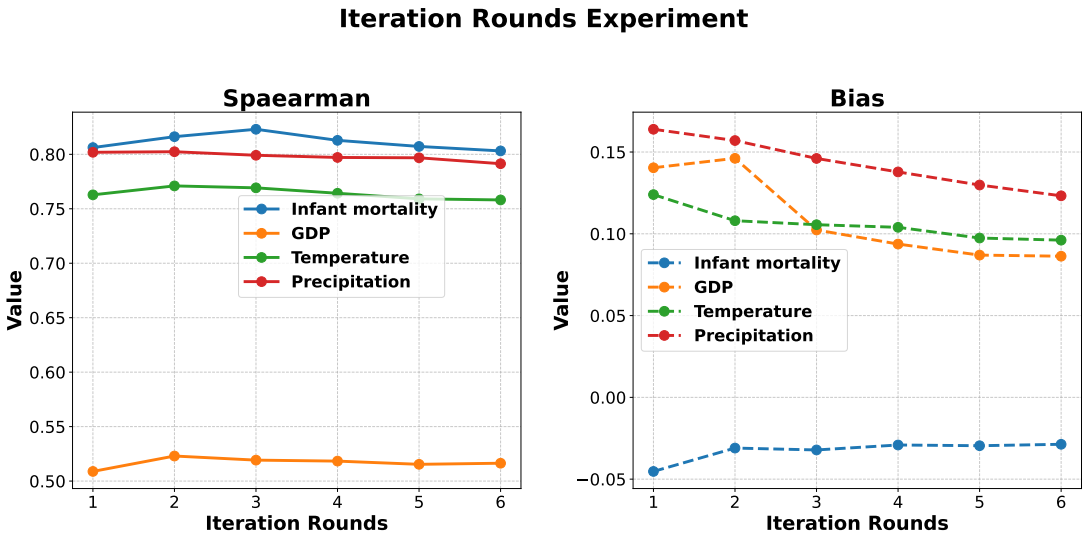


Figure 15: Iteration Rounds Experiment(GeoGR²-enhanced GeoGPT).

Table 13: WorldClim bioclimatic variables used in this study.

ID	Variable	Ecological relevance
Temperature-related (energy)		
bio1	Annual Mean Temperature	Baseline thermal energy
bio2	Mean Diurnal Range	Daily temperature stability
bio3	Isothermality	Shape of thermal variability
bio4	Temperature Seasonality	Intensity of thermal pulses
bio5	Max Temperature of Warmest Month	Acute heat stress
bio6	Min Temperature of Coldest Month	Acute cold stress
bio7	Temperature Annual Range	Climatic buffering capacity
bio8	Mean Temp. of Wettest Quarter	Thermal conditions under high moisture
bio9	Mean Temp. of Driest Quarter	Thermal conditions under low moisture
bio10	Mean Temp. of Warmest Quarter	Prolonged heat exposure
bio11	Mean Temp. of Coldest Quarter	Prolonged cold exposure
Precipitation-related (water)		
bio12	Annual Precipitation	Total water input
bio13	Precipitation of Wettest Month	Moisture extremes
bio14	Precipitation of Driest Month	Moisture extremes
bio15	Precipitation Seasonality	Drought risk
bio16	Precipitation of Wettest Quarter	Seasonal water allocation
bio17	Precipitation of Driest Quarter	Seasonal water allocation
bio18	Precipitation of Warmest Quarter	Temperature–precipitation coupling
bio19	Precipitation of Coldest Quarter	Temperature–precipitation coupling

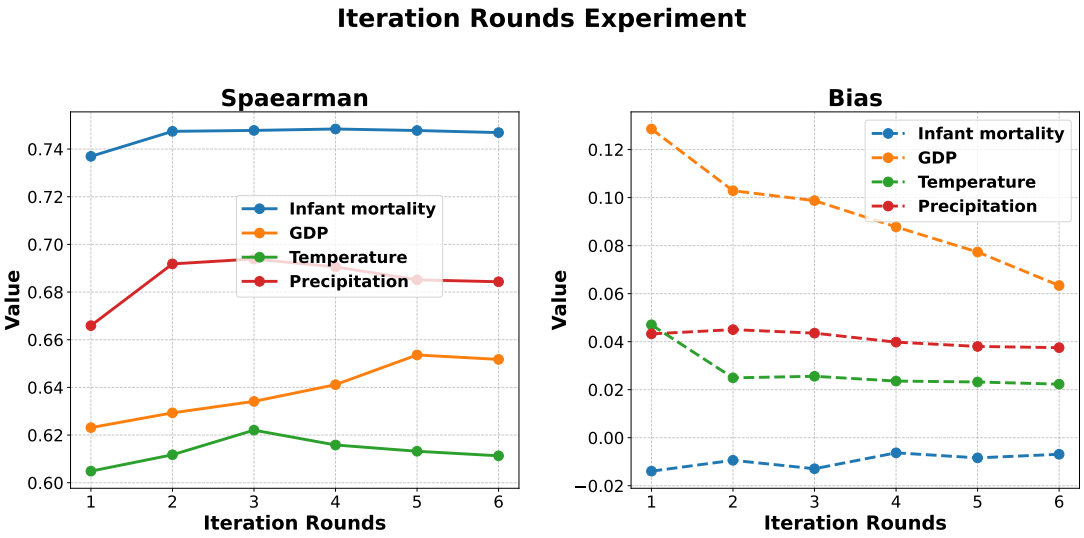


Figure 16: Iteration Rounds Experiment(GeoGR²-enhanced GPT-3.5-Turbo).

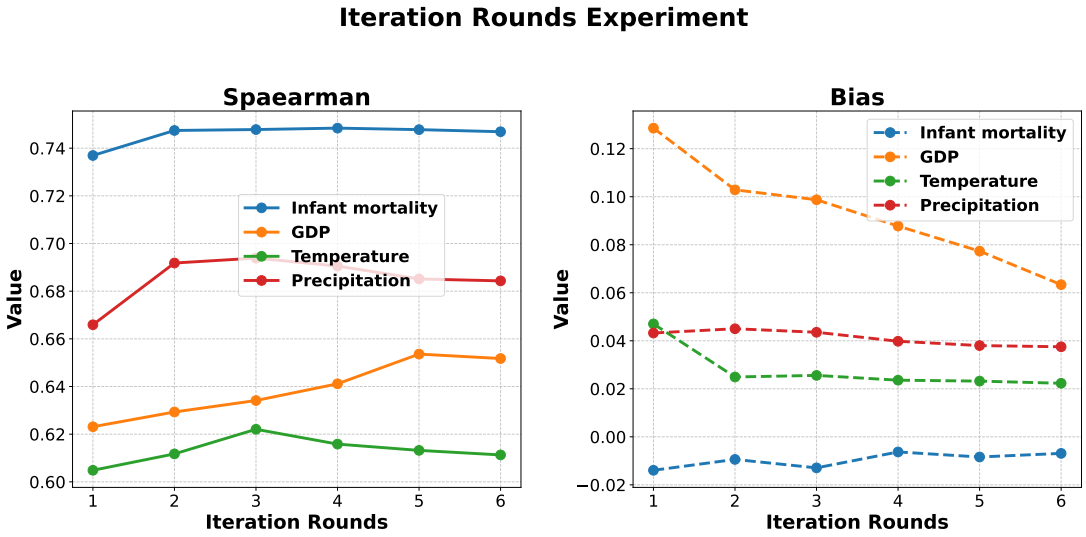


Figure 17: Iteration Rounds Experiment(GeoGR²-enhanced GPT-4o-mini).

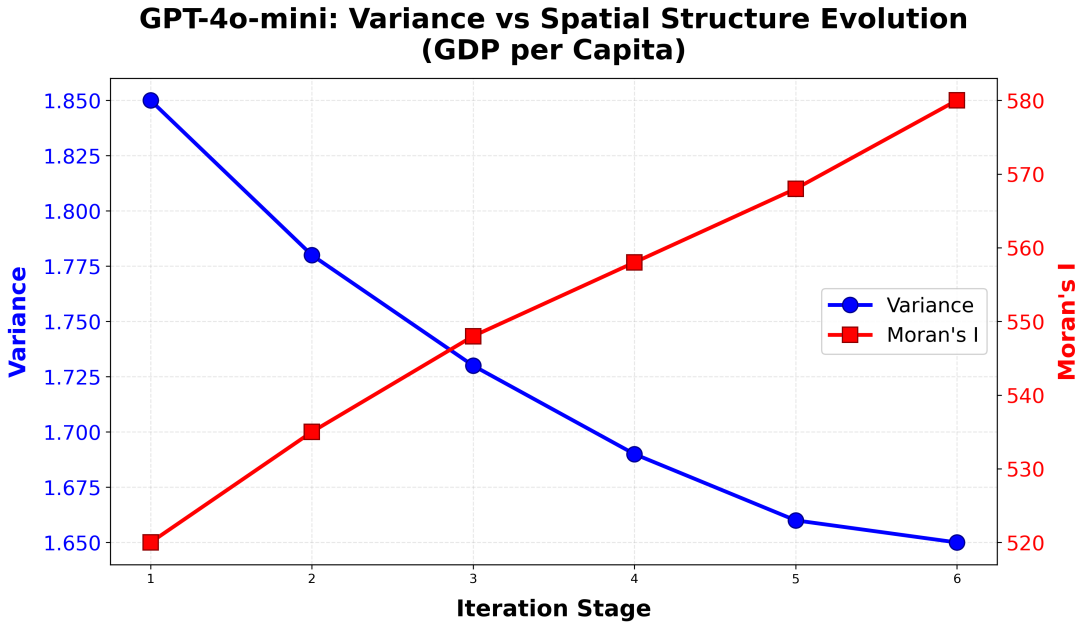


Figure 18: patial heterogeneity and convergence metrics across refinement rounds (GDP task, GPT-4o- mini).

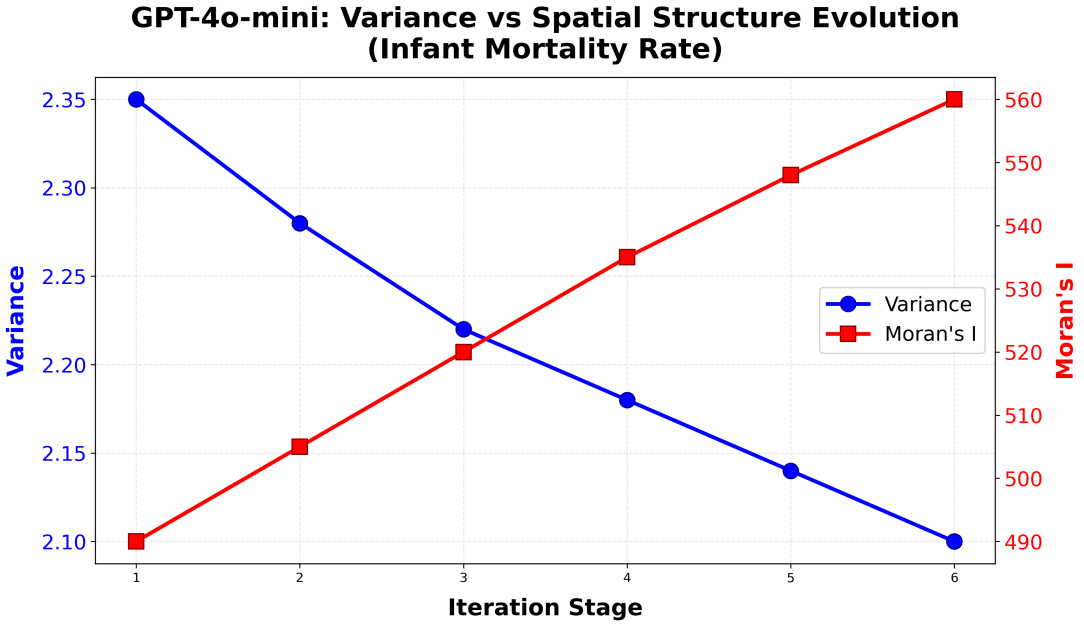


Figure 19: patial heterogeneity and convergence metrics across refinement rounds (Infant Mortality task, GPT-4o-mini).

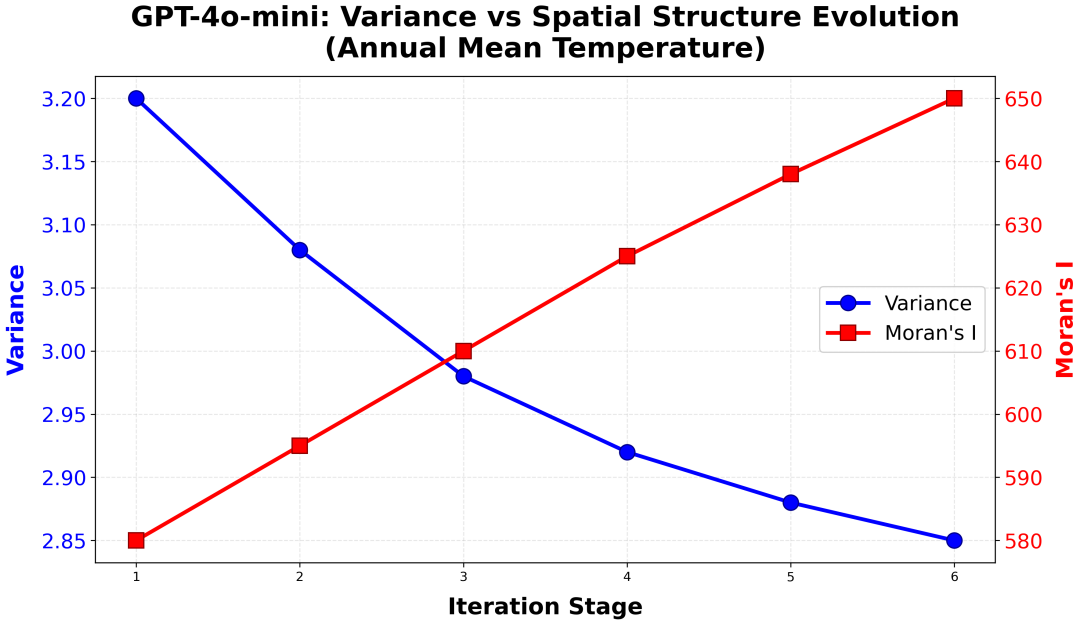


Figure 20: patial heterogeneity and convergence metrics across refinement rounds (Temperature task, GPT-4o-mini).

Table 14: Ablation Study on External Variables for DeepSeek-V3

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o external variables	Spearman	0.825	0.628	0.579	0.799
	Bias	-0.036	0.093	0.101	0.086
GeoGR²	Spearman	**0.834**	**0.644**	**0.563**	**0.813**
	Bias	**-0.026**	**0.063**	**0.097**	**0.061**

Table 15: Ablation Study on External Variables for GPT-4o-mini

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o external variables	Spearman	0.710	0.554	0.524	0.601
	Bias	-0.033	0.091	0.062	0.059
GeoGR²	Spearman	**0.762**	**0.585**	**0.517**	**0.634**
	Bias	**-0.011**	**0.087**	0.085	**0.011**

Table 16: Ablation Study on External Variables for GeoGPT

Method Variant		Tasks			
Method	Metric	Infant mort.	GDP	Temp.	Precip.
GeoGR² w/o external variables	Spearman	0.801	0.517	0.781	0.800
	Bias	-0.043	0.095	0.134	0.116
GeoGR²	Spearman	**0.816**	**0.523**	0.771	**0.802**
	Bias	**-0.031**	0.146	**0.108**	0.157