POLITECNICO DI TORINO
Repository ISTITUZIONALE

Climate Knowledge in Large Language Models

Original
Climate Knowledge in Large Language Models / Kuznetsov, I., Grassi, J., Pantiukhin, D., Shapkin, B., Jung, T., Koldunov, N.. - In: ARTIFICIAL INTELLIGENCE FOR THE EARTH SYSTEMS. - ISSN 2769-7525. - 5:4(2026). [10.1175/aies-d-26-0017.1]

Availability:
This version is available at: 11583/3016227 since: 2026-10-27T13:32:20Z

Publisher:
American Meteorological Society

Published
DOI:10.1175/aies-d-26-0017.1

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

15 November 2026

Climate Knowledge in Large Language Models

IVAN KUZNETSOV,[a] JACOPO GRASSI,[a,b] DMITRII PANTIUKHIN,[a] BORIS SHAPKIN,[a] THOMAS JUNG,[a,c] AND NIKOLAY KOLDUNOV[a]

[a] *Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany*
[b] *Department of Environment, Land, and Infrastructure Engineering, Politecnico di Torino, Turin, Italy*
[c] *Institute of Environmental Physics, University of Bremen, Bremen, Germany*

(Manuscript received 25 February 2026, in final form 5 July 2026, accepted 21 September 2026)

ABSTRACT: Large language models (LLMs) are increasingly deployed for climate-related applications, where understanding internal climatological knowledge is important for assessing their reliability. Despite growing adoption, the capacity of LLMs to recall climate normals from parametric knowledge (information stored in the model's weights, without external retrieval) remains largely uncharacterized. We investigate this capacity focusing on a prototypical query: July mean 2-m air temperature, 1991–2020, at specified locations. We construct a global grid of queries at 1° resolution land points, providing coordinates and location descriptors, and validate responses against ERA5 reanalysis. We further extend the evaluation to the 12 monthly normals for 4 of our best-performing LLMs on a subset of the grid. Results show that LLMs reproduce broad geographic regularities in climatological temperature fields, capturing latitudinal and topographic patterns, with July root-mean-square errors of 3°–6°C and biases of ±1°C; errors are comparable across the annual cycle. However, spatially coherent errors remain: Performance degrades sharply above 1500 m, where root-mean-square error (RMSE) reaches 5°–13°C compared to 2°–4°C at lower elevations. Including geographic context (country, city, region) reduces errors by 27% on average. While models capture the global mean magnitude of observed warming between 1950–74 and 2000–24, they fail to reproduce spatial patterns of temperature change. This limitation highlights that while LLMs may capture present-day climate distributions, they struggle to represent the regional and local expression of long-term shifts in temperature essential for understanding climate dynamics. Our evaluation framework provides a reproducible benchmark for quantifying parametric climate knowledge in LLMs.

SIGNIFICANCE STATEMENT: People increasingly ask artificial intelligence chatbots about climate and weather at specific places. We tested whether these systems can recall basic climate facts, specifically average July temperatures worldwide, from what they absorbed during training, without access to any external databases. Chatbots capture broad temperature patterns (warm tropics, cool poles) with typical errors of 3°–6°C but struggle in mountainous areas where errors can exceed 10°C. Providing place names alongside coordinates reduced errors by about 27%, and accuracy was similar across all 12 months. Critically, while chatbots approximate the overall magnitude of recent global warming, they cannot reproduce its spatial structure. These findings highlight that chatbot-generated climate information should be verified against authoritative data sources.

KEYWORDS: Atmosphere; Climate; Artificial intelligence; Machine learning; Other artificial intelligence/machine learning

1. Introduction

Large language models (LLMs) have demonstrated substantial capacity to store and recall factual information acquired during pretraining (Kandpal et al. 2023; Yuan et al. 2024; Wang et al. 2024). As these systems are increasingly deployed as assistants in scientific domains, understanding the scope and reliability of their internal knowledge becomes critical.

In climate science, users frequently seek factual information about temperature, precipitation, or other variables at specific locations and time periods. Modern climate applications of LLMs have begun to address this through retrieval-augmented generation, where external authoritative sources ground model responses (Vaghefi et al. 2023; Koldunov and Jung 2024; Kuznetsov et al. 2025; Bulian et al. 2024). For instance, ChatClimate integrates content from IPCC reports to answer climate questions (Vaghefi et al. 2023), while ClimSight combines LLM reasoning with climate model databases to provide localized projections (Koldunov and Jung 2024; Kuznetsov et al. 2025). These tool-augmented systems underscore the value of external knowledge—but many real-world interactions with LLMs occur in closed-book (no retrieval of external data) settings where the model relies solely on parametric knowledge.

Although consumer chat interfaces increasingly trigger a web or tool call for such queries, this behavior is neither universal nor guaranteed: Many LLM-based applications operate without retrieval, and even a tool-enabled model may forgo a call, for reasons of latency or cost, and answer directly from parametric memory. Closed-book responses are moreover the default for the open-weight and fine-tuned local deployments (Llama, Gemma, Mistral, GPT-OSS). Retrieving an authoritative value is in any case nontrivial: It requires an account and accepted

Denotes content that is immediately available upon publication as open access.

Corresponding author: Ivan Kuznetsov, ivan.kuznetsov@awi.de

DOI: 10.1175/AIES-D-26-0017.1 e260017
© 2026 American Meteorological Society. This published article is licensed under the terms of a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

license with a data provider, such as Copernicus, Earthmover, or hosted ERA5 collections, identification of the appropriate dataset, and a retrieval step that can take minutes to hours; end-to-end systems of this kind are only beginning to be built (Pantiukhin et al. 2026).

This raises a fundamental question: What climatological information does an LLM implicitly encode from its training data, and how reliable is it? Recent work has examined LLMs' factual recall capabilities across diverse domains (Kandpal et al. 2023; Yuan et al. 2024; Wang et al. 2024; Min et al. 2023; Zhang et al. 2023; Tonmoy et al. 2024; Chang et al. 2024). Studies show that while LLMs can retrieve many facts stored in their parameters, accuracy degrades for rare or fine-grained information, the so-called long tail of knowledge (Wang et al. 2024; Chang et al. 2024). Models also exhibit challenges in maintaining factual consistency (Min et al. 2023; Zhang et al. 2023) and can hallucinate when uncertain (Tonmoy et al. 2024). However, existing evaluations of LLMs' geoscientific knowledge focus on general geographical facts (e.g., populations, elevations) or climate information retrieved via external sources; to our knowledge, no prior work has systematically assessed LLMs' ability to recall location-specific numeric climate data, such as climate normals, from their parametric memory.

We investigate this question by asking contemporary LLMs a simple, prototypical query: "What is the mean July 2-m air temperature at a location (latitude, longitude) for the period 1991–2020?" This query mimics a common request to a climate data archive but is formulated in natural language to elicit a single numeric answer. By posing such questions at 15 395 land locations on a global 1° grid (ocean points excluded) and comparing responses to reanalysis data, we can assess the extent to which these models have absorbed climate norms during pretraining. This constitutes a closed-book climatology challenge where each location has an unambiguous correct value that can be verified against reference datasets. We use ERA5 reanalysis (1991–2020) as our ground truth (Hersbach et al. 2023). We then test whether this July skill generalizes across the annual cycle by querying the 12 monthly means for 4 of our best-performing models on a representative 1540-point land subset, retaining the full-grid July experiment as the primary benchmark.

The motivation is both scientific and practical. If LLMs encode substantial climate statistics, they might correctly answer common climate queries without external tools. If they do not, any confident answers could be misleading. Understanding this balance is important for deploying LLMs in climate-related applications where misinformation carries high stakes. By isolating the LLM in closed-book mode (no tools or retrieval allowed), our evaluation specifically probes parametric knowledge, the information about climate stored in internal model weights.

Numerous frameworks now evaluate LLMs on diverse capabilities (Liang et al. 2023; Chiang et al. 2024; Zheng et al. 2023; Srivastava et al. 2022), with recent efforts focusing on scientific domains (Feng et al. 2024; Sun et al. 2023). However, climate-specific evaluations have primarily assessed textual understanding and communication (Bulian et al. 2024; Thulke et al. 2024) rather than quantitative recall of climate variables. Our work complements these by requiring exact numeric answers for objectively verifiable facts, enabling straightforward spatial error analysis. Our demonstration combines a full-grid July benchmark with a 12-month subset benchmark for 4 of our best-performing models, and the same framework extends naturally to additional variables (e.g., precipitation, wind, extremes) and to comparisons between closed-book and retrieval-augmented modes. We provide this as a reproducible evaluation framework that others can adopt and extend.

Our evaluation offers a new lens to examine LLMs on quantitative scientific knowledge. We emphasize that our goal is to characterize what the model knows, not to claim any predictive capability or physical understanding. The framework facilitates objective comparisons across models and can track progress as newer models are released or as domain-specific models are fine-tuned on climate texts (Thulke et al. 2024).

2. Methods

We generated a global grid at 1° latitude–longitude resolution spanning longitudes from −180° to 180° and latitudes from −60° to 85°. Antarctica, which lies south of 60°S, is therefore excluded from the evaluation domain, as are land areas north of 85°N. From this grid, we selected 15 395 land points, excluding ocean cells. For each grid point, we extracted from ERA5 reanalysis (Hersbach et al. 2023) July monthly mean 2-m air temperature for the period 1991–2020 to serve as ground truth. We used a geographic reverse-geocoding service to associate each coordinate with the nearest country, state/region, and city when available (Nominatim 2025).

We queried models with standardized prompts (Fig. 1) that requested only a numeric answer in degrees Celsius. No few-shot examples or chain-of-thought prompting was used. To assess variability, we queried each model exactly 10 times per location. If any individual request failed to return a valid numeric response, we requeried that specific request up to three additional times. Nonnumeric or out-of-range responses (<-90°C or >60°C) were flagged during this process. The final prediction for each location was taken as the mean of the 10 successful responses. We tested models at two sampling temperature settings, where the sampling temperature controls how random the output is: temperature = 0 (deterministic), which always returns the single most probable value, and temperature = 0.3 (slightly stochastic), which allows small variation between repeated queries.

For evaluation metrics reported in Table 2, we computed root-mean-square error (RMSE) and mean bias directly between model predictions and ERA5 values across all land points without area weighting. This provides a straightforward assessment of pointwise accuracy. We treat RMSE as the primary accuracy metric because the domain-mean bias can be small even when large regional errors are present but cancel each other due to opposite signs; mean bias is therefore reported as a complementary indicator of a net warm or cold tendency, and the spatial structure of the bias is examined through the bias maps (Fig. 4). For spatial visualization in bias maps (Fig. 4), we employed neighborhood-averaged bias calculations following standard practice in climate model evaluation.

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

You are a climate data expert. Given the location coordinates and address information below, provide the mean July temperature for the period 1991-2020. Location Information: - Longitude: {{longitude}} - Latitude: {{latitude}} - Country: {{country}} - State/Region: {{state}} - City: {{city}} Provide ONLY the mean July temperature at 2m above surface (°C) for this location for the climatological period 1991-2020. IMPORTANT: Return ONLY a single number (float) representing the mean July temperature in Celsius. No text, no JSON, just the number. Example: 25.4	You are a climate data expert. Given the location coordinates and address information below, calculate the change in mean July temperature between the periods 1950-1974 and 2000-2024. Location Information: - Longitude: {{longitude}} - Latitude: {{latitude}} - Country: {{country}} - State/Region: {{state}} - City: {{city}} Provide ONLY the temperature difference (2000-2024 minus 1950-1974) at 2 m above surface (°C) for this location. IMPORTANT: Return ONLY a single number (float) representing the temperature difference in Celsius. No text, no JSON, just the number. Example: 1.8

FIG. 1. Prompt templates. (left) Prompt for climatological baseline (1991–2020). (right) Prompt for climate change (temperature difference, 2000–24 minus 1950–74).

Specifically, we assessed spatial consistency using a neighborhood scheme: For each land grid point, metrics were computed over a 5×5 window (radius = 2 grid cells) with latitude clipping, longitudinal wraparound, and a requirement of at least 6 valid neighbors (Gilleland et al. 2009). Windows not meeting the neighbor threshold were set to missing, and ocean cells were excluded. This spatial smoothing highlights coherent regional error patterns and accounts for the varying area represented by grid cells at different latitudes.

To evaluate seasonal generalization, we repeated the baseline experiment for all 12 calendar months for GPT-5, GPT-OSS 120B, Mistral-small-3.1 24B, and Gemma 27B. To limit cost, the monthly experiment used 1540 land points obtained by retaining every 10th point of the original 15 395-point grid; because this subset is nested within the original grid, the published July runs can be restricted to the same points to assess representativeness. The monthly prompt requested the 1991–2020 mean 2-m air temperature for January–December as 12 comma-separated values, and responses were retained only when all 12 values were numeric and physically plausible. Each point was queried 10 times, and the prediction was taken as the mean of valid responses. The monthly runs used the application programming interface (API)-default sampling setting, with the sampling-temperature parameter left unset, and predictions were compared pointwise with the ERA5 1991–2020 monthly climatology at the nearest grid cell.

To interpret the error magnitude relative to observed variability, we also computed the ERA5 interannual variability at the 1540 points. For each point and month, individual yearly monthly means were expressed as departures from the 1991–2020 monthly mean, and the root mean square of these departures was pooled over points. This quantity describes the natural year-to-year variability of the target month and is used as an observational reference scale, not as a model-error metric; averaged over all point–month pairs, it is 2.00°C.

To assess the influence of the sampling setting and of the subset design, we repeated the July task at three settings of the sampling temperature (the LLM parameter controlling output randomness): 0, 0.3, and the API default, where available, comparing both single-month July prompts and the July component of the joint 12-month prompt.

To support reproducibility, Table 1 records the exact local image identifiers and quantization of the open-weight models, which were run locally through Ollama. For closed-weight models, reproducibility was pinned to the provider-exposed model identifier and snapshot/version rather than to local image digests. The GPT-5 family was accessed as gpt-5, gpt-5-mini, and gpt-5-nano, all corresponding to version 7 August 2025. Gemini was accessed as gemini-2.5-flash-lite-preview-09-2025, the September 2025 Flash-Lite preview endpoint. Provider-internal build identifiers, parameter counts, and quantization schemes were not publicly disclosed.

3. Results

a. *Present-day climatology*

The LLM-based July mean 2-m air temperature climatology is shown in Fig. 2, with GPT-5 as a representative model alongside the ERA5 reference. We evaluated 17 model configurations across multiple model families on the July temperature task (Table 2). Our evaluation reveals that current LLMs contain a nontrivial representation of Earth's climate, though with significant quantitative limitations. Model predictions capture the broad expected structure of Earth's climate: warmer tropics

TABLE 1. Open-weight models: local image identifiers and quantization.

Model	Local image (Ollama)	Parameters	Quantization
GPT-OSS 120B	gpt-oss:120b (f7f8e2f8f4e0)	116.8 B	MXFP4
GPT-OSS 20B	gpt-oss:20b (aa4295ac10c3)	20.9 B	MXFP4
Mistral-small-3.1 24B	mistral-small3.1:24b (b9aaf0c2586a)	24.0 B	Q4_K_M
Gemma 27B	gemma3:27b (a418f5838eaf)	27.4 B	Q4_K_M
Gemma 12B	gemma3:12b (f4031aab637d)	12.2 B	Q4_K_M
Llama3.1 8B	llama3.1:8b (46e0c10c039e)	8.0 B	Q4_K_M
Mistral-nemo 12B	mistral-nemo:12b (e7e06d107c6c)	12.2 B	Q4_0
Llama2 70B	llama2:70b (e7f6c06ffef4)	69.0 B	Q4_0
eci-io-climategpt 70B	eci-io-climategpt-70b (jjansen)	69.0 B	Q4_0

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

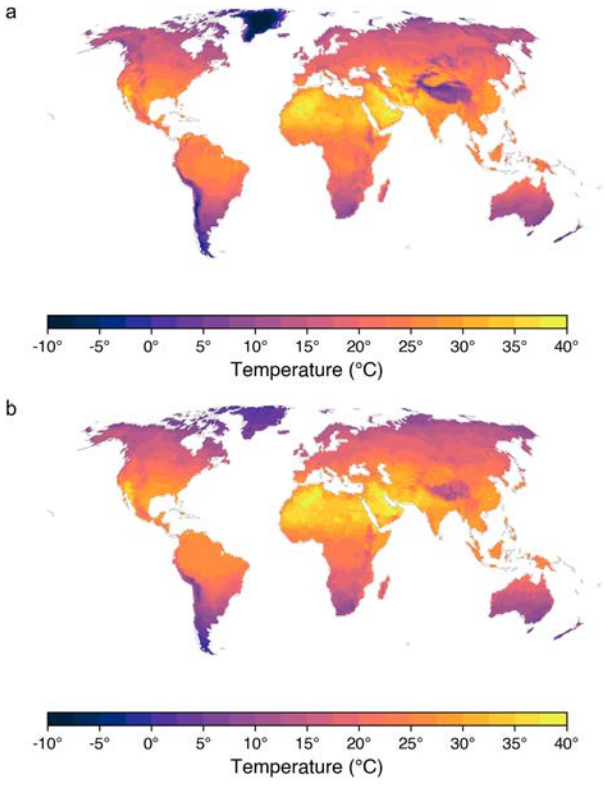


FIG. 2. Mean Jul 2-m air temperature (1991–2020). (a) ERA5 reanalysis. (b) LLM GPT-5 predictions. Both panels are shown on a 1° grid.

TABLE 2. RMSE and mean bias for LLM temperature predictions. Models are sorted by minimum RMSE. All experiments used a 1.0° global land grid with July mean temperature (1991–2020). "No address" indicates experiments where country/region/city information was withheld. The temp = 0.3 denotes the sampling temperature, a parameter controlling the randomness of LLM outputs, lower values produce more deterministic predictions.

Model/configuration	RMSE/bias (°C)	RMSE/bias (°C), no address
GPT-5, temp = 0.3	3.5/0.9	—
GPT-5-mini, temp = 0.3	3.7/−0.4	5.6/−0.8
GPT-5-mini	3.7/−0.4	—
GPT-OSS 120B, temp = 0.3	3.7/−0.6	5.3/−0.6
GPT-OSS 120B	4.0/−0.4	—
Mistral-small-3.1 24B, temp = 0.3	4.1/1.0	—
Mistral-small-3.1 24B	4.1/1.0	—
Gemma 27B, temp = 0.3	4.5/1.3	7.1/2.2
Gemma 27B	4.6/1.5	—
GPT-OSS 20B, temp = 0.3	5.1/−1.3	6.5/−1.3
Gemma 12B, temp = 0.3	5.2/1.1	—
Gemma 12B	5.2/1.2	6.3/0.0
GPT-5-nano, temp = 0.3	5.3/−0.8	—
GPT-5-nano	5.2/−0.8	6.9/−0.7
GPT-OSS 20B	5.5/−1.4	—
Gemini-2.5-flash-lite, temp = 0.3	5.5/−1.2	5.6/−1.2
Gemini-2.5-flash-lite	5.8/−1.2	5.8/−1.2
Llama3.1 8B, temp = 0.3	5.9/1.8	5.9/1.8
Llama3.1 8B	6.0/1.6	5.9/1.6
Mistral-nemo 12B, temp = 0.3	6.9/0.4	6.9/0.4
Mistral-nemo 12B	7.5/0.5	7.5/0.5
Llama2 70B	8.0/3.8	7.9/4.8
eci-io-climategpt 70B	10.2/3.9	—

and hot subtropical deserts, cooler high latitudes and mountain regions (Fig. 2). Visual comparison between ERA5 observations (Fig. 2a) and a representative model output (Fig. 2b, GPT-5 with sampling temperature 0.3) reveals qualitative agreement in large-scale patterns. However, quantitative metrics reveal substantial deviations. Global RMSEs ranged from 3.5°C (GPT-5, temp = 0.3) to 10.2°C (eci-io-climategpt 70b), with a median near 5°C across models (Table 2); most models fell within 3°–6°C. Mean biases were on the order of ±1°C, with spatially coherent error structures.

Interestingly, these error magnitudes are comparable to biases in contemporary physical climate models when evaluated against reanalysis at similar spatial scales (Eyring et al. 2016), though we stress that this comparison is purely contextual—LLMs lack physical understanding, and their errors arise from fundamentally different mechanisms.

The best-performing LLM showed mean biases ranging from −0.6° to +0.9°C, while the full ensemble spanned from −1.4° to +3.9°C. Importantly, individual models maintained consistent bias sign across different configurations, whether changing sampling parameters or adding or removing location information (country, city, region). This indicates that the systematic tendency to underpredict or overpredict temperatures is an intrinsic characteristic of each model rather than an artifact of specific prompt variations.

A regional close-up over Europe, North Africa, and the Caspian region (Fig. 3) illustrates the spatial structure of these deviations at the gridcell level. GPT-5 reproduced the large-scale temperature distribution of ERA5 (Figs. 3a,b), including the northwest–southeast gradient from the cool Atlantic seaboard and Scandinavia toward the hot Sahara and Middle East. The difference map (Fig. 3c), however, revealed a systematic warm bias over mountainous terrain (the Alps, the Atlas, the Anatolian Plateau, the Caucasus, and the Zagros), reaching up to +14.5°C, where the model does not resolve elevation-driven cooling. In contrast, biases remained small, mostly within ±1°C, over the low-lying and densely populated parts of western and central Europe, such as the British Isles, the North European Plain, and the major river valleys. Errors were also visibly larger and noisier across the sparsely populated Sahara than across Europe, consistent with the place-name anchoring discussed below: Densely populated regions are richly described in training corpora, whereas remote desert and high-mountain locations are represented far more sparsely.

Larger models within the same family generally exhibited lower errors. For instance, within the GPT-5 family, GPT-5 (RMSE 3.5°C) outperformed GPT-5-mini (3.7°C) and GPT-5-nano (5.2°C). Similarly, Gemma 27B (4.5°C) surpassed Gemma 12B (5.2°C). Notably, the direction of the bias (warm or cold) was shared by all size variants within a family: Every GPT-5 variant was cold-biased and every Gemma variant was warm-biased.

Among the best-performing configurations, GPT-5 (temp = 0.3), GPT-5-mini (temp = 0.3), GPT-OSS 120B (temp = 0.3),

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

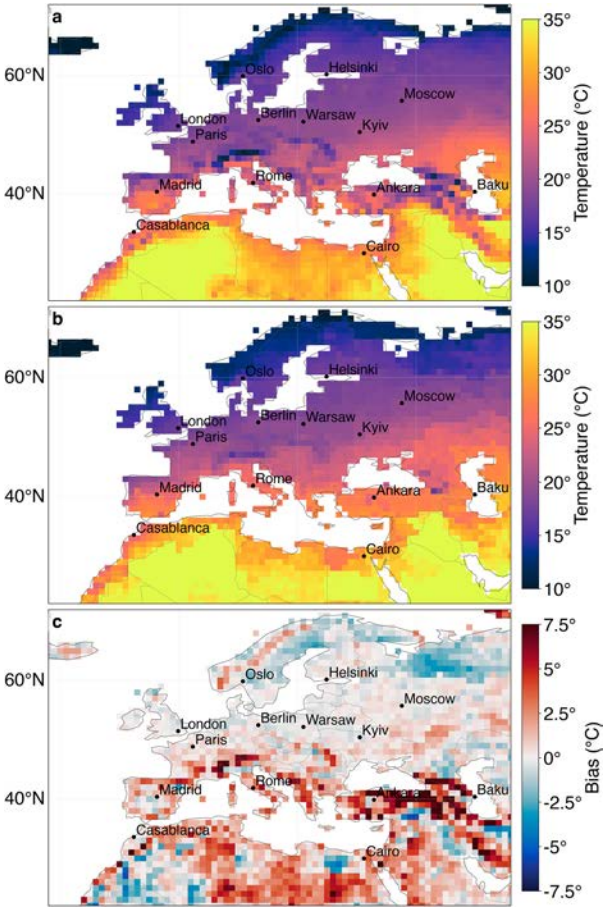


FIG. 3. Mean Jul 2-m air temperature (1991–2020). (a) ERA5 reanalysis. (b) LLM GPT-5 predictions. (c) Bias (GPT-5 minus ERA5). All panels are shown on a 1° grid.

and Mistral-small-3.1 24B (temp = 0.3) all achieved RMSEs below 4.1°C. The specialized climate model eci-io-climategpt 70B showed the poorest performance (RMSE 10.2°C), suggesting that domain-specific fine-tuning on textual climate content does not necessarily improve numeric climate recall.

Notably, this specialized model eci-io-climategpt 70B performed worse than the general-purpose Llama2 70B from which it was derived (RMSE 10.2° vs 8.0°C): Continued pretraining and instruction tuning on curated climate text degraded rather than improved closed-book numeric recall, a specialization-generalization trade-off (sometimes termed catastrophic forgetting) in which adapting the base weights toward qualitative climate discourse appears to erode their stored numeric associations. The model is moreover intended for text-based climate question answering with retrieval augmentation rather than unaided numeric recall, and its warm bias is already present in the base model (+3.8°C), only slightly amplified after fine-tuning (+3.9°C), consistent with the overrepresentation of warming in climate-change text.

To test whether the RMSE differences are significant, we used a paired bootstrap over 10° × 10° spatial blocks, which accounts for the spatial autocorrelation of the errors; a naive point test on roughly 15 000 land points reports almost any difference as significant. Gaps of about 0.3°C or less are not resolved at the 95% level. Within a family, the leading sizes are statistically indistinguishable (GPT-5 vs GPT-5-mini, 0.19°C, 95% CI [−0.44, +0.05]), and the four best models lie within 0.6°C of one another and cannot be separated (e.g., GPT-5-mini vs GPT-OSS 120B, 0.01°C, CI [−0.21, +0.25]). Larger steps in size are significant (GPT-5-mini vs GPT-5-nano, 1.61°C; GPT-OSS 120B vs GPT-OSS 20B, 1.44°C).

Skill does not follow size alone: The newer Llama3.1 8B significantly outperforms the larger Llama2 70B by about 2°C, indicating that model generation might matter more than parameter count. These intrafamily differences are also small relative to the roughly 2°C interannual variability of the target.

Bias maps reveal spatially coherent error patterns (Fig. 4). GPT-5 exhibited cold biases in central Asia and parts of South America, with notable warm biases in Greenland and parts of the Sahara (Fig. 4a). GPT-OSS 120B showed similar regional structures but with opposite signs in some areas (Fig. 4b). Mistral-small-3.1 24B and Gemma 27B displayed warm biases across much of the Northern Hemisphere midlatitudes and cold biases in parts of the Southern Hemisphere (Figs. 4c,d). High-elevation regions such as the Himalayas, Tibetan Plateau, and Greenland exhibited particularly large errors across all models, indicating difficulty encoding temperature–elevation relationships, partly due to less frequently available information ("remote regions"). These mountainous areas are also known to show large biases in contemporary physical climate models (Eyring et al. 2016), suggesting a common challenge in representing complex topography at relatively coarse resolution. We also note that ERA5, although the standard global reference and the gridded product closest to observations, carries larger uncertainty over complex terrain and in data-sparse high latitudes (Delhasse et al. 2020; Muñoz-Sabater et al. 2021); part of the elevation-dependent discrepancy may therefore reflect uncertainty in the reference rather than LLM error alone. At high elevations, the models are warm-biased, and because reanalyses tend to smooth the highest terrain and represent mountain surfaces as lower and warmer than they are, ERA5 may itself share part of this warm tendency; the true warm bias of the LLMs at these points may therefore be larger than the LLM-minus-ERA5 difference reported here.

In fact, our analysis shows that most biases appear in areas with altitudes exceeding 1500 m for nearly all models. To quantify this altitude dependence, we divided results into nine elevation bins (Table 3). At low elevations (0–500 m, 60.5% of land points), RMSE ranged from 1.9° to 3.1°C across the four best-performing models, with biases near zero for GPT-5 (−0.14°C) and Mistral-small-3.1 24B (−0.02°C), but reaching −1.4°C for GPT-OSS 120B and +0.23°C for Gemma 27B. Performance degrades systematically with altitude: RMSE increased to 3.1°–4.2°C at 1000–1500 m and then rose sharply to 5.2°–6.2°C at 1500–2000 m. Above 2500 m (4.3% of points), errors became severe (RMSE > 7°C, biases +5° to +13°C). Notably, 90% of locations lie below 1500 m where performance is markedly better, indicating that altitude-stratified metrics are more informative than global averages.

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

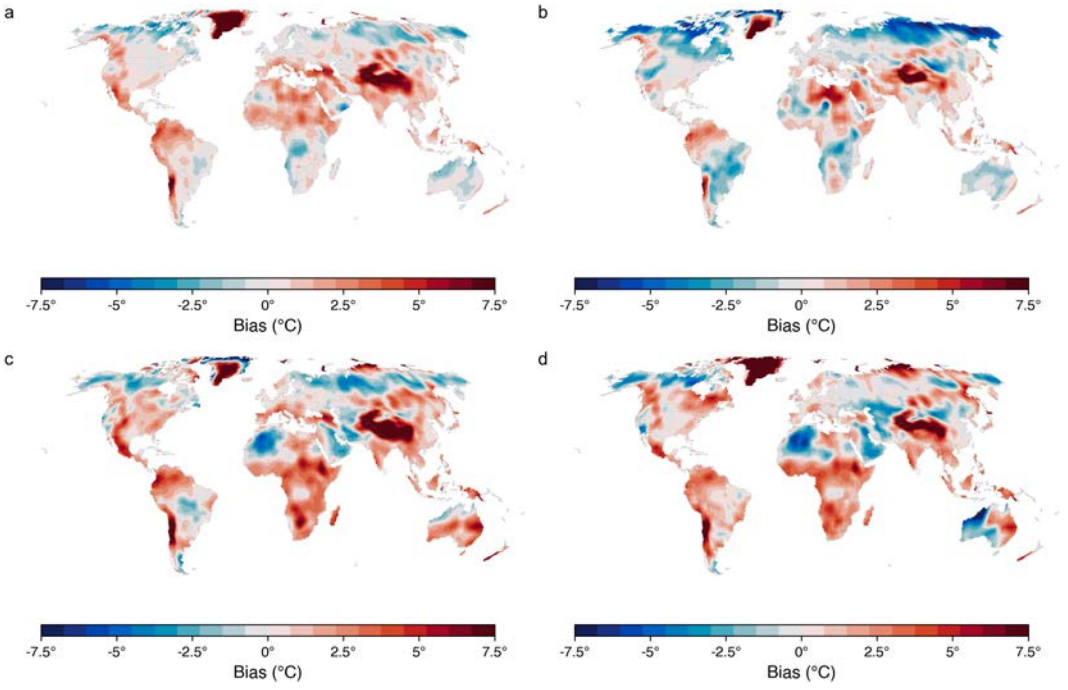


FIG. 4. Bias in mean Jul 2-m air temperature (1991–2020), defined as LLM predictions minus ERA5 reanalysis. (a) GPT-5, (b) GPT-OSS 120B, (c) Mistral-small-3.1 24B, and (d) Gemma 27B. All simulations were run with sampling temperature set to 0.3. Bias was calculated using a neighborhood (radius = 2) scheme (see methods).

Scatterplots of predicted versus observed temperatures (Fig. 5) show strong overall correlation but also systematic deviations. GPT-OSS 120B (Fig. 5b) achieved near-unity slope (1.012) with relatively tight clustering around the 1:1 line, whereas GPT-5 (Fig. 5a), Mistral-small-3.1 24B (Fig. 5c), and Gemma 27B (Fig. 5d) exhibited greater scatter. At the coldest temperatures (below 10°C), all models tended to underpredict the magnitude of cold, consistent with their difficulty in high-latitude and high-elevation regions.

The sampling temperature is a model setting that controls how random the output is: A low value makes the model favor its single most likely answer and behave almost deterministically, whereas a higher value produces more varied responses. We tested each model at temperature = 0, at which it is deterministic and returns its single most likely answer, and at a slightly higher temperature of 0.3, which introduces a small amount of variation. Across most models, using a temperature of 0.3 yielded marginally improved performance. For example, GPT-OSS 120B improved from RMSE 4.0°C (temp = 0) to 3.7°C (temp = 0.3), and GPT-5-nano showed a slight degradation from 5.2° to 5.3°C. These changes are small, and their sign is not consistent across models, indicating that the sampling temperature has only a minor effect on accuracy.

Providing additional geographic descriptors (country, state/region, city) alongside coordinates significantly improved model accuracy. For models where we tested both conditions, RMSE increased substantially when address information was omitted (Table 2, "no address" column). Across all 13 model

TABLE 3. Model performance across elevation bins. RMSE and mean bias for July mean temperature predictions (1991–2020) at different altitudes. Percentages indicate the fraction of land grid points within each altitude range. All models were run with a sampling temperature 0.3.

Model|Altitude	GPT-5 RMSE|Bias		oss120b RMSE|Bias		Mistral-small-3.1 24B RMSE|Bias		Gemma 27B RMSE|Bias	
0–500 m, 60.5%	1.9	−0.1	3.1	−1.4	3.0	0.0	3.1	0.2
500–1000 m, 19.2%	2.3	0.6	3.3	−0.3	3.4	0.8	3.3	1.1
1000–1500 m, 10.3%	3.1	1.5	3.4	0.2	3.9	1.9	4.2	2.0
1500–2000 m, 3.9%	5.2	3.8	4.3	0.9	4.8	3.1	6.2	4.1
2000–2500 m, 2.3%	8.5	7.3	5.6	3.0	7.0	5.8	10.3	8.8
2500–3000 m, 1.4%	11.7	10.6	7.4	5.1	8.6	7.7	14.1	12.8
3000–3500 m, 0.6%	12.4	10.9	9.5	7.0	11.7	10.4	14.0	12.3
3500–4000 m, 0.5%	11.2	8.6	10.3	5.4	12.7	11.0	11.5	9.7
4000 m+, 1.5%	11.3	9.5	9.5	5.7	13.9	12.8	11.6	9.2

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

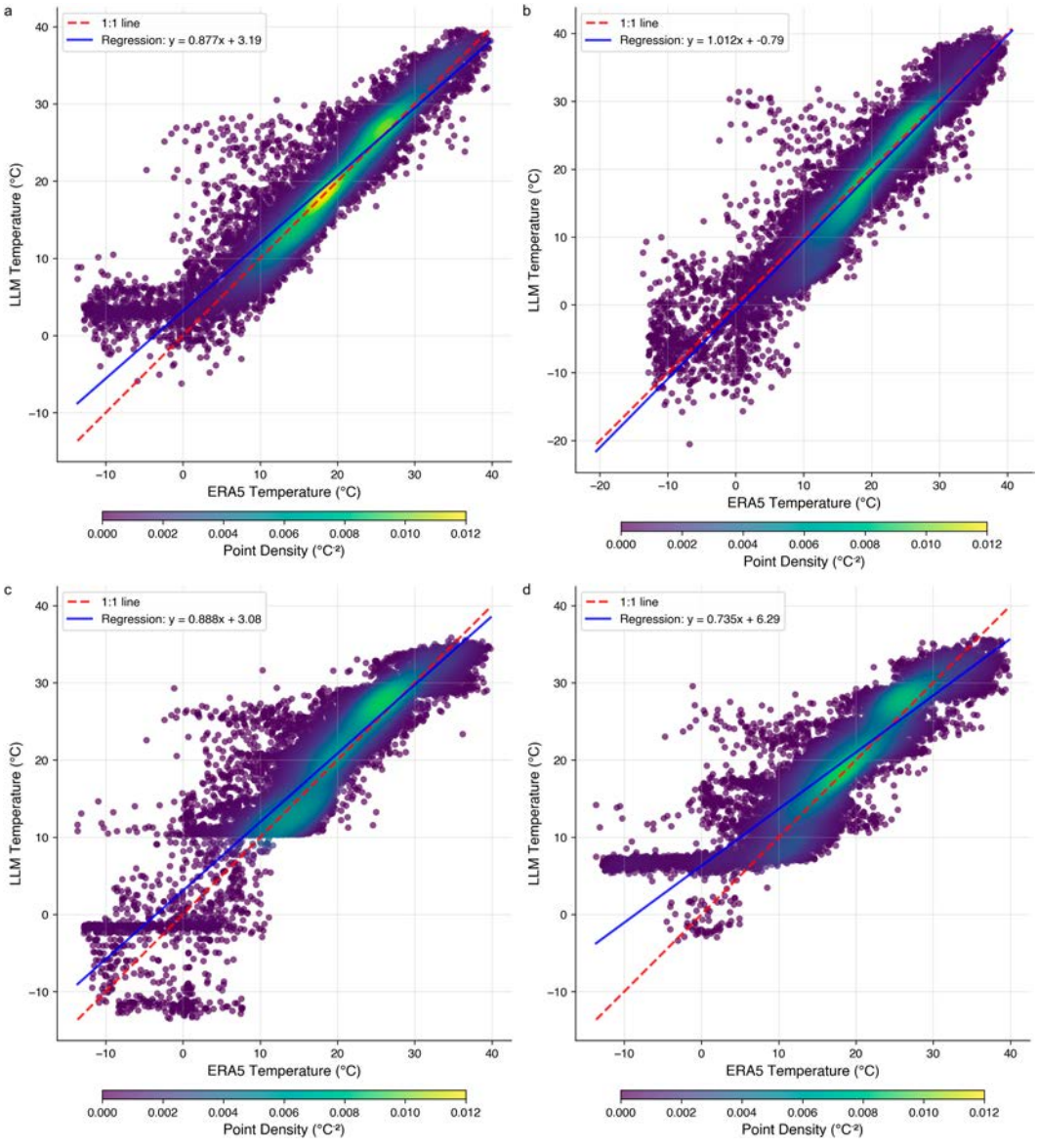


FIG. 5. Scatterplots of LLM-predicted vs ERA5 2-m air temperature (1991–2020). (a) GPT-5, (b) GPT-OSS 120B, (c) Mistral-small-3.1 24B, and (d) Gemma 27B. Point density was estimated using a Gaussian kernel density estimator with the default Scott bandwidth (Scott 2015).

configurations tested with and without address information, RMSE increased by an average of 27% when geographic descriptors were omitted. For instance, GPT-5-nano RMSE rose from 5.2°C (with address) to 6.9°C (without address), representing a 32% increase in error, GPT-5-mini from 3.7° to 5.6°C (+51%), and Gemma 27B from 4.5° to 7.1°C (+57%).

This result indicates that LLMs leverage recognizable place names to anchor their temperature predictions, likely because training corpora associate locations with climatic descriptions more strongly than raw coordinates. The observed skill should therefore be interpreted as closed-book parametric recall of climatological values and broad spatial regularities, partly aided by place-name associations when such descriptors are provided, rather than as evidence of physical understanding. Critically, mean bias remained consistent in sign and magnitude regardless of address provision, reinforcing that omitting context increases random error rather than introducing systematic bias shifts.

b. Seasonal extension and sampling sensitivity

Extending the evaluation to all 12 months preserved the July model ranking and revealed a coherent seasonal error structure (Fig. 6). Pooled over 18 480 point–month pairs, RMSE was 3.73°C for GPT-5, 4.00°C for GPT-OSS 120B, 4.94°C for Mistral-small-3.1 24B, and 6.10°C for Gemma 27B. GPT-5 was the most nearly season-independent, with monthly

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

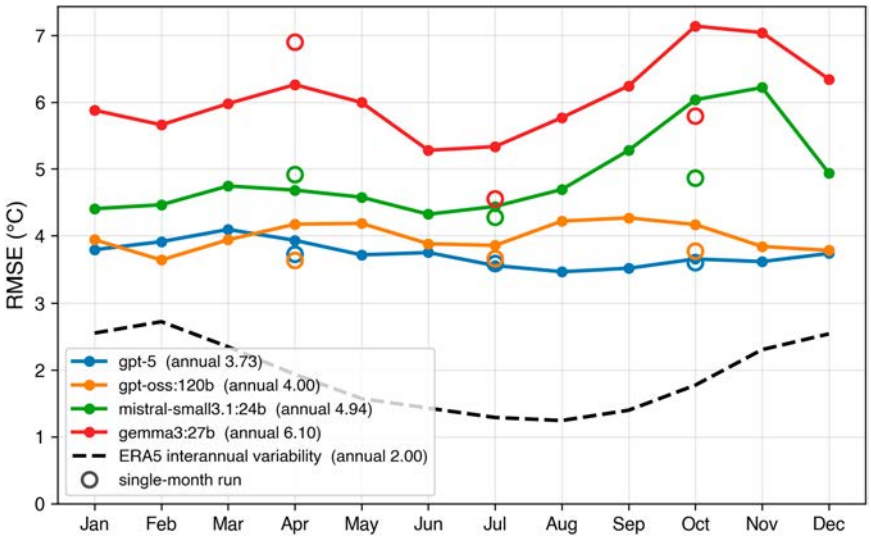


FIG. 6. Monthly RMSE of closed-book LLM predictions of mean 2-m air temperature (1991–2020) against ERA5, for GPT-5, GPT-OSS 120B, Mistral-small-3.1 24B, and Gemma 27B at 1540 land points (every 10th point of the 1° land grid). Colored lines show model RMSE for each calendar month; legend values give the RMSE pooled over all 18 480 point–month pairs. The dashed black line shows the ERA5 interannual variability at the same points (the typical year-to-year spread of each month about its 1991–2020 mean). Open circles show the corresponding single-month runs. All LLM runs used the API-default sampling setting.

RMSE ranging from 3.46° to 4.09°C and an annual bias of −0.33°C. The other three models were warm biased in the annual mean (from +0.86° to +1.77°C) and showed larger errors in boreal autumn; Mistral-small-3.1 24B reached 6.22°C in November and Gemma 27B reached 7.14°C in October. ERA5 interannual variability at the same points averaged 2.00°C over the year, larger in boreal winter and smaller in boreal summer, so the monthly LLM errors are typically about 2–3 times the observed year-to-year variability of monthly temperature (Fig. 6).

Sampling-temperature sensitivity was small relative to intermodel differences (Fig. 7). On the 1540-point July subset, RMSE changed by at most 0.45°C and bias by at most 0.18°C among the temperature = 0, temperature = 0.3, and API-default settings; the largest change occurred for GPT-OSS 120B, whose RMSE decreased from 4.11°C at temperature = 0 to 3.66°C at the API default, while the other three models varied by 0.15°C or less. The subset also reproduced the full-grid July values at temperature = 0.3, with GPT-OSS 120B at 3.68°C on the subset versus 3.67°C on the full grid and Gemma 27B at 4.48° versus 4.49°C. These results indicate that the monthly experiment is not materially controlled by the sampling-temperature setting or by the 10% grid subset.

c. *Climate change*

We extended the evaluation to assess whether LLMs can estimate observed climate change by querying the temperature difference between two historical periods: 1950–74 and 2000–24. Results were poor across all tested models. Smaller models (e.g., Mistral-small-3.1 24B) produced nearly uniform warming of approximately 0.8°C everywhere with minimal spatial variability, while GPT-5 predicted a global mean warming of 1.0°C (Fig. 8). These global mean values are reasonably close to the ERA5-derived warming of 1.1°C for land points over this period, indicating that LLMs capture the overall magnitude of observed warming. However, models failed to reproduce the spatial distribution of temperature change: Larger models such as GPT-5 exhibited some spatial variability in predicted changes but showed near-zero correlation with ERA5-derived spatial patterns of temperature trends over these periods (Fig. 8).

This failure to capture regional patterns can be explained by the fact that point-specific temperature trends are rarely discussed in accessible literature outside scientific publications, and hence LLMs have limited exposure to such granular temporal data during training. The results also suggest that LLMs cannot generalize information to construct implicit time series for individual locations from the broader corpus of climate information. They do not automatically extract temporal patterns from scattered mentions of a location across different time periods and synthesize them into coherent trends. This contrasts sharply with their success at recalling climatological baselines, where both global statistics and spatial patterns (e.g., latitudinal gradients, topographic effects) are reasonably well reproduced.

Our results show that while LLMs capture the global climate change signal, they fail to represent its local and regional distribution, underscoring the need for external data when temporal or spatial details of climate change dynamics are required.

4. Discussion

Our closed-book evaluation reveals that contemporary LLMs possess nontrivial internal representations of Earth's mean climate. In qualitative terms, model-generated temperature maps

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

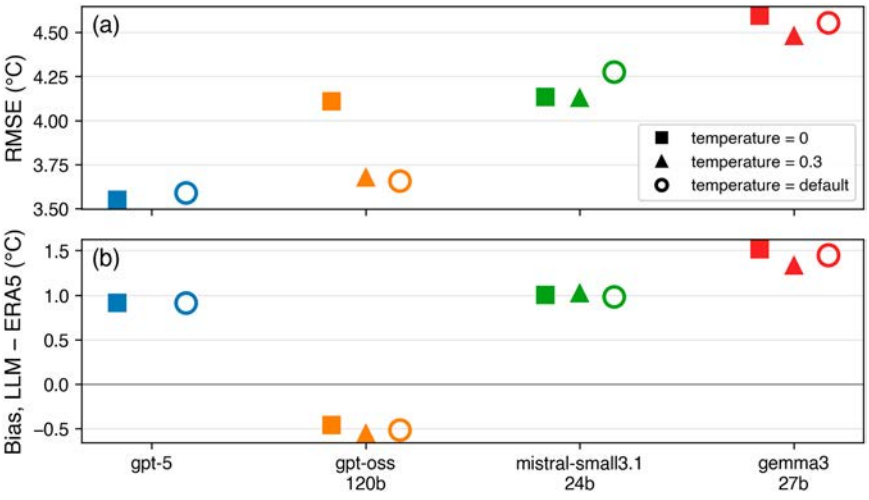


FIG. 7. Sensitivity of July temperature skill to the sampling-temperature setting at the 1540-point subset. (a) RMSE and (b) mean bias (LLM minus ERA5) for July mean 2-m air temperature (1991–2020). Markers denote the sampling setting, where available: filled square, temperature = 0; filled triangle, temperature = 0.3; open circle, API default (sampling-temperature parameter left unset). Squares and triangles are the full-grid July runs restricted to the subset, and open circles are independent single-month July runs at the API default.

capture broad expected patterns: warmer tropics and hot subtropics, cooler high latitudes and mountains.

Computed over all 1540 land points and 12 months, the predicted and ERA5 monthly climatologies correlate at 0.95–0.98.

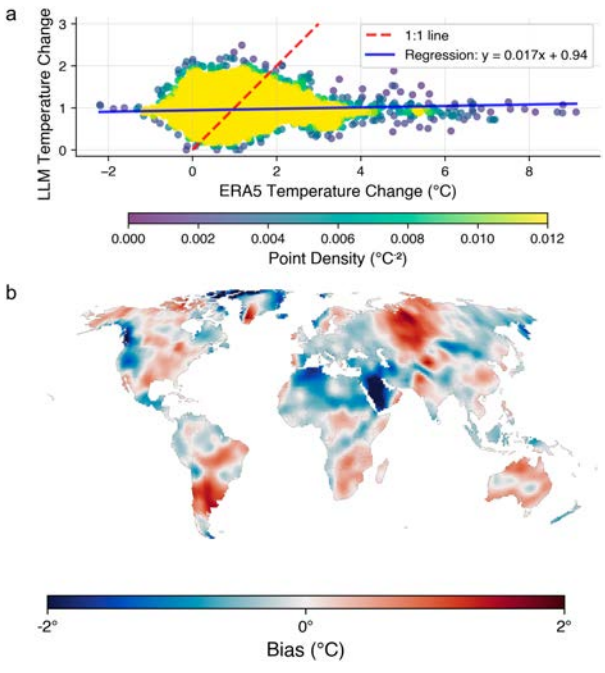


FIG. 8. GPT-5 predictions of temperature change between 1950–74 and 2000–24. (a) Scatterplot of LLM-predicted vs ERA5-derived change in Jul mean 2-m air temperature. (b) Spatial bias map (LLM minus ERA5). Simulation was run with sampling temperature 0.3. Bias calculated using a neighborhood (radius = 2) scheme (see methods).

This correlation is high mainly because the models follow the seasonal cycle closely at each location and place temperatures correctly across the globe; as the evaluation spans the entire land surface, it further shows that the models resolve regionally specific behavior rather than a single global seasonal cycle, reproducing the reversal of the seasons between the Northern and Southern Hemispheres. Because correlation reflects agreement of pattern and not of magnitude, it coexists with the RMSE and bias of several degrees reported above.

Quantitatively, however, there are significant deviations from true values, with substantial errors (RMSEs of 3°–6°C) and systematic regional biases that limit the reliability of unaided LLM climate recall for practical applications. The elevation-dependent analysis reveals that this limitation is concentrated: 90% of land points lie below 1500 m where errors remain moderate (RMSE 2°–4°C), while the remaining high-elevation locations drive much of the global error statistics.

Because a domain-mean bias can mask compensating regional errors of opposite sign, the regionally resolved biases (e.g., over plateaus, deserts, and coastal zones) are more informative than the global mean. A systematic analysis of these regional biases, and of how they depend on the training data and its geographic distribution across models, would be a valuable extension that lies outside the scope of the present study.

The finding that geographic context significantly enhances accuracy has practical implications. When applications or systems query LLMs for climate information using coordinates, providing additional place names (country, region, city) substantially improves response accuracy. This suggests that climate-related applications integrating LLMs should include rich geographic descriptors alongside any coordinate data to maximize accuracy. The failure of LLMs to capture temporal changes, despite their partial success with climatological means, highlights a fundamental limitation: Models encode static snapshots of

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

knowledge but struggle with dynamic trends absent explicit training signals.

The proliferation of LLM benchmarks across domains underscores the need for task-specific evaluation (Liang et al. 2023; Chiang et al. 2024; Zheng et al. 2023; Srivastava et al. 2022). While general-purpose benchmarks measure broad capabilities, domain benchmarks like ours provide targeted assessment of scientific knowledge recall. Repeating the evaluation for all 12 months on the nested subset did not materially change the picture obtained from July alone. The relative ordering of the models by RMSE and the sign of their biases were preserved across the annual cycle, indicating that a single well-chosen month already captures the relative performance of the models. The models with fewer parameters tended to degrade more when the query became more demanding; for example, returning the full seasonal cycle in a single response inflated the error of Gemma 27B most strongly, which suggests that smaller models are more easily confused as the task grows more complex, although model-specific factors cannot be excluded. The same closed-book approach can be applied directly to other key climate variables, such as precipitation or wind. Tracking performance on this benchmark as models evolve can quantify progress in parametric climate knowledge.

We emphasize that accurate numeric recall does not imply physical understanding. LLMs remain statistical pattern matchers without causal models of atmospheric processes. Unlike physically based models, or even simple empirical predictors that compute temperature from latitude, elevation, and season, they do not derive a value from physical laws or geographic predictors but reproduce it from statistical associations learned from text. Their errors, while similar in magnitude to some physical model biases (Eyring et al. 2016), arise from different mechanisms: incomplete or noisy training data rather than approximations to governing equations. The direction and spatial structure of these residual biases, including the warm errors over Greenland and the Himalaya, cannot be established from our experiments alone; one may speculate that locations with denser place-tagged numeric reporting are better represented in the training corpora, or that the prominence of warming in climate-change text influences predictions for frequently discussed cold regions, but testing such hypotheses would require analysis of the models' training data, which lies beyond the scope of the present study. This distinction matters when assessing reliability. For mission-critical climate applications, tool-augmented systems that ground LLM reasoning in authoritative data sources (Vaghefi et al. 2023; Koldunov and Jung 2024; Kuznetsov et al. 2025) are essential.

For LLM research, our results position closed-book climate recall as a quantitative, objectively verifiable probe of parametric knowledge: Unlike question-answering accuracy, it yields a continuous error with interpretable spatial structure, and it cleanly separates two abilities. Contemporary LLMs recall static facts that are abundantly stated in text (the present-day climatology of a place, including its latitudinal and topographic structure), yet they do not synthesize implicit temporal trends that are rarely stated point by point, such as local rates of change. The strong dependence on place names (a 27% average reduction in error when country, region, and city are supplied) indicates that much of this skill is lexical association rather than a learned physical or geographic model, and the failure of a domain-adapted model (eci-io-climategpt 70B), which underperforms its own base model, shows that fine-tuning on climate text does not by itself improve numeric recall. The framework provides a reproducible benchmark for tracking these abilities as models evolve.

For climate applications, the message for practitioners is twofold. First, unaided closed-book values cannot be trusted at the local scale: Errors of several degrees, growing sharply over high terrain and reaching 2–3 times the interannual variability of the target, exceed what most quantitative applications tolerate. Second, and more consequentially, LLMs reproduce the global-mean magnitude of recent warming but not its regional expression, which is precisely the spatial detail that matters for impacts and adaptation. Both limitations point to the same conclusion: Climate-facing applications should ground LLM reasoning in authoritative data through retrieval or tools, and should supply rich geographic context, rather than rely on parametric recall alone. This matters in practice because an agentic system may be built without such tools, or may decline to invoke them, and silently substitute a parametric estimate, and our results quantify the error that such substitution incurs. This is also why systems that query authoritative databases directly and carry out the computation through dedicated tools rather than within the language model, such as those of Kuznetsov et al. (2025) and Pantiukhin et al. (2026), are important for reliable local climate information.

Acknowledgments. This work was supported by the European Union's Destination Earth Initiative and relates to tasks entrusted by the European Union to the European Centre for Medium-Range Weather Forecasts implementing part of this Initiative with funding by the European Union. This work is also supported by projects S1: Diagnosis and Metrics in Climate Models of the Collaborative Research Centre TRR 181 "Energy Transfer in Atmosphere and Ocean," funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project 274762653). It is also supported by the EERIE project (Grant Agreement 101081383), funded by the European Union. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Climate Infrastructure and Environment Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Data availability statement. Code is available at https://github.com/CliDyn/geo_benchmark. ERA5 monthly averaged data on single levels are available from the Copernicus Climate Data Store (Hersbach et al. 2023) (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means) (accessed 7 July 2025). Contains modified Copernicus Climate Change Service information 2025. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

REFERENCES

Bulian, J., and Coauthors, 2024: Assessing large language models on climate information. *ICML'24: Proceedings of the 41st*

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC

International Conference on Machine Learning, JMLR.org, 4884–4935, https://dl.acm.org/doi/10.5555/3692070.3692264.

Chang, H., J. Park, S. Ye, S. Yang, Y. Seo, D. Chang, and M. Seo, 2024: How do large language models acquire factual knowledge during pretraining? *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 60 626–60 668, https://dl.acm.org/doi/10.5555/3737916.3739855.

Chiang, W.-L., and Coauthors, 2024: Chatbot arena: An open platform for evaluating LLMs by human preference. *ICML'24: Proceedings of the 41st International Conference on Machine Learning*, JMLR.org, 8359–8388, https://dl.acm.org/doi/abs/10.5555/3692070.3692401.

Delhasse, A., C. Kittel, C. Amory, S. Hofer, D. van As, R. S. Fausto, and X. Fettweis, 2020: Brief communication: Evaluation of the near-surface climate in ERA5 over the Greenland Ice Sheet. *Cryosphere*, **14**, 957–965, https://doi.org/10.5194/tc-14-957-2020.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Feng, K., X. Shen, W. Wang, X. Zhuang, Y. Tang, Q. Zhang, and K. Ding, 2024: SciKnowEval: Evaluating multi-level scientific knowledge of large language models. arXiv, 2406.09098v4, https://doi.org/10.48550/arXiv.2406.09098.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Hersbach, H., and Coauthors, 2023: ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), accessed 11 March 2025, https://doi.org/10.24381/cds.f17050d7.

Kandpal, N., H. Deng, A. Roberts, E. Wallace, and C. Raffel, 2023: Large language models struggle to learn long-tail knowledge. *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, JMLR.org, 15 696–15 707, https://dl.acm.org/doi/10.5555/3618408.3619049.

Koldunov, N., and T. Jung, 2024: Local climate services for all, courtesy of large language models. *Commun. Earth Environ.*, **5**, 13, https://doi.org/10.1038/s43247-023-01199-1.

Kuznetsov, I., A. A. Jost, D. Pantiukhin, B. Shapkin, T. Jung, and N. Koldunov, 2025: Transforming climate services with LLMs and multi-source data integration. *npj Climate Action*, **4**, 97, https://doi.org/10.1038/s44168-025-00300-y.

Liang, P., and Coauthors, 2023: Holistic Evaluation of Language Models. arXiv, 2211.09110v2, https://doi.org/10.48550/arXiv.2211.09110.

Min, S., and Coauthors, 2023: FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. *Proc. 2023 Conf. on Empirical Methods in Natural Language Processing*, Singapore, Association for Computational Linguistics, 12 076–12 100, https://doi.org/10.18653/v1/2023.emnlp-main.741.

Muñoz-Sabater, J., and Coauthors, 2021: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021.

Nominatim, 2025: Open-source geocoding with OpenStreetMap data. Accessed 11 March 2025, https://nominatim.org/.

Pantiukhin, D., B. Shapkin, I. Kuznetsov, T. Jung, and N. Koldunov, 2026: CMIP-Forge: An agentic system that retrieves, computes, and self-reviews climate science. arXiv, 2606.17076v1, https://doi.org/10.48550/arXiv.2606.17076.

Scott, D. W., 2015: *Multivariate Density Estimation: Theory, Practice, and Visualization.* 2nd ed. Wiley, 384 pp., https://doi.org/10.1002/9781118575574.

Srivastava, A., and Coauthors, 2022: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv, 2206.04615v3, https://doi.org/10.48550/arXiv.2206.04615.

Sun, L., Y. Han, Z. Zhao, D. Ma, Z. Shen, B. Chen, L. Chen, and K. Yu, 2023: SciEval: A multi-level Large Language Model evaluation benchmark for scientific research. arXiv, 2308.13149v2, https://doi.org/10.48550/arXiv.2308.13149.

Thulke, D., and Coauthors, 2024: ClimateGPT: Towards AI synthesizing interdisciplinary research on climate change. arXiv, 2401.09646v1, https://doi.org/10.48550/arXiv.2401.09646.

Tonmoy, S. M. T. I., S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, 2024: A comprehensive survey of hallucination mitigation techniques in large language models. arXiv, 2401.01313v3, https://doi.org/10.48550/arXiv.2401.01313.

Vaghefi, S. A., and Coauthors, 2023: ChatClimate: Grounding conversational AI in climate science. *Commun. Earth Environ.*, **4**, 480, https://doi.org/10.1038/s43247-023-01084-x.

Wang, Y., Y. Chen, W. Wen, Y. Sheng, L. Li, and D. Zeng, 2024: Unveiling factual recall behaviors of Large Language Models through Knowledge Neurons. *Proc. 2024 Conf. on Empirical Methods in Natural Language Processing*, Miami, FL, Association for Computational Linguistics, 7388–7402, https://doi.org/10.18653/v1/2024.emnlp-main.420.

Yuan, J., L. Pan, C.-W. Hang, J. Guo, J. Jiang, B. Min, P. Ng, and Z. Wang, 2024: Towards a holistic evaluation of LLMs on factual knowledge recall. arXiv, 2404.16164v1, https://doi.org/10.48550/arXiv.2404.16164.

Zhang, Y., and Coauthors, 2023: Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv, 2309.01219v3, https://doi.org/10.48550/arXiv.2309.01219.

Zheng, L., and Coauthors, 2023: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 46595–46623, https://dl.acm.org/doi/10.5555/3666122.3668142.

Unauthenticated | Downloaded 10/27/26 01:12 PM UTC