Empowering LLM Agents with Geospatial Awareness: Toward Grounded Reasoning for Wildfire Response

Anonymous ACL submission

Abstract

Effective disaster response is essential for safeguarding lives and property. Existing statistical approaches often lack semantic context, generalize poorly across events, and offer limited interpretability. While Large language models (LLMs) provide few-shot generalization, they remain text-bound and blind to geography. To bridge this gap, we introduce a **Geospatial Awareness Layer (GAL)** that grounds LLM agents in structured earth data. Starting from raw wildfire detections, GAL automatically retrieves and integrates infrastructure, demographic, terrain, and weather information from external geodatabases, assembling them into a concise, unit-annotated *perception script*. This enriched context enables agents to produce evidence-based resource-allocation recommendations (e.g., personnel assignments, budget allocations), further reinforced by historical analogs and daily change signals for incremental updates. We evaluate the framework in real wildfire scenarios across multiple LLM models, showing that geospatially grounded agents can outperform baselines. The proposed framework can generalize to other hazards such as floods and hurricanes.

1 Introduction

Natural hazards like wildfires, floods, earthquakes, and hurricanes evolve rapidly, imposing intense time and space constraints on the disaster decision-making process (Ma et al., 2024; Li et al., 2023). Wildfires in particular illustrate these challenges: they can spread quickly across vast areas, with propagation behavior influenced by complex topography and weather characteristics (Finney, 1998; Abatzoglou and Williams, 2016). Effective disaster response is critical for protecting lives, property, and enhancing community resilience under these extreme conditions (Cutter, 2016). However, emergency managers often face the difficulty of estimating personnel, equipment, and financial resources across diverse regions due to incomplete and uncertain information. While existing approaches in wildfire contexts, such as hotspot counts, fire radiative power (FRP) statistics, and time-series predictors, provide useful signals, they often ignore context and lack generalizability and interpretability (Giglio et al., 2016; Wooster et al., 2005). These limitations underscore an urgent need for effective disaster decision-support systems, especially in wildfire settings where operational decisions are made under both environmental uncertainty and rapidly shifting information conditions (Dunn et al., 2017; Thompson et al., 2019; Martell, 2015; Chen et al., 2026).

Large language models (LLMs) offer a promising direction, with strong few-shot generalization and the capacity to integrate diverse sources of knowledge (Brown et al., 2020; Touvron et al., 2023; Lei et al., 2025). More broadly, recent work on foundation models for environmental science has highlighted their potential to unify heterogeneous Earth data and support downstream scientific and decision-making tasks (Yu et al., 2025a). Despite this potential, current LLMs remain weak in geospatial grounding. They are prone to hallucinating in numerical operations and often struggle to reliably localize, interpret, or reason over the spatial context that is crucial for appraising hazard impacts and response needs (Huang et al., 2025; Li et al., 2025b). This disconnect between linguistic adaptability and real-world awareness represents a central limitation of current LLM agents. Inspired by recent advances in augmenting LLMs with external knowledge, together with the emerging need for retrieval mechanisms tailored to Earth and environmental sciences (Lewis et al., 2020; Borgeaud et al., 2022; Yu et al., 2025b), we pose an important question: can LLM agents draw on structured geospatial data from our physical world to support disaster response decisions?

Turning this idea requires bridging LLMs with

the physical world along several dimensions. First, the agent needs to actively retrieve relevant geospatial evidence at appropriate spatial and temporal scales. Second, the retrieved information has to be represented in a form that LLMs can interpret without being overwhelmed by raw geometries or large tabular data. Third, the system should generate stable and actionable outputs that remain consistent across days and events.

We address these challenges with a novel framework that equips LLM agents with a **Geospatial Awareness Layer (GAL)**, providing a structured interface to access data from the physical world. First, for retrieval, we build a PostGIS–raster database that (Obe and Hsu, 2021), given only hotspot coordinates and timestamps, automatically extracts infrastructure, demographic, terrain, and weather attributes using spatial joins and zonal statistics. Second, for representation, the layer encodes these heterogeneous signals into a compact, unit-annotated perception script with fixed fields, abstracting away raw geometries and large tables into a stable form that LLMs can readily reason over. Third, for decision stability, it enforces schema validation, unit normalization, and bounded ranges, while also reinforcing outputs with historical analogs and daily change signals. Our key contributions include:

- **Geospatially grounded framework.** We introduce a novel framework that enables LLM agents to see and interpret the physical world. This layer actively retrieves and interprets spatial evidence to ground decision-making in real-world conditions.
- **Reasoning analysis on real wildfires.** We evaluate grounded LLMs on multiple California wildfire events to analyze how spatial grounding affects reasoning behavior, finding improved interpretability and alignment with operational outcomes.
- **Open benchmark.** We conduct a cross-model evaluation of LLMs within this framework, establishing one of the first benchmarks for assessing geospatially grounded reasoning and decision accuracy in wildfire response.

2 Related Work

2.1 Traditional Wildfire Modeling

Traditional wildfire modeling spans physical simulation, statistical analysis, and remote sensing. Physical fire spread models, such as the Rothermel equation (Rothermel, 1972) and its implementation in FARSITE (Finney, 1998), provide mechanistic predictions of fire growth. Coupled fire–atmosphere systems like WRF-Fire (Coen et al., 2013) further integrate weather dynamics (e.g., wind, humidity) with combustion processes. Statistical and time-series approaches estimate wildfire occurrence or extent from historical patterns (Preisler and Westerling, 2007; Laube and Hamilton, 2021), and other data-driven methods predict burned area or assess fire impacts through epidemic-style models (Ma et al., 2024). Remote sensing techniques complement these efforts by detecting and forecasting wildfire activity using satellite observations (Chaparro et al., 2016; Huot et al., 2022). Despite their value for scientific understanding, these approaches are often slow and lack contextual awareness.

2.2 LLMs for Disaster Response

Compared with traditional techniques, LLMs offer unique advantages in processing multimodal inputs, analyzing unstructured text, and generating interpretable outputs to support decision-making in dynamic environments (Li et al., 2024, 2025b; Chen et al., 2024). Recent studies have explored their potential for disaster response (Chen et al., 2025; Otal et al., 2024; Goecks and Waytowich, 2023). For instance, pretrained LLMs have been used to extract location-based descriptors from social media, showing that incorporating geographic knowledge can improve performance (Hu et al., 2023). Other works integrate maps and geospatial data into LLM workflows to enhance disaster-relevant analysis (Hu et al., 2023; Yin et al., 2025; Zguir et al., 2025). Despite these advances, most LLMs operate in text space and cannot reliably perceive or reason about the geospatial conditions that govern real-world hazards (Huang et al., 2025; Li et al., 2025b).

LLM-based approaches to disaster prediction and simulation are particularly relevant to our focus (Tang et al., 2025; Li et al., 2025a; Chen et al., 2025; Zhong et al., 2024). For example, Li et al. (2025a) evaluate LLMs as world models to simulate human-perceived seismic risk using seismic data and Google Street views, revealing geospatial patterns consistent with "Did You Feel It?" maps (Quitoriano and Wald, 2020). Within wildfire studies, domain-specific models such as WildfireGPT (Xie et al., 2024) have been introduced for response tasks, and later evaluated for quantitative fire spread prediction (Ramesh et al., 2025). It has been observed that while general LLMs can de-

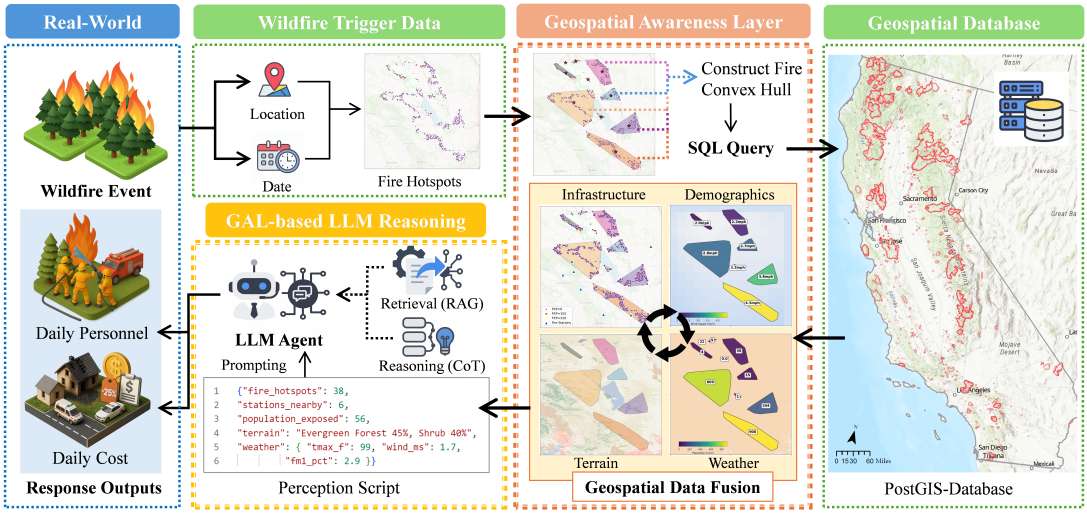


Figure 1: An illustration of the research framework.

scribe wildfire behavior qualitatively, they struggle with accurate quantitative forecasts (e.g., fire radiative power). LLMs have also been used for evacuation prediction (Dang et al., 2025; Chen et al., 2025); for instance, Chen et al. (2025) develop a framework that outperforms traditional theory-based models in modeling evacuation behaviors.

Our work differs from prior studies by introducing a GAL that dynamically grounds LLM agents in structured Earth data. GAL can perform inference-time retrieval of geospatial attributes from live databases and organizes them into a stable, unit-annotated representation. This dynamic grounding enables LLMs to produce geographically coherent and operationally actionable outputs across evolving disaster contexts.

3 Methodology

3.1 Framework Design

We propose to shift language models from closed linguistic systems toward agents capable of perceiving and reasoning about the physical world. Existing LLM-based tools remain confined to text, being powerful in text reasoning and summarization, yet blind to geography. To bridge this gap, we introduce a **Geospatial Awareness Layer (GAL)** that serves as an environmental sensor that provides active retrieval and integration of spatial evidence.

As illustrated in Figure 1, the proposed framework links real wildfire events to actionable response decisions through geospatially grounded reasoning. Each event is characterized by its detection date and hotspot locations, which trigger our GAL. GAL clusters hotspots into fire footprints, constructs convex hulls, and issues SQL queries to retrieve spatial attributes—such as infrastructure, demographics, terrain, and weather—from a PostGIS–raster database. These layers are fused into a compact, unit-annotated perception script that summarizes local exposure, accessibility, and environmental conditions. The script then drives the GAL-based LLM Reasoning module, where the model performs retrieval-augmented and chain-of-thought reasoning over current states and historical analogs to produce evidence-grounded recommendations for daily personnel and cost.

3.2 Data Preparation

We integrate multiple open geospatial datasets, including fire detection, infrastructure, demographics, terrain, and weather. For event initialization, we use satellite-derived wildfire detections from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensors onboard the Suomi National Polar-orbiting Partnership (S-NPP), NOAA-20, and NOAA-21 satellites. The VIIRS active fire data, with a spatial resolution of 375m, are obtained from NASA's Fire Information for Resource Management System (FIRMS).[1] Administrative and population attributes are derived from county boundaries and demographic statistics provided by the U.S. Census Bureau's American Community Survey (ACS) and the CA-POP dataset.[2] Critical facilities such as fire stations and infrastructure are obtained from the Homeland Infrastructure Foundation-Level Data (HIFLD) repository. Land surface information is provided by the National

[1] https://firms.modaps.eosdis.nasa.gov/usfs/active_fire/

[2] https://doi.org/10.5281/zenodo.5874927

Land Cover Database (NLCD)[3] and a digital elevation model (DEM) for terrain metrics.

Daily weather variables, including Burning Index (BI; a composite fire-danger metric used by agencies to gauge escalation risk), temperature, wind speed, and 1-hour fuel moisture (FM1; moisture content of fine fuels that governs ignition and spread rate), are sourced from the gridMET (METDATA) dataset (Abatzoglou, 2013). For evaluation, we further process daily situation reports from ICS-209-Plus (St. Denis et al., 2023) (standardized daily logs of personnel assignments, cumulative costs, and containment status filed by incident command), extracting daily personnel and cost records. These records are cleaned and aggregated into per-event time series for ground-truth validation.

The fundamental unit of analysis is a *fire–day*: one wildfire incident on one calendar date, paired with a perception script and a ground-truth target pair (daily personnel and cost from ICS-209-Plus). Dynamic inputs (VIIRS hotspots, gridMET weather, daily footprint) are synchronized to that date, while static layers (NLCD, ACS, HIFLD) are joined to the day's footprint via spatial overlays and zonal statistics. Fire–days with missing or inconsistent ICS-209-Plus records are excluded without imputation; absent geospatial fields are flagged `NA` rather than zero-filled, and short weather gaps are filled with a one-to-two day rolling window to avoid leaking future information.

3.3 Geospatial Awareness Layer (GAL)

GAL acts as a geospatial encoder that converts raw spatial signals into interpretable symbolic representations, allowing the LLM to reason over structured context rather than unbounded numeric inputs. By linking real-world coordinates to semantically meaningful fields, GAL provides the spatial grounding that conventional LLMs lack.

Event detection and normalization. GAL takes as input the detection date and hotspot coordinates with basic attributes. It clusters hotspots using DBSCAN (Ester et al., 1996) (radius = 3 km, MinPts ≥ 3). When fewer points are available, the module builds simple geometric buffers—a line for two points or a small circle for one—to preserve spatial continuity. For each cluster, FRP-weighted centroids are calculated, and polygonal footprints are generated using convex hulls or buffers to outline the fire's active area. The resulting standardized event footprints are stored in PostGIS with daily timestamps for subsequent spatial queries.

Geospatial information retrieval. Given the normalized fire footprints, GAL acts as an automatic retrieval interface that converts spatial context into semantically structured attributes usable by the LLM. It composes parameterized SQL queries over a PostGIS–raster store to gather relevant evidence while preserving spatial and temporal alignment. Each query retrieves a specific dimension of the wildfire environment: (i) **Fire hotspots & stations:** nearest-k fire stations (geodesic distance, $k = 3$), station density within radius, and intra-day hotspot counts/FRP per cluster. (ii) **Demographics (exposure):** population sum/density within each footprint and county identifiers for cross-jurisdiction reasoning. (iii) **Terrain (operational complexity):** NLCD land-cover composition per cluster, Shannon diversity and fragmentation indices, classification of vegetation into high/medium/low spread-risk classes, estimated fractions of continuous fuels and natural/artificial barriers, and an area-weighted spread-potential score. (iv) **Weather (escalation risk):** FRP-weighted Burning Index, wind speed, air temperature (Tmax/Tmin), and 1-hr fuel moisture (FM1) aggregated over the footprint and aligned to the analysis day. Scale adapts per query (cluster footprint vs. county buffer) and per variable (daily vs. hourly products). All joins and raster stats are executed server-side to minimize token and bandwidth costs.

Feature consolidation. GAL consolidates retrieved signals into cluster-level features and a global snapshot. Cluster features include point count, sum FRP, max brightness, hotspot-weighted weather, land-cover mix, exposed population, and nearest-station distances. The global snapshot contains totals (points/FRP), extremes (max FRP, max brightness), and robust distributional measures (e.g., median FRP-per-cluster, $p95$). For $t > 1$, GAL computes temporal anchors: 3-day/7-day rolling means and maxima, as well as qualitative deltas ("↑" or "≈" or "↓") for points, FRP, burned area, and key weather fields. These stabilize incremental decisions without requiring cluster lineage.

3.4 GAL-based LLM Reasoning

We transform geospatial information into a structured, language-ready context that guides LLMs through grounded reasoning. The process combines three elements: a permutation-invariant, unit-

[3] https://doi.org/10.5066/P9JZ7AO3

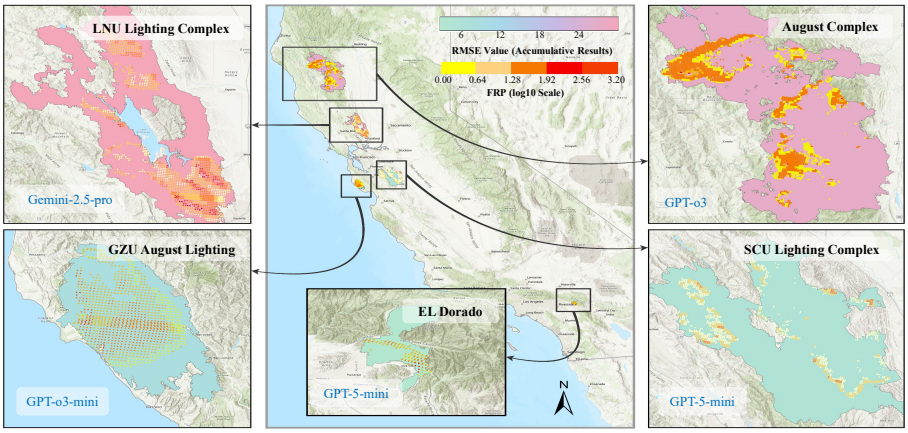


Figure 2: Predictive accuracy across five 2020 California wildfires and hotspot distributions colored by FRP.

locked perception script that standardizes spatial inputs; a retrieval-augmented module that supplies historical analogs as priors; and a rubric-guided chain-of-thought (CoT) that organizes reasoning into interpretable and physically consistent steps.

Perception script. GAL renders heterogeneous geospatial evidence into a compact, unit-annotated perception script with fixed slots and permutation invariance, so the agent reasons over stable, semantically aligned inputs rather than raw geometries or long tables. The script contains: (i) **Top-K cluster summaries** for the most consequential activity areas; (ii) a **global snapshot** of the event day; and (iii) a **unit lock/normalization** step (temperatures in Kelvin, wind in m/s, FM1 in %, standardized magnitudes). Missing fields are marked `NA` with defaults to reduce hallucination and numeric drift.

Analogs scaffold. To supply scale priors, we attach a shared historical-analogs module based on retrieval-augmented generation (Lewis et al., 2020; Yu et al., 2025b). This module represents each event–day snapshot as a standardized numeric vector comprising z-scored features for fire activity, exposure, terrain, and weather, along with categorical flags. Weighted cosine similarity retrieves the top-k prior analog days with unique-by-fire de-duplication. For active fire days, the vector is built from current GAL signals; for no-hotspot days, a rolling 3/7-day trajectory vector is used instead. These ranges provide soft bounds for personnel and cost estimation.

Prompted reasoning with rubric-guided CoT. We transform the script and analog priors into an auditable decision via a rubric-guided CoT (Wei et al., 2022). The prompt requires brief justifications on terrain, weather, fire intensity, exposure, and access, and produces a structured JSON with six intermediate indicators {spread difficulty, deployment access, weather risk, terrain complexity, exposure density, station coverage} on a five-level scale. Two modes share one interface: *Day-1* reasons from the current script plus top-k analogs; *Incremental* additionally consumes yesterday's outputs and compact change cues, prompting a short comparative rationale before adjustment. A validator enforces JSON schema, unit correctness, and bounded ranges; failures trigger targeted re-prompts. Stability comes from fixed field slots, normalized magnitudes, and analog-anchored bounds. Example prompts are provided in **Appendix A**.

4 Experiments

4.1 Dataset

Following the data sources described in Section 3.2, we evaluate the proposed framework on fourteen large wildfire events that occurred in California during the 2020 season. The dataset spans diverse ignition sources, terrains, and durations (see Table 3 in **Appendix B**), covering incidents from late July through early October. Among them, five major events (CZU AUG LIGHTNING, EL DORADO, LNU LIGHTNING COMPLEX, SCU LIGHTNING COMPLEX, and AUGUST COMPLEX) are held out for evaluation, while the remaining nine incidents constitute the historical corpus used by the RAG module and for training non-LLM baselines. This split ensures spatial and temporal disjointness between training and evaluation while maintaining realistic wildfire diversity.

4.2 Benchmark Models

To establish a unified benchmark for wildfire response, our evaluation consists of the following:
Baselines. The Physical baseline converts fire radiative power and area into flame length via fireline intensity, assigns an NWCG flame-length class,

Models	CZU		ELDO		LNU		SCU		AUGC	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Physical	3.2178	3.6681	1.2623	1.4817	4.4987	4.9380	1.9013	2.1603	2.6976	3.6299
LSTM	5.5422	5.9243	2.7959	3.1506	6.8444	7.1553	3.3554	3.8586	4.3175	4.9412
Gemini-2.5-Flash	4.4517	5.1462	1.0977	1.3941	4.6493	5.3697	2.0380	2.8660	3.2682	3.9792
GPT-4o-mini	4.0891	4.6469	1.9234	2.1657	3.9590	4.5367	**1.0770**	**1.5466**	3.8893	4.3974
GPT-4o	4.0409	4.6471	1.2134	1.4672	3.6717	4.3117	1.8040	2.4694	3.2277	4.0495
GPT-4.1-mini	3.9966	4.6667	1.5977	1.9122	4.5579	5.2680	1.8740	2.8371	2.6463	3.3665
GPT-4.1	4.0974	4.8116	1.7389	1.9735	4.3107	5.0133	2.7487	3.9777	3.6742	4.5958
GPT-5-mini	2.4600	3.0067	**0.9086**	**1.0845**	2.6693	3.1717	3.0500	3.3324	2.5408	**3.3326**
GPT-5	2.8482	3.2180	0.9126	1.1722	3.0859	3.7891	1.5530	1.7762	2.8172	3.6462
Gemini-2.5-Pro	4.8717	5.7561	1.9129	2.4405	4.1486	4.9068	3.1723	4.7006	3.2702	3.9927
GPT-o3-mini	**2.2826**	**2.8561**	1.6263	1.8525	**1.2259**	**1.6785**	1.5093	1.6869	**2.5165**	3.3576
GPT-o3	3.2583	3.7656	1.2737	1.5457	4.0952	4.7052	1.3927	1.6506	2.7599	3.4331

Abbrev.: CZU = CZU Aug. Lightning; ELDO = El Dorado; LNU = LNU Lightning; SCU = SCU Lightning; AUGC = August Complex.

Table 1: Daily personnel prediction results (MAE/RMSE) on five held-out wildfires.

Models	CZU		ELDO		LNU		SCU		AUGC	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Physical	1.2484	1.3979	0.9421	1.4042	2.9998	3.5513	1.6222	1.9078	1.9755	2.5039
LSTM	1.2009	1.7702	1.3117	1.9195	3.7605	4.4340	2.2523	2.7913	1.8712	2.6876
Gemini-2.5-flash	1.4029	1.7572	**0.7971**	1.3868	2.7750	3.5688	1.6066	1.9672	2.2081	2.6126
GPT-4o-mini	1.3520	1.6307	0.8006	1.4770	2.9398	3.7167	1.2701	1.5829	2.1567	2.5263
GPT-4o	1.3071	1.5944	0.8062	1.3992	2.4031	3.2454	1.3549	1.7213	1.8794	2.3381
GPT-4.1-mini	1.2128	1.5345	0.7999	1.4345	3.0051	3.8043	1.4321	1.7667	1.8045	2.3087
GPT-4.1	1.5083	1.8928	0.8706	1.5239	2.9922	3.8010	2.1056	2.7367	2.0662	2.5178
GPT-5-mini	1.0229	1.3740	0.7973	1.3633	2.5250	**3.1504**	**0.9999**	**1.3888**	2.0941	2.4502
GPT-5	1.0230	1.3284	0.8072	1.4238	2.6781	3.3963	1.2822	1.5542	1.7897	**2.2270**
Gemini-2.5-pro	1.5842	2.1840	1.3254	2.0212	**2.3426**	3.3043	2.0097	2.5964	2.2865	2.6452
GPT-o3-mini	**0.9530**	**1.2599**	0.8418	**1.3535**	2.7487	3.3869	1.6596	1.9789	1.8401	2.2779
GPT-o3	1.2494	1.5501	0.8804	1.5253	2.9812	3.7631	1.7996	2.1649	**1.7874**	2.2328

Abbrev.: CZU = CZU Aug. Lightning; ELDO = El Dorado; LNU = LNU Lightning; SCU = SCU Lightning; AUGC = August Complex.

Table 2: Daily cost prediction results (MAE/RMSE) on five held-out wildfires.

constructs a workload score (perimeter $\times$ class, with minor cluster adjustments), and linearly calibrates it to the target variables. The LSTM baseline is a many-to-one sequence model trained on recent multi-day inputs (wildfire activity and weather), with prediction for next-day personnel and cost.

LLMs. We evaluate two major families of large language models under a consistent setup. Regular models emphasize throughput and efficiency, including Gemini 2.5 Flash, GPT-4o, GPT-4.1, and GPT-5 (and their mini variants). Reasoning models target multi-step decision-making and compositional reasoning, including Gemini 2.5 Pro, GPT-o3, and GPT-o3-mini. All models share the same prompts, RAG module, and output schema to ensure comparability, with no fine-tuning applied.

4.3 Evaluation Protocol

The task is daily, event-level forecasting: given same-day detections and GAL-derived features, the model predicts (i) required personnel and (ii) daily budget. For a fire of length N with predictions $\hat{y}_i$ and ground truth y_i, we report MAE and RMSE. MAE measures the average absolute deviation, whereas RMSE highlights larger errors that show operational under- or over-allocation.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \quad \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}.$$

4.4 Model Performance

Figure 2 summarizes the best cumulative predictive performance achieved by these models across the five held-out wildfires. Tables 1–2 report the MAE and RMSE for daily personnel and cost forecasts.

LLMs outperform baselines. Across all evaluations, the lowest error in every column is achieved by an LLM. For daily personnel, GPT-o3-mini, GPT-5-mini, and GPT-4o-mini each lead on different events, while the Physical and LSTM baselines never achieve the top rank. For daily cost, GPT-o3-mini and GPT-5-mini again dominate most incidents, with Gemini-2.5-Pro and Gemini-2.5-Flash showing competitive results in select cases. These outcomes highlight the strong generalization of geospatially grounded LLMs.

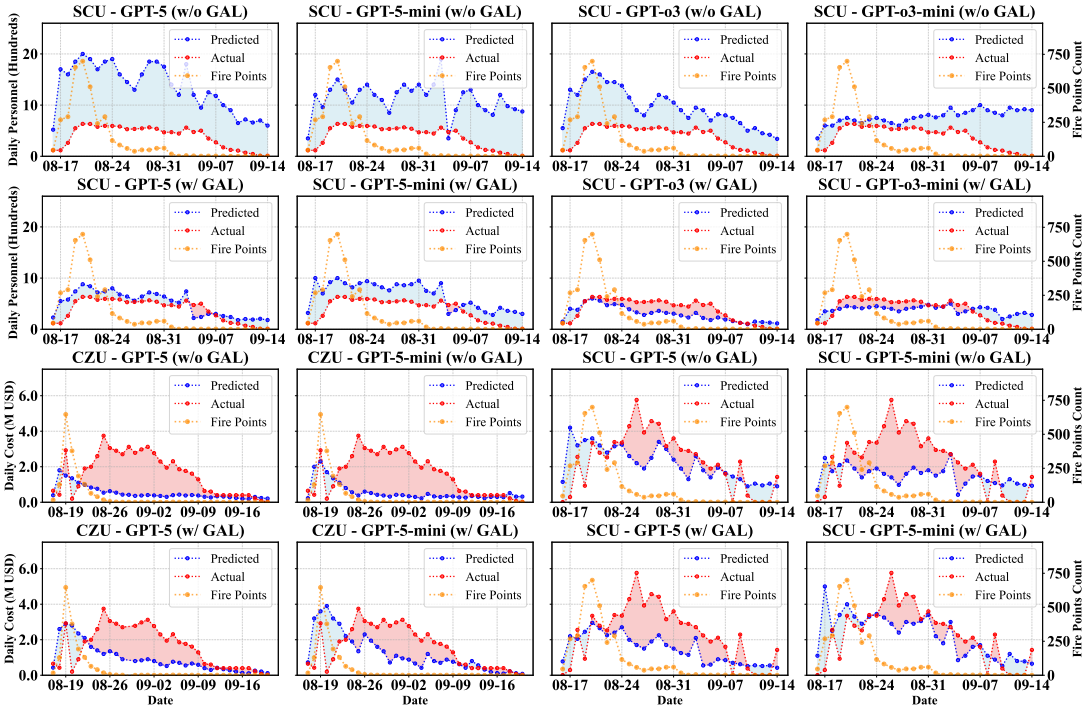


Figure 3: Ablation of the Geospatial Awareness Layer on daily personnel and daily cost forecasting.

GAL enhances robustness for complex fires. As shown in Figure 2, complex and spatially fragmented fires such as LNU and August exhibit inherently higher uncertainty, reflected by larger RMSE values. Even under these challenging settings, grounded LLMs maintain competitive performance and avoid catastrophic errors, while achieving lower variance across smaller and more localized fires such as El Dorado and SCU. This suggests that GAL improves the model's stability and situational consistency, rather than overfitting to particular spatial configurations.

Compact reasoning models remain competitive. Smaller variants such as GPT-o3-mini, GPT-4o-mini, and GPT-5-mini often match or even surpass their larger counterparts across multiple fires. This counterintuitive pattern arises because the structured and low-entropy inputs provided by GAL reduce linguistic ambiguity and emphasize quantitative reasoning. Under these stable, unit-normalized conditions, compact models adhere more consistently to schema and numerical cues, achieving higher stability and precision even without the capacity overhead of larger models.

4.5 Ablation Studies

We isolate the contribution of each module through two controlled settings: (i) w/o GAL versus w/ GAL, where the former exposes the model only to raw fire detections and coarse daily summaries, and

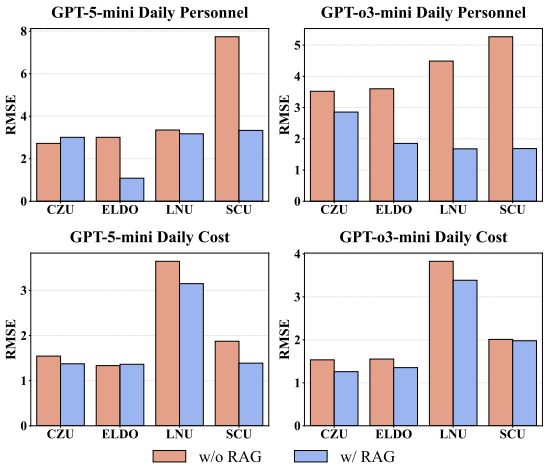


Figure 4: Ablation of the RAG module.

the latter provides the full perception script from the Geospatial Awareness Layer; and (ii) w/o RAG versus w/ RAG, where historical analogs are omitted or appended to the prompt. Prompts, decoding settings, and evaluation protocols are held constant across variants and models.

Figure 3 visualizes the effect of GAL on daily personnel (top two rows) and daily cost (bottom two rows) forecasts across representative wildfire events. In each panel, yellow lines denote fire hotspot counts, red lines show ground-truth records, and blue lines represent model predictions; the shaded areas indicate the magnitude of prediction error. The upper rows compare models without GAL (w/o GAL) and the lower rows show results with GAL (w/ GAL) under identical prompting and

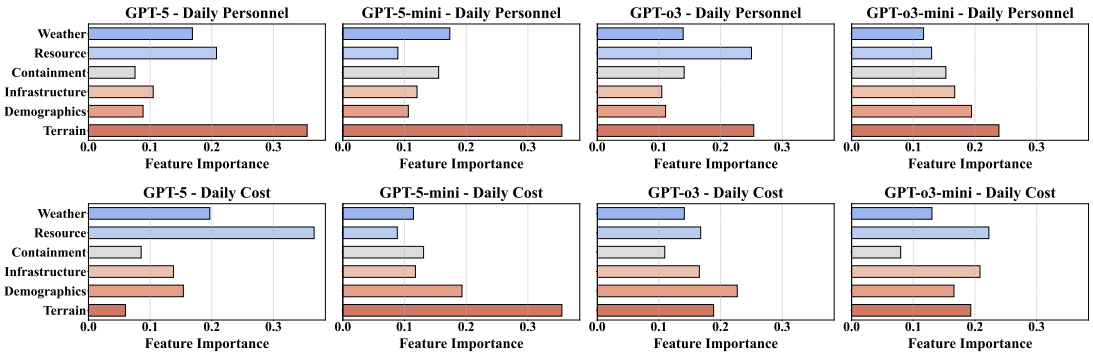


Figure 5: Feature importance analysis of LLM reasoning across four models.

evaluation settings. Overall, introducing GAL produces clearer temporal alignment, smoother amplitude transitions, and more realistic decay behavior across both tasks. For SCU personnel, w/o-GAL models substantially overestimate resource levels, maintaining inflated staffing long after the fire subsides. In contrast, w/ GAL predictions follow the actual red curve more closely, with synchronized rise (fall phases and a visibly reduced error region around the incident peak), suggesting better recognition of operational draw-down. For CZU and SCU costs, similar stabilization effects emerge: baseline models tend to underreact early or sustain overly long cost plateaus, whereas w/ GAL variants better capture the inertia of spending.

Figures 4 present RMSE bar charts for daily personnel and cost, comparing w/o RAG (orange) and w/ RAG (blue) across four held-out fires. Lower blue bars indicate improvements from historical retrieval. Overall, RAG reduces error in most cases, though the magnitude varies by model and event. For daily personnel, GPT-o3-mini consistently improves across all fires, with the largest gains on LNU and SCU, while GPT-5-mini achieves clear reductions on ELDO and SCU but shows a slight regression on CZU, likely due to distinctive fire dynamics that make retrieved analogs less relevant. For daily cost, both models achieve steady and robust gains across events. These results suggest that RAG enables models to incorporate cross-event experience to stabilize predictions under sparse or uncertain conditions, enhancing temporal consistency while maintaining geospatial grounding.

4.6 Output Reasoning Analysis

To better understand what drives the models' decisions, we train a lightweight post-hoc regressor for each (model, task) pair using the feature labels extracted by the LLMs from the GAL perception script and referenced in their reasoning traces. Features are grouped into six categories: terrain, demographics, infrastructure, containment, resource, and weather. Figure 5 reports normalized feature importances. First, for daily personnel prediction, terrain is typically the dominant driver across all four models. Second, for the daily cost task, the influence shifts in a model-dependent way rather than being dominated by a single factor: GPT-5-mini retains terrain as its leading cue, and GPT-o3 variants allocate weight more evenly, with visible contributions. Notably, GPT-5 shows the clearest transition in focus, from terrain in personnel forecasting to resource in cost forecasting, suggesting that larger models may learn to separate operational capacity factors from environmental constraints when estimating expenditures. Additional qualitative examples are provided in **Appendix D**.

5 Conclusion

This study bridges LLMs and geospatial realities by introducing a Geospatial Awareness Layer (GAL) that grounds reasoning in structured spatial evidence. From minimal wildfire inputs, GAL retrieves infrastructure, demographic, terrain, and weather information, fuses them into unit-locked perception scripts, and pairs them with historical analogs for auditable resource decisions. Evaluations across California fires show that geospatially grounded agents outperform physical and LSTM baselines in forecasting daily personnel and cost, with higher accuracy and temporal stability. Results indicate that structured spatial grounding matters more than model scale, while spatial-temporal cues and analog retrieval improve robustness. Feature attribution reveals that personnel forecasts depend primarily on spread potential and exposure, whereas cost estimation emphasizes accessibility and logistical demand, reflecting real-world operational priorities. Beyond the wildfire context, GAL offers a generalizable interface for other hazards such as floods, hurricanes, and earthquakes.

6 Limitations

Several limitations warrant further consideration. First, our evaluation focuses on a subset of large 2020 California wildfires. Although GAL is hazard-agnostic and easily extendable, validating its generality across other regions, years, and hazard types (e.g., floods, earthquakes) will require additional data connectors and local calibration. Second, the framework depends on public geospatial datasets (e.g., FIRMS, ACS, HIFLD, NLCD), which may contain detection noise, temporal delays, or resolution mismatches. Our normalization and aggregation steps mitigate these effects, but data quality still constrains absolute accuracy. Third, the feature-importance analysis is post-hoc and descriptive rather than causal. It reveals how GAL-derived factors correlate with model outputs, but further controlled studies are needed to establish causal mechanisms in real operational settings.

7 Ethics

This study uses only public, aggregated datasets without personal information. Demographic variables serve exposure estimation, not prioritization. All outputs are logged for auditability, and compact models are used to reduce computational cost. Potential misuse or bias is mitigated through uncertainty displays and human oversight. Code and documentation will be released for reproducible, non-surveillance use in public safety contexts.

8 Statement of AI Assistant Use

During the development of this paper, we leveraged LLM-based assistants in the following aspects: (i) Code assistance: coding agents were employed to support the development and debugging of the project's code base. (ii) Grammar check: LLMs were used for grammar checking and stylistic polishing to improve the clarity and readability of the manuscript.

References

John T. Abatzoglou. 2013. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131.

John T. Abatzoglou and A. Park Williams. 2016. Impact of anthropogenic climate change on wildfire across western us forests. *Proceedings of the National Academy of Sciences*, 113(42):11770–11775.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

David Chaparro, Merce Vall-Llossera, Maria Piles, Adriano Camps, Christoph Rüdiger, and Ramon Riera-Tatché. 2016. Predicting the extent of wildfires using remotely sensed soil moisture and temperature trends. *IEEE journal of selected topics in applied earth observations and remote sensing*, 9(6):2818–2829.

Minze Chen, Zhenxiang Tao, Weitong Tang, Tingxin Qin, Rui Yang, and Chunli Zhu. 2024. Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, 113:104804.

Ruxiao Chen, Chenguang Wang, Yuran Sun, Xilei Zhao, and Susu Xu. 2025. From perceptions to decisions: Wildfire evacuation decision prediction with behavioral theory-informed LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29754–29778, Vienna, Austria. Association for Computational Linguistics.

Yiheng Chen, Alina Hagen, Fan Yang, Ratna B Dougherty, Zihui Ma, Lingyao Li, and Runlong Yu. 2026. Politicized attention shifts amplify polarization in the information ecosystem during california wildfires. *arXiv preprint arXiv:2603.19536*.

Janice L Coen, Marques Cameron, John Michalakes, Edward G Patton, Philip J Riggan, and Kara M Yedinak. 2013. Wrf-fire: coupled weather–wildland fire modeling with the weather research and forecasting model. *Journal of Applied Meteorology and Climatology*, 52(1):16–38.

Susan L. Cutter. 2016. The landscape of disaster resilience indicators in the usa. *Natural Hazards*, 80:741–758.

Pei Dang, Jun Zhu, Weilian Li, Yakun Xie, and Heng Zhang. 2025. Large-language-model-driven agents for fire evacuation simulation in a cellular automata environment. *Safety Science*, 191:106935.

Christopher J. Dunn, David E. Calkin, and Matthew P. Thompson. 2017. Towards enhanced risk management: Planning, decision making and monitoring of us wildfire response. *International Journal of Wildland Fire*, 26(7):551–556.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Mark A. Finney. 1998. *FARSITE, Fire Area Simulator–model development and evaluation*. 4.

Louis Giglio, Wilfrid Schroeder, and Christopher O. Justice. 2016. The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, 178:31–41.

Vinicius G Goecks and Nicholas R Waytowich. 2023. Disasterresponsegpt: Large language models for accelerated plan of action development in disaster response scenarios. *arXiv preprint arXiv:2306.17271*.

Yingjie Hu, Gengchen Mai, Chris Cundy, Kristy Choi, Ni Lao, Wei Liu, Gaurish Lakhanpal, Ryan Zhenqi Zhou, and Kenneth Joseph. 2023. Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*, 37(11):2289–2318.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Fantine Huot, R Lily Hu, Nita Goyal, Tharun Sankar, Matthias Ihme, and Yi-Fan Chen. 2022. Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13.

Ryan Laube and Howard J Hamilton. 2021. Wildfire occurrence prediction using time series classification: A comparative study. In *2021 ieee international conference on big data (big data)*, pages 4178–4182. IEEE.

Zhenyu Lei, Yushun Dong, Weiyu Li, Rong Ding, Qi R Wang, and Jundong Li. 2025. Harnessing large language models for disaster management: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14528–14551.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc.

Bowen Li, Zhaoyu Li, Qiwei Du, Jinqi Luo, Wenshan Wang, Yaqi Xie, Simon Stepputtis, Chen Wang, Katia P. Sycara, Pradeep Kumar Ravikumar, Alexander Gray, Xujie Si, and Sebastian Scherer. 2024. LogiCity: Advancing neuro-symbolic ai with abstract urban simulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 69840–69864.

Lingyao Li, Michelle Bensi, and Gregory Baecher. 2023. Exploring the potential of social media crowdsourcing for post-earthquake damage assessment. *International Journal of Disaster Risk Reduction*, 98:104062.

Lingyao Li, Dawei Li, Zhenhui Ou, Xiaoran Xu, Jingxiao Liu, Zihui Ma, Runlong Yu, and Min Deng. 2025a. Llms as world models: Data-driven and human-centered pre-event simulation for disaster impact assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3096.

Lingyao Li, Runlong Yu, Qikai Hu, Bowei Li, Min Deng, Yang Zhou, and Xiaowei Jia. 2025b. From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models. *arXiv preprint arXiv:2508.01608*.

Zihui Ma, Lingyao Li, Libby Hemphill, Gregory B. Baecher, and Yubai Yuan. 2024. Investigating disaster response for resilient communities through social media data and the susceptible-infected-recovered (sir) model: A case study of 2020 western u.s. wildfire season. *Sustainable Cities and Society*, 106:105362.

David L. Martell. 2015. A review of recent forest and wildland fire management decision support systems research. *Current Forestry Reports*, 1(2):128–137.

Regina Obe and Leo Hsu. 2021. *PostGIS in action*. Simon and Schuster.

Hakan T Otal, Eric Stern, and M Abdullah Canbaz. 2024. Llm-assisted crisis management: Building advanced llm platforms for effective emergency response and public collaboration. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 851–859. IEEE.

Haiganoush K Preisler and Anthony L Westerling. 2007. Statistical model for forecasting monthly large wildfire events in western united states. *Journal of Applied Meteorology and Climatology*, 46(7):1020–1030.

Vincent Quitoriano and David J Wald. 2020. Usgs "did you feel it?"—science and lessons from 20 years of citizen science-based macroseismology. *Frontiers in Earth Science*, 8:120.

Meghana Ramesh, Ziheng Sun, Yunyao Li, Li Zhang, Sai Kiran Annam, Hui Fang, and Daniel Tong. 2025. Assessing wildfiregpt: a comparative analysis of ai models for quantitative wildfire spread prediction. *Natural Hazards*, 121(11):13117–13130.

Richard C. Rothermel. 1972. A mathematical model for predicting fire spread in wildland fuels. Research Paper Res. Pap. INT-115, U.S. Department of Agriculture, Intermountain Forest and Range Experiment Station, Ogden, UT.

Lise A. St. Denis, Karen C. Short, Kathryn McConnell, Maxwell C. Cook, Nathan P. Mietkiewicz, Mollie Buckland, and Jennifer K. Balch. 2023. All-hazards dataset mined from the us national incident management system 1999–2020. *Scientific Data*, 10(1):112.

Yinzhou Tang, Huandong Wang, Xiaochen Fan, and Yong Li. 2025. Predicting human mobility in disasters via llm-enhanced cross-city learning. *arXiv preprint arXiv:2507.19737*.

Matthew P. Thompson, Yu Wei, David E. Calkin, Christopher D. O'Connor, Christopher J. Dunn, Nathaniel M. Anderson, and John S. Hogland. 2019. Risk management and analytics in wildfire response. *Current Forestry Reports*, 5(4):226–239.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. abs/2307.09288.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Martin J. Wooster, Gareth Roberts, George L. W. Perry, and Yoram J. Kaufman. 2005. Retrieval of biomass combustion rates and totals from fire radiative power observations: Frp derivation and calibration relationships between biomass consumption and fire radiative energy release. *Journal of Geophysical Research: Atmospheres*, 110(D24).

Yangxinyu Xie, Tanwi Mallick, Joshua David Bergerson, John K. Hutchison, Duane R. Verner, Jordan Branham, M. Ross Alexander, Robert B. Ross, Yan Feng, Leslie-Anne Levy, and Weijie J. Su. 2024. Wildfiregpt: Tailored large language model for wildfire analysis. abs/2402.07877.

Wenping Yin, Yong Xue, Ziqi Liu, Hao Li, and Martin Werner. 2025. LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of hurricane harvey. *International Journal of Applied Earth Observation and Geoinformation*, 137:104423.

Runlong Yu, Shengyu Chen, Yiqun Xie, and Xiaowei Jia. 2025a. A survey of foundation models for environmental science. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 39–57. Springer.

Runlong Yu, Shiyuan Luo, Rahul Ghosh, Lingyao Li, Yiqun Xie, and Xiaowei Jia. 2025b. Rag for geoscience: What we expect, gaps and opportunities. *arXiv preprint arXiv:2508.11246*.

Ahmed El Fekih Zguir, Ferda Ofli, and Muhammad Imran. 2025. Roadmind: Towards a geospatial ai expert for disaster response. *arXiv preprint arXiv:2509.19354*.

Xiaofeng Zhong, Yinfeng Xiang, Fang Yi, Chao Li, and Qinmin Yang. 2024. Hmp-llm: Human mobility prediction based on pre-trained large language models. In *2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 687–692. IEEE.

Appendix

The appendix provides additional materials to support reproducibility, interpretability, and extended evaluation. We first present the complete prompt design used in the GAL-based reasoning pipeline, including both system and user prompts, to make the model-facing decision interface transparent. We then summarize the wildfire events used in the study and provide qualitative case analyses that illustrate how geospatial grounding affects model behavior over time, including representative failure modes without GAL. To further contextualize the main results, we report an extended benchmark against additional time-series baselines and LLM variants. Finally, we include token usage and API cost statistics to characterize the practical efficiency of different model configurations.

LLM-based assistants were used only for code assistance and minor language polishing. Our experiments evaluated the capabilities of several commercial LLMs, such as OpenAI's GPT and Google's Gemini. The total API expenditure for these experiments was approximately $300.

A Prompt Design

[SYSTEM PROMPT]

```
You are a wildfire analysis and resource
    management expert. You must return ONLY
    a valid JSON object following the exact
    schema provided below.

### Global Guidelines
- The task is to estimate TODAY's required
    daily_personnel and daily_budget.
```

```
- reasoning must explain how terrain,
    weather, fire intensity, population
    exposure, and resource accessibility
    shape your judgment, considering both
    current conditions and previous analysis
     context.
- daily_personnel is the total integer
    headcount assigned today (all crews/
    engines/aviation modules plus command/
    overhead/support).
- daily_budget is the **new cost incurred
    today only**, in USD.

### Resource Estimation Principles
- If the fire surges, remember resources are
     finite-do not assume cost and personnel
     can scale proportionally.
- When the fire eases, non-suppression needs
     persist (patrol, mop-up, rehab,
    logistics); budget and staffing may
    still be required.
- In "stable" periods, account for
    cumulative costs and crew fatigue-
    budgets and crews are not unlimited.
- No detected hotspots $\neq$ full
    extinguishment; avoid indiscriminate
    cuts and maintain a prudent baseline.
- Weigh these trade-offs and produce a
    balanced, defensible recommendation for
    today's personnel and today's spend.
    Include any key assumptions and risks.
- **Common pitfall**: after you've committed
     resources and the fire is "under
    control" but not yet stable, that
    actually signals under-resourcing-
    maintain or increase resources until
    true stability is confirmed.

### Analysis Approach
- Analyze the fire situation holistically,
    considering today's conditions and
    changes from the previous analysis.
- Provide updated estimates for required
    daily_personnel and daily_budget based
    on your professional judgment.

### Output Schema (STRICT JSON; no extra
    keys; no comments)
{
  "analysis_reasoning": {
    "situation_comparison": "<2-3 sentences
    comparing today vs yesterday>",
    "personnel_reasoning": "<2-3 sentences
    explaining daily_personnel changes>",
    "budget_reasoning": "<2-3 sentences
    explaining daily_budget changes>",
    "overall_reasoning": "<2-3 sentences
    with overall change assessment>"
  },
  "resource_requirements": {
    "daily_personnel": {
      "value": "<integer>",
      "unit": "people"
    },
    "daily_budget": {
      "value": "<integer>",
      "unit": "USD"
    }
  },
  "confidence": {
    "score": "<1-5 integer>"
  },
  "intermediate_indicators": {
    "spread_containment_difficulty": "<
    minimal|low|moderate|high|critical>",
    "resource_access_deployment": "<minimal|
    low|moderate|high|critical>",
    "weather_escalation_risk": "<minimal|low
    |moderate|high|critical>",
    "terrain_operational_complexity": "<
    minimal|low|moderate|high|critical>",
    "population_exposure_density": "<minimal
    |low|moderate|high|critical>",
    "fire_station_coverage": "<minimal|low|
    moderate|high|critical>"
  }
}
```

[USER PROMPT]

```
## Previous Analysis Context
- Previous personnel: <int> people
- Previous daily budget: <int>
- Total cumulative cost: $<int>
- Previous reasoning: <text>

## Cumulative Context
- Total cumulative cost: $<int>
- Total cumulative personnel-days: <int>
- Days since fire start: <int>
- 3-day rolling avg daily cost: $<int>
- 3-day rolling avg daily personnel: <int>
- 7-day rolling avg daily cost: $<int>
- 7-day rolling avg daily personnel: <int>
- Recent cost trend: <increasing|decreasing|
    stable>
- Recent personnel trend: <increasing|
    decreasing|stable>

## Fire Intensity Rolling Metrics
- 3-day avg fire points: <float>
- 3-day max fire points: <float>
- 7-day avg fire points: <float>
- 7-day max fire points: <float>
- Current fire points vs historical max: <
    percent>%
- Global max fire points: <float> (<days>
    days ago)
- 3-day avg total area: <float> acres
- 7-day avg total area: <float> acres
- Current area vs historical max: <percent>%
- Global max area: <float> acres (<days>
    days ago)

## Fire Overview vs Yesterday
- Current date: <MM-DD>
- Total Fire Points: <int> (up/down <delta>)
- Num Clusters: <int> (up/down <delta>)
- Total FRP: <float> MW (up/down <delta>, <
    percent>%)
- Total area: <float> acres (up/down <delta>,
     <percent>%)
- Max FRP/Brightness: <float> (up/down <
    delta>)
```

```
- Weather conditions: BI=<float>, Tmax/Tmin
    =<float>, Wind=<float>, FM1=<float>%

## Affected Areas vs Yesterday
- Counties: removed {<names>}; now <int>
- Total Population Affected: <int> (up/down
    <delta>)
- Fire stations in area: <int> (up/down <
    delta>)
- Nearest station: <float> mile (up/down <
    delta>)

## Historical Context (RAG)
- [<FIRE> <DATE>] sim=<float> | Personnel=<
    float>, Daily_Budget=$<float>
- [repeat for top-K similar cases]

## Cluster Details
- Cluster <ID>:
  fire[points=<int>, frp=<float>, brightness
    =<float>, area=<float>]
  weather[BI=<float>, tmax=<float>, tmin=<
    float>, wind=<float>, FM1=<float>]
  location[<County>, <State>, pop=<int>,
    station_1/2/3=<float> mile]
  terrain: NLCD analysis - land cover,
    vegetation coverage, fire spread
    potential
```

B Wildfire Events and Statistics

We summarize the fourteen large California wildfire events used in this study in Table 3. The dataset covers incidents from late July through early October 2020. The top section lists the nine fires used as the historical corpus for model training and RAG retrieval, while the bottom section lists the five events reserved for evaluation.

C Failure case analysis

To illustrate why GAL helps, we present two representative cases in which removing GAL causes systematic behavioral failures.

SCU Lightning Complex: personnel over-allocation. Figure 6 (left) compares GPT-5 with and without GAL on SCU personnel during 8/16–8/20. Without GAL, the model relies primarily on FRP and area counts, reasoning: *"Two-cluster fire remains highly energetic... no clear reduction in heat... maintain robust staffing until hotspot density and heat signatures trend down"*, and over-allocates by a factor of 2–3 relative to ground truth. With GAL, the model explicitly references fuel type, cluster structure, and population exposure: *"Population exposure remains low... Given finite regional resources and limited population density, we stop short of mega-fire levels"*, producing a moderate, resource-constrained trajectory close to observations. This illustrates a failure mode where FRP-only signals are insufficient to calibrate operationally realistic staffing in low-exposure regions.

Wildfire Event	Start Date (2020)	Duration (days)
Training / Historical Corpus		
CREEK	Sep 5	84
DOLAN	Aug 18	51
FORK	Sep 8	32
NORTH COMPLEX	Aug 17	83
RED SALMON COMPLEX	Jul 28	103
SLATER	Sep 8	61
SQF COMPLEX	Aug 24	111
SLINK	Aug 29	40
WOODWARD	Aug 18	28
Evaluation / Test Set		
CZU AUG LIGHTNING	Aug 17	35
EL DORADO	Sep 5	35
LNU LIGHTNING COMPLEX	Aug 18	29
SCU LIGHTNING COMPLEX	Aug 16	30
AUGUST COMPLEX	Aug 17	83

Table 3: Summary of 2020 California wildfire events.

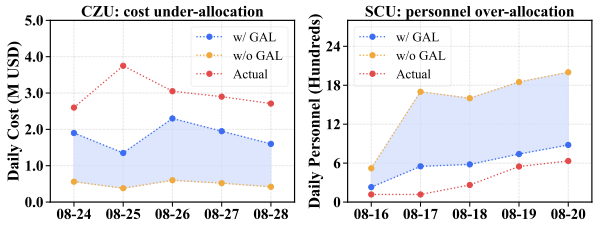


Figure 6: Failure case analysis.

CZU August Lightning: cost under-allocation. Figure 6 (right) compares GPT-5-mini with and without GAL on CZU daily cost during 8/24–8/28. Without GAL, the model is driven by changes in area and FRP alone, remaining conservative in magnitude and underestimating true cost by more than a factor of four on peak days. With GAL, the model directly uses terrain and exposure structure—*"fuel continuity is very high in evergreen-dominated terrain... population exposure grew markedly as the large cluster expanded in Santa Cruz"*—and justifies substantial cost increases for aviation, overtime, and structure protection, yielding a trajectory that tracks the high observed costs.

In summary, these cases show that GAL corrects two systematic failures of text-only agents: over-allocation in low-exposure regions and under-allocation in high-risk terrain, yielding operationally meaningful gains.

D Case Study: Daily Forecast Trajectory

Figure 7 illustrates the daily forecast trajectories of GPT-o3-mini under the GAL-based reasoning setup. The top panel plots the predicted daily

personnel and daily cost. The numbered callouts [1]–[9] are the model's auditable rationales, emitted under our rubric-guided CoT. What follows are concrete behaviors that emerge only because the LLM is reasoning over GAL's structured context and analog priors. These outputs demonstrate how GAL supports grounded and interpretable reasoning in dynamic disaster contexts.

Early phase (08–17 to 08–29). The model links initial staffing decisions to high vegetation risk, multi-county spread, and limited station coverage, maintaining moderate personnel despite modest FRP levels. When activity shifts toward more populated areas (08–21), the agent increases staffing even as FRP declines, prioritizing exposure and accessibility over fire intensity. This behavior reflects multi-factor reasoning enabled by GAL's structured context, which aligns spatial and demographic factors on comparable scales.

Stabilization and drawdown (08–25 to 09–02). As fire intensity eases, the model gradually reduces personnel and cost without oscillations. Rationales cite "crew fatigue" and "logistical readiness," showing that the LLM draws on GAL's temporal anchors and analog priors rather than reacting to daily noise. The reduction proceeds smoothly, maintaining essential logistics—evidence that the analog-conditioned soft bounds effectively regularize quantitative outputs.

Localized escalation (09–06). When the footprint becomes more spatially concentrated yet closer to populated zones, the model justifies a small resource increase, citing "higher intensity and rising risk to nearby populations." This response indicates that GAL's geometric and demographic signals—such as buffer-based proximity and exposure density—guide context-sensitive adjustments that reflect operational priorities rather than purely numerical intensity.

Data-sparse periods (09–10, 09–14). On days with minimal or missing hotspot detections, the model maintains a conservative baseline for patrol and mop-up operations. Its explanations explicitly reference contingency readiness, demonstrating that analog-based priors provide stability and prevent collapse to zero allocations when satellite observations are incomplete. This robustness is crucial for real-world deployment, where sensing delays and cloud cover are common.

Re-escalation (09–18). When the footprint expands again, the model raises both personnel and cost, supported by references to "augmented aerial support" and "logistical coordination." These justifications align with renewed increases in FRP and spatial extent, showing that the model adapts promptly when environmental signals re-intensify.

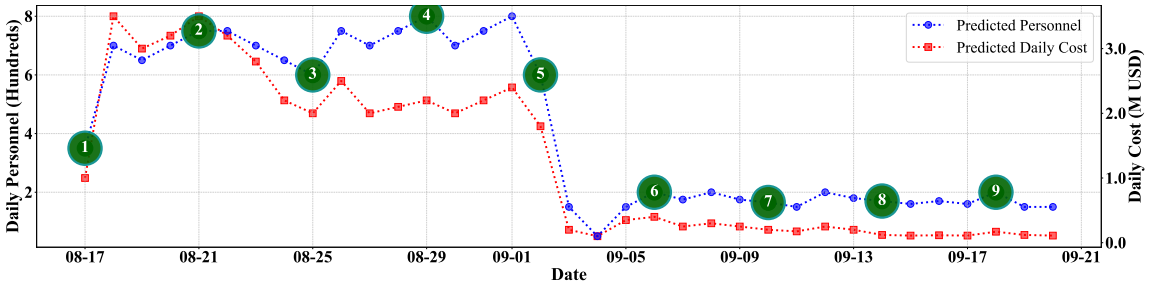
Discussion and remark. These observations provide clear evidence that GAL enhances the reasoning reliability of LLMs in disaster response. It enables explicit multi-objective trade-offs: the agent increases staffing when exposure or access risks outweigh intensity, and scales back when trends and logistics allow—supported by GAL's unified, unit-locked representation. The 3/7-day anchors and Incremental mode promote temporal stability without sacrificing responsiveness, while analog-conditioned priors calibrate personnel and cost magnitudes to realistic ranges, preventing overreaction or drift. Even under missing signals, the model produces conservative, rationale-grounded baselines instead of collapsing to zero. Finally, GAL improves operational explainability, as each chain-of-thought links decisions to verifiable GIS cues such as counties, stations, fuels, and weather.

E Extended Benchmark: Time-Series Baselines and Additional LLMs

To further validate the robustness of our conclusions, we report results on a stricter four-fire subset—CZU August Lightning, El Dorado, LNU Lightning Complex, and SCU Lightning Complex—against an expanded set of competitors across two tasks: daily total personnel and estimated daily cost-to-date.

Non-LLM baselines. Beyond the Physical and LSTM baselines reported in the main paper, we evaluate four strong time-series models: DLinear, TimesNet, PatchTST, and TimeLLM. These methods represent a broad spectrum of modern sequence-forecasting approaches, from simple linear decomposition to patch-based transformers and LLM-augmented forecasters. We also include a **Text-RAG (no GAL)** baseline, which uses the same historical-analog retrieval module as our full system but feeds only unstructured textual summaries of recent FRP and area history to the LLM, without any structured geospatial perception script. This baseline directly tests whether generic retrieval alone—without structured spatial grounding—is sufficient to match GAL-grounded agents.

Additional LLMs. Our main evaluation already spans a diverse set of model families and sizes, in-



[1] 08-17
Personnel: The required daily personnel is estimated at a moderate level given the persistent high-risk vegetation and multiple clusters spread across two counties. Despite low population exposure, the high potential for rapid fire spread and the distances to supporting fire stations justify maintaining or slightly increasing crew numbers.
Budget: The daily budget reflects the need for rapid aerial and ground responses, intensive incident command support, and equipment mobilization. Given the historical cost references and current fuel and weather dynamics, a higher expenditure than routine mop-up operations is warranted for timely intervention.

[2] 08-21
Personnel: Despite the reduction in fire points, the emergence of an additional county and increased community exposure require bolstered staffing for effective perimeter control and support tasks. This leads to a modest increase in daily personnel from yesterday's count to ensure enhanced command, safety, and mop-up operations.
Budget: While the lower FRP and brightness suggest a potential reduction in aggressive suppression costs, the large and continuous fuel-loaded environment along with extended operations and higher population risks drive the need for a higher daily budget. Additional resources are necessary for logistics and support functions given the sustained maximum area and multi-agency coordination.

[3] 08-25
Personnel: With the overall intensity reducing, a moderate reduction in personnel is justified to reflect the current demand while still preserving a robust suppression capacity due to the high fuel risk. The reduction from 650 to 600 maintains operational readiness and addresses crew fatigue concerns from multiple days of activity.
Budget: Given the improved fire parameters, the daily budget is slightly reduced relative to the previous day’s $2200000 while still accommodating the costs of logistics and support required for high-risk operational zones. The adjusted expenditure supports sustained operational readiness without overcommitting resources.

[4] 08-29
Personnel: Given the abrupt FRP increase and the high-risk, continuous fuel conditions of evergreen forests, a modest increase in personnel is warranted to boost aggressive suppression and aerial coordination. Although overall exposure is lower, the challenges of accessing the area due to increased distances to fire stations necessitate extra hands on the ground.
Budget: The intensified fire performance indicators require additional aerial suppression and specialized teams, which in turn drives up operational costs. While recent trends indicate a slight decrease in costs, today's elevated FRP and required rapid response justify a budget increase despite stable personnel trends.

[5] 09-02
Personnel: Yesterday saw 800 personnel deployed during higher fire intensity, and today a reduced presence is justified given the apparent drop in fire activity. Still, a substantial crew is maintained to address potential data gaps, mop-up operations, and crew fatigue from prolonged operations.
Budget: With fire metrics at zero, daily expenditures can be scaled back from the previous $2400000 while still covering essential logistics and support operations. The reduced daily budget reflects both the lower immediate fire suppression needs and the necessity to sustain operational readiness.

[6] 09-06
Personnel: Although the fire perimeter appears more concentrated, the higher intensity and rising risk to nearby populations justify an increase in on-ground staffing. Increasing from the previous 150 people to around 200 will allow for intensified suppression efforts, protective operations, and continued mop-up to prevent further escalation.
Budget: The spike in fire intensity and the need to secure assets and populations suggest a slight increase in daily expenditure over the prior $350000. Allocating about $400000 today can support the required tactical boosts and rapid response deployments without overshooting finite resources.

[7] 09-10
Personnel: Given the current absence of fire points yet factoring in the need for continuous monitoring and readiness against unexpected flare-ups, a modest reduction in personnel compared to recent averages is justified. The recommendation maintains a sufficient baseline to cover patrol, mop-up, and rapid response while managing crew fatigue.
Budget: With fire activity minimal today, the daily budget can be reduced from previous higher levels yet still allocated to essential operations, surveillance, and support services. This balanced approach ensures funds are available for contingency actions should conditions change unexpectedly.

[8] 09-14
Personnel: Even with minimal detected fire activity, baseline personnel remain essential for patrol, mop-up, and rapid response in case of a resurgence. The slight reduction from previous levels reflects the current lull while ensuring sustained readiness.
Budget: The daily budget is decreased compared to prior expenditures due to the absence of active fire suppression needs, yet it still covers ongoing logistics, command overhead, and community safety measures. This adjustment helps manage operational costs while providing for immediate response if conditions change.

[9] 09-18
Personnel: The previous staffing level of 160 is inadequate given the large jump in affected area and complexity of fire behavior. Increasing personnel to 200 allows for more robust ground operations, effective mop-up, and better safety oversight in this expanded incident.
Budget: With the significant increase in fire area and the need for augmented aerial support and logistical coordination, today's budget must rise from $110000 to ensure comprehensive suppression efforts. The recommended $170000 accommodates these additional costs while aligning with recent operational trends.

Figure 7: Daily forecast trajectory for the CZU Lightning Complex using GPT-o3-mini with GAL setup.

cluding Gemini 2.5 Flash/Pro, GPT-4/4.1/4o/5, and the o3 series. In additional runs, we further evaluate DeepSeek-V3.2, Gemini-3-Pro, GPT-5.1, and a GPT-5 configuration with web-enabled retrieval (GPT-5-web), the last of which approximates a simpler web-search–based RAG baseline without structured geospatial context.

Results. As shown in Tables 4–5, modern sequence models such as DLinear, TimesNet, and PatchTST perform comparably to or worse than the Physical baseline, underscoring the difficulty of temporal extrapolation without spatial context. TimeLLM achieves more competitive results on cost prediction, yet still trails the best GAL-grounded LLMs on most fires. The Text-RAG configuration consistently underperforms full GAL agents, confirming that structured geospatial grounding contributes beyond generic retrieval alone. Among the newly added LLMs, GPT-5-web and Gemini-3-Pro are competitive in select settings, while DeepSeek-V3.2 and GPT-5.1 show moderate performance. Across both tasks and all four fires, GAL-grounded LLM agents consistently match or outperform all baselines, reinforcing that structured geospatial grounding adds value beyond generic retrieval or temporal extrapolation alone.

F Token Usage and API Cost per Fire

Table 6 reports the average token usage and estimated API cost per fire for each model evaluated on the four-fire subset (CZU, ELDO, LNU, SCU). These figures reflect the total tokens consumed and dollar cost accumulated across all days of each fire. The results show that the strongest GAL-grounded configurations (e.g., GPT-o3-mini, GPT-4o, GPT-5-mini) achieve competitive predictive performance at modest per-event costs, which is important for real-time or large-scale deployment.

Models	CZU		ELDO		LNU		SCU	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Text-RAG (no GAL)	2.8957	3.4167	1.8742	2.1816	4.2935	4.9984	1.6869	1.9685
DLinear	5.9859	6.3614	3.4593	3.7735	7.3808	7.7136	3.9831	4.4761
TimesNet	4.9538	5.7916	4.2132	5.3438	6.2882	6.9969	5.2743	6.9569
PatchTST	3.5742	3.9399	1.3680	1.6111	4.7462	5.1846	2.2235	2.4021
TimeLLM	6.7984	7.7380	1.6207	1.9352	6.1624	6.5885	2.8362	2.9376
Gemini-2.5-Flash	4.4517	5.1462	1.0977	1.3941	4.6493	5.3697	2.0380	2.8660
GPT-4o-mini	4.0891	4.6469	1.9234	2.1657	3.9590	4.5367	**1.0770**	**1.5466**
GPT-5-mini	2.4600	3.0067	0.9086	**1.0845**	2.6693	3.1717	3.0500	3.3324
GPT-5	2.8482	3.2180	0.9126	1.1722	3.0859	3.7891	1.5530	1.7762
GPT-5-web	2.5851	2.8872	**0.8886**	1.1016	3.0172	3.6741	4.1640	4.3954
GPT-5.1	4.5651	5.1041	1.5174	1.7251	4.0810	4.7041	1.9830	2.2630
Gemini-2.5-Pro	4.8717	5.7561	1.9129	2.4405	4.1486	4.9068	3.1723	4.7006
Gemini-3-Pro	4.0097	4.4446	1.2997	1.4568	3.8729	4.3779	3.1807	4.4384
DeepSeek-V3.2	3.5914	4.1162	1.3285	1.6747	4.1000	4.7346	1.9467	2.4835
GPT-o3-mini	**2.2826**	**2.8561**	1.6263	1.8525	**1.2259**	**1.6785**	1.5093	1.6869
GPT-o3	3.2583	3.7656	1.2737	1.5457	4.0952	4.7052	1.3927	1.6506

Abbrev.: CZU = CZU Aug. Lightning; ELDO = El Dorado; LNU = LNU Lightning; SCU = SCU Lightning.

Table 4: Daily personnel prediction results (MAE/RMSE) on four held-out wildfires.

Models	CZU		ELDO		LNU		SCU	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Text-RAG (no GAL)	1.2283	1.3927	1.1878	1.6621	2.5054	3.1979	1.4997	1.8252
DLinear	1.9814	2.2738	1.3117	1.9195	3.7606	4.4340	2.4192	2.8928
TimesNet	1.9797	2.2729	1.3114	1.9198	3.7585	4.4326	2.4198	2.8917
PatchTST	1.9676	2.2618	1.2988	1.9100	3.7502	4.4223	2.4097	2.8812
TimeLLM	0.9768	**1.2018**	1.0556	1.5233	2.7634	3.2914	1.8859	2.0891
Gemini-2.5-Flash	1.4029	1.7572	0.7971	1.3868	2.7750	3.5688	1.6066	1.9672
GPT-4o-mini	1.3520	1.6307	0.8006	1.4770	2.9398	3.7167	1.2701	1.5829
GPT-5-mini	1.0229	1.3740	0.7973	1.3633	2.5250	3.1504	0.9999	1.3888
GPT-5	1.0230	1.3284	0.8072	1.4238	2.6781	3.3963	1.2822	1.5542
GPT-5-web	1.0185	1.2800	**0.7783**	1.3844	2.6985	3.4130	**0.8963**	**1.1402**
GPT-5.1	1.3679	1.6940	0.8313	1.5050	2.9297	3.7163	1.8093	2.1514
Gemini-2.5-Pro	1.5842	2.1840	1.3254	2.0212	2.3426	3.3043	2.0097	2.5964
Gemini-3-Pro	1.3816	1.7301	0.8267	1.4926	**2.2283**	**3.0762**	1.3474	1.5935
DeepSeek-V3.2	1.2644	1.5260	0.8518	1.4380	2.7602	3.5321	1.4208	1.6561
GPT-o3-mini	**0.9530**	1.2599	0.8418	**1.3535**	2.7487	3.3869	1.6596	1.9789
GPT-o3	1.2494	1.5501	0.8804	1.5253	2.9812	3.7631	1.7996	2.1649

Abbrev.: CZU = CZU Aug. Lightning; ELDO = El Dorado; LNU = LNU Lightning; SCU = SCU Lightning.

Table 5: Daily cost prediction results (MAE/RMSE) on four held-out wildfires.

Model	CZU		ELDO		LNU		SCU	
	Tokens	Cost ($)	Tokens	Cost ($)	Tokens	Cost ($)	Tokens	Cost ($)
GPT-4o-mini	74 244	0.02	76 087	0.02	72 405	0.02	80 827	0.02
GPT-4o	74 099	0.27	75 350	0.28	72 319	0.25	80 974	0.28
GPT-4.1-mini	77 512	0.05	79 295	0.05	75 064	0.04	84 225	0.05
GPT-4.1	78 276	0.25	79 754	0.25	75 445	0.22	85 020	0.25
GPT-5-mini	120 626	0.13	122 823	0.13	110 436	0.11	122 551	0.12
GPT-5	152 329	0.97	153 593	0.97	134 258	0.79	152 889	0.91
GPT-5.1	80 770	0.25	82 097	0.25	77 264	0.22	86 662	0.24
GPT-o3-mini	137 115	0.40	139 130	0.40	129 254	0.36	137 646	0.37
GPT-o3	97 130	0.40	98 463	0.41	89 818	0.34	101 459	0.39
Gemini-2.5-Flash	75 903	0.06	76 725	0.06	71 372	0.05	75 656	0.06
Gemini-2.5-Pro	74 807	0.24	76 364	0.25	70 005	0.20	77 538	0.22
Gemini-3-Pro	75 192	0.68	75 911	0.68	70 859	0.64	77 861	0.70
DeepSeek-V3.2	164 132	0.06	165 856	0.06	146 846	0.05	157 016	0.06

Abbrev.: CZU = CZU Aug. Lightning; ELDO = El Dorado; LNU = LNU Lightning; SCU = SCU Lightning.

Table 6: Average token usage and API cost ($) per fire per model on four held-out wildfires.