Generating Synthetic Intelligence Reports with LLMs via Structured and Recursive Prompting

Anonymous submission

Abstract

We propose a structured framework for generating and evaluating synthetic intelligence reports using large language models (LLMs), specifically GPT-3.5 and GPT-4. Our approach integrates recursive prompting with symbolic and spatial grounding via knowledge graphs and map metadata to produce multi-perspective, JSON-formatted reports that emulate real-world intelligence workflows. To assess quality, we conduct a human evaluation using a rubric based on five analytic dimensions: clarity, objectivity, comprehensiveness, rigor, and relevance. Results show that GPT-4 produces more coherent and reliable outputs, while GPT-3.5, when scaffolded with structured input, performs competitively in analytical depth and relevance. Our framework extends prior LLM benchmarks by targeting long-form synthesis and structured reasoning in complex, mission-oriented domains.

1 Introduction

Advances in large language models (LLMs) have enabled new capabilities for simulating complex analytical workflows through synthetic text generation. In high-stakes domains such as intelligence analysis, access to real-world data is often restricted due to classification and privacy concerns Falis et al. (2024); Yu et al. (2024). As a result, synthetic datasets offer a promising alternative for training, evaluating, and benchmarking LLMs in structured reasoning tasks.

This paper presents a structured generation and evaluation framework for producing synthetic intelligence reports using LLMs, specifically GPT-3.5 and GPT-4. Our approach integrates multiple components: scenario-driven prompting to define fictional yet coherent geopolitical narratives; symbolic grounding through knowledge graphs; spatial grounding via procedurally generated map metadata; and recursive prompting to simulate multi-turn analytical workflows. Each report includes structured metadata, follows a controlled prompt format, and optionally embeds tables for improved information retrieval. To assess report quality, we implement a rubric-based human evaluation strategy focused on five key criteria relevant to intelligence work: clarity, objectivity, comprehensiveness, rigor, and relevance.

To illustrate these capabilities, Table 1 presents a sample report generated by our framework. It documents battlefield injuries sustained by Mamba Force Opal during the fictional Battle of Salazar Marshlands. The report includes structured metadata such as the author persona, classification label, timestamp, and geospatial coordinates, while the body summarizes injury types (e.g., gunshots, IED trauma) using domain-relevant language. Feedback from fictional readers and distribution to fictional government entities further anchor the report within its symbolic scenario. At the end of the report, a follow-up question is generated, as shown in Table 1: “Can we get more specific information on the types of injuries sustained by our forces?” This question serves as a seed for recursive prompting, allowing the model to explore subtopics such as medical diagnostics or tactical procedures, thereby enabling iterative and inquiry-driven expansion of the scenario.

Our framework addresses limitations in existing LLM benchmarks. While datasets such as TruthfulQA Lin et al. (2021), WebGPT Nakano et al. (2022), and LegalBench Guha et al. (2023) emphasize factual correctness or discrete task performance, they do not evaluate long-form, context-rich synthesis grounded in structured symbolic and spatial input. We define long-form as outputs that combine narrative explanation, analytical reasoning, metadata alignment, and stylistic realism. Our framework fills this gap by incorporating scenario construction, symbolic and spatial grounding, recursive prompting, structured prompt formatting, and metadata-aware generation. Additional design elements including fictional personas, geocoordinates, classification labels, table generation, and source integration further align synthetic reports with real-world intelligence conventions. This paper investigates the following research questions:

- Can LLMs generate synthetic intelligence reports that resemble real documents in tone, structure, and informational depth?
- What is the impact of recursive prompting on the coherence, continuity, and scalability of report synthesis?
- How do structured inputs such as knowledge graphs and geographic metadata affect report realism and analytical value?
- How do GPT-3.5 and GPT-4 differ in schema adherence, entity grounding, and output robustness within this framework?

Table 1: Example report generated by our framework, combining medical intelligence, narrative detail, and recursive prompting via follow-up questions.

Title: *Injury Report: Mamba Force Opal in the Battle of Salazar Marshlands* **Timestamp:** 2023-06-17T10:30:00 **Event Date:** 2023-06-14 **Author:** Patricia “Pat” Wilson **Readers:** Charlie “Champ” Thompson, Dr. Richard Thornton **Feedback Score:** 4.5 **Feedback Comments:** Charlie "Champ" Thompson: 4 – "Detailed and comprehensive report." Dr. Richard Thornton: 5 – "Excellent use of medical documents." **Source Type:** Medical Intelligence **Classification:** CONFIDENTIAL CHIPMUNK **Geo Coordinates:** Salazar Marshlands (Latitude: -15.2, Longitude: -177.7) **Tags:** Operation Koila, Salazar Marshlands, Mamba Force Opal, Injury Report **Distribution:** SCADSLand Ministry of Defense, National Security Council, Ministry of Health
Body: This report summarizes injuries sustained by Mamba Force Opal during the Battle of Salazar Marshlands. Of 42 wounded, 25 suffered gunshot wounds, 10 sustained shrapnel injuries from IEDs, and 7 were injured by other causes including burns and blunt trauma. While body armor reduced the severity of most gunshots, the IED-related injuries were critical, including limb loss and severe burns. Rapid medical response limited fatalities, but the findings highlight a need for improved protective gear, IED training, and specialized burn treatment capabilities. **Follow-Up Question:** *Can we get more specific information on the types of injuries sustained by our forces?*

Through these research questions, we aim to establish both a robust generation pipeline and a principled evaluation framework for synthetic intelligence reporting. Our findings show that LLMs guided by structured input produce higher-quality outputs and that GPT-3.5, when properly scaffolded, can approach the performance of GPT-4 in structured, long-form reporting tasks.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 presents our methodology. Section 4 introduces our evaluation rubric, describes the human evaluation setup, and reports results comparing GPT-3.5 and GPT-4. Section 5 situates our approach within existing LLM benchmarks. Section 6 concludes the paper. Section 7 discusses limitations, while Section 8 covers ethical considerations.

2 Related Work

Large language models (LLMs) such as GPT-3.5 and GPT-4 Achiam et al. (2023) have demonstrated strong capabilities in generating coherent and contextually grounded text. These capabilities have enabled a wide range of applications in synthetic data generation, particularly in domains where real data is scarce or sensitive Nguyen-Mau et al. (2024). Initial studies have applied LLMs to domain-specific generation tasks including question answering Kalpakchi and Boye (2023), dialogue simulation Abdullin et al. (2024), and stylized text production Popescu-Belis et al. (2023). LLMs have also been explored in medical and cybersecurity domains, where they are used to generate discharge summaries Falis et al. (2024), simulate longitudinal health records Pang et al. (2024), model synthetic traffic data Kholgh and Kostakos (2023), and augment biomedical signals Bird et al. (2021).

Recent work has shifted attention toward long-form generation, with an emphasis on planning, structure, and multi-stage synthesis. Liang et al. (2024) introduced a planning-based framework that guides LLMs to organize and reason before generating full-length documents. Zhu et al. (2024) proposed a segment-level diffusion approach to improve coherence and controllability in long-form generation. Tan et al. (2024) developed ProxyQA, a benchmark that evaluates long-form outputs by measuring human evaluator accuracy on proxy questions derived from model-generated responses.

In parallel, techniques for symbolic and structured prompting have been developed to improve the reasoning capabilities of LLMs. Cognitive prompting has been shown to encourage structured thinking and stepwise reasoning in complex problem-solving tasks Kramer and Baumann (2024). Wang et al. (2024) proposed a symbolic working memory mechanism that enhances LLM performance on rule-based applications. Challenging traditional assumptions, Gubelmann (2024) argue that LLMs can achieve grounded understanding through pragmatic interaction norms rather than explicit symbol-referent mappings.

As generation quality improves, evaluation strategies have evolved to reflect the needs of high-stakes, long-form generation tasks. Hashemi et al. (2024) introduced LLM-Rubric, a calibrated framework for scoring generated outputs across multiple qualitative dimensions. Rubicon Biyani et al. (2024) applies rubric-based evaluation to domain-specific human–AI conversations, combining human and automated scoring. Similarly, Farzi and Dietz (2024) proposed a rubric-driven evaluation pipeline for retrieve-and-generate systems, offering a structured approach to assessing both content and context alignment.

Despite these advancements, most prior work remains focused on short-form generation or narrow task-specific outputs. Long-form generation that incorporates recursive prompting, symbolic grounding, and spatial metadata remains underexplored, particularly in domains such as intelligence analysis that demand analytical depth, narrative continuity, and structured metadata integration.

Our work addresses this gap by introducing a framework for long-form synthetic intelligence report generation. We combine recursive prompting with structured inputs such as knowledge graphs and map metadata, and evaluate outputs using a rubric-based human assessment aligned with the needs of mission-critical analysis. This contributes to ongoing efforts to expand the scope of LLM evaluation beyond factuality and classification into the realm of context-rich, multi-perspective synthesis.

3 Methodology

We present a structured methodology for generating synthetic intelligence reports using large language models (LLMs), with a focus on GPT-3.5 and GPT-4. Our framework simulates analyst workflows through three tightly integrated components: (1) scenario construction, which provides a coherent fictional backdrop for report generation; (2) symbolic and spatial grounding, which embeds entities and geocoordinates through knowledge graphs and map metadata; and (3) recursive prompting, which produces multi-layered reports by chaining outputs through follow-up questions. The generation process is further enriched through structured metadata design, prompt formatting, and controlled variability. Figure 1 illustrates the complete generation pipeline, from scenario setup and metadata initialization to recursive report generation and rubric-based evaluation. These components work together to create long-form, multi-perspective reports that reflect the analytical rigor and continuity found in real intelligence workflows.

3.1 Scenario Construction

To establish a coherent fictional setting for report generation, we adapted a passage from *Traffic Analysis and the Zendian Problem* Callimahos (1989), modifying names, affiliations, and events while preserving the structural logic of the original narrative. This scenario, used as an illustrative example throughout the paper, features fictional characters, military factions, government ministries, and geopolitical flashpoints within the nation of Zendia. Care was taken to ensure that all fictional entities were unique and did not correspond to real individuals or organizations. This structured backdrop enables the generation of multi-layered intelligence reports that simulate realistic analytical workflows.

3.2 Symbolic Grounding

To simulate the structured reasoning typical of intelligence reporting, we constructed a symbolic knowledge graph consisting of entities such as individuals, organizations, events, and locations. These entities were created using prompt templates with predefined attributes (e.g., geopolitical affiliation, type, and quantity). Relationships between entities were likewise generated and normalized to reduce redundancy and maintain coherence.

This knowledge graph served as a grounding tool during report generation. Entity and relationship data were embedded in prompts to guide model behavior. In practice, GPT-4 demonstrated a higher rate of entity incorporation and relational consistency compared to GPT-3.5, which often omitted or altered elements unless explicitly reinforced. Additional examples and statistics are included in Appendix B.

3.3 Spatial Grounding

To complement symbolic grounding, we integrated procedurally generated map data, including landforms, boundaries, and geospatial references.

These were manually aligned with the knowledge graph and embedded into the model input in a structured format. This approach improved the spatial consistency of generated content, enabling more realistic references to terrain, regions, and distances. A visual example of the map configuration is provided in Section 3.8.3.

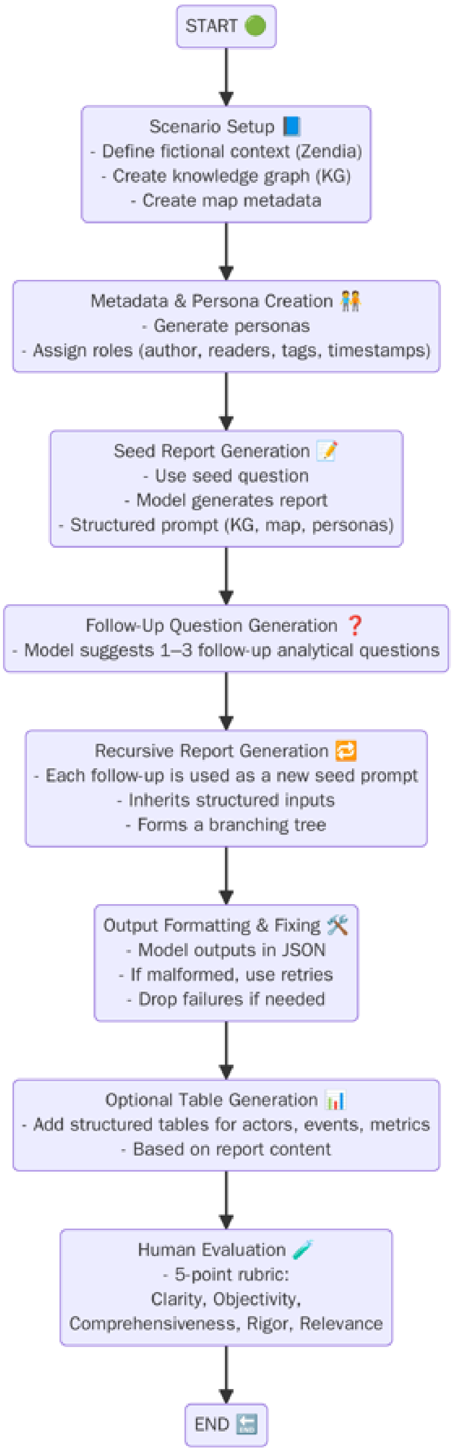


Figure 1: Overview of our synthetic intelligence report generation pipeline, combining scenario setup, symbolic and spatial grounding, recursive prompting, and rubric based evaluation.

3.4 Token Optimization and Generation Trade-offs

Structured inputs such as maps and knowledge graphs significantly enhanced the coherence of the generated reports but also increased token consumption. To maintain scalability, we separated background information from report prompts and summarized it before injection. GPT-4 yielded the most coherent results under full structure but incurred higher generation costs. GPT-3.5 was more cost-efficient and still benefited considerably from structured inputs. Detailed comparisons of token usage and model cost are included in Appendix A.

3.5 Recursive Prompting

Our report generation framework builds on recursive prompting to simulate how intelligence analysts refine assessments over time. Given a structured seed scenario, each report leads to follow-up questions that drive the next layer of generation, resulting in a tree-like structure of interconnected reports. This recursive process encourages multi-perspective reasoning, temporal continuity, and analytical depth across generations.

3.5.1 Generation Strategy

The system initiates each report using a two-prompt cycle. First, the model is asked to generate an intelligence report in response to a guiding question. Second, it is prompted to suggest follow-up questions a government analyst might pose after reviewing the report. These follow-up questions are recursively used to generate additional reports, forming a hierarchical generation tree. Figure 2 visualizes the recursive prompting flow, where each report leads to a new layer of analytical follow-up. Each layer expands the narrative by exploring new analytical paths grounded in prior context.

3.5.2 Controlling Report Scope and Volume

The framework supports several generation parameters: the number of top-level seed questions (**S**), the number of follow-up prompts per report (**X**), the number of recursive layers (**Y**), and a regeneration factor (**R**) that introduces variation by generating multiple outputs per node. Together, these parameters define the breadth and depth of the reporting tree and enable flexible control over the volume and structure of the dataset. An illustrative generation tree is provided in Appendix C.

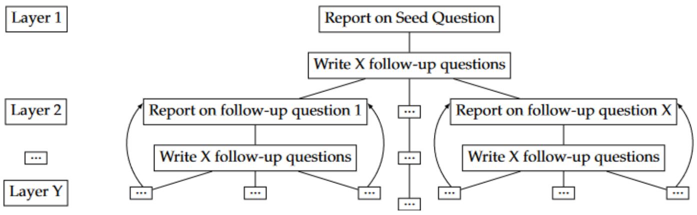


Figure 2: recursive prompting structure for a single seed question with X children and Y layers.

3.5.3 Scalability and Parallelization

To enable large-scale generation, we use a breadth-first strategy that processes all nodes at each layer in parallel. This improves efficiency, while the regeneration factor adds further parallelism by creating multiple report variants per node.

3.5.4 Simulating Variability

To better reflect the heterogeneity of real intelligence reporting, we introduce controlled variability into the generation process. Drawing from an analysis of 1,000 real-world reports, we designed prompts to include stochastic constraints on length, structure, and metadata. Rather than relying on vague directives (e.g., "make it random"), we use Python-generated parameters embedded into the prompt (e.g., "length around 600 tokens"). This approach yields more natural length distributions across characters, words, and sentences. Additional randomness was introduced across metadata fields to support diversity and reduce repetitiveness.

3.5.5 Structured Output Format (JSON)

To facilitate downstream analysis and reuse, we implemented two output strategies. The first parses plain text reports into structured JSON via post-processing. The second instructs the model to generate reports directly in JSON format using an explicit schema. While the latter simplifies the pipeline, schema compliance is not guaranteed. Minor inconsistencies (e.g., variation in field names) required a validation and correction loop. When invalid output was detected, corrective prompts were issued until a valid structure was obtained. Most errors were resolved within two iterations; persistent failures were discarded to preserve dataset quality.

3.6 Prompt Structure

Prompt design not only guides the model's narrative but also ensures alignment with formatting constraints, enabling seamless downstream processing and rubric-based evaluation. Prompt structure plays a central role in guiding large language models toward generating consistent, grounded, and analytically useful reports. Our framework relies on structured multi-turn prompting to simulate the behavior of intelligence analysts, with distinct system, user, and assistant roles to maintain continuity across recursive generations.

Each generation session begins with a system message that establishes the model's persona and operational context. For instance, the model is positioned as an analyst working for the fictional government of SCADSLand and tasked with monitoring geopolitical developments in the nation of Zendia. This message includes the base scenario as well as structured metadata (e.g., knowledge graphs, personas, and maps) prepared during the pre-processing phase. The prompting workflow proceeds in five stages:

1. Pre-generation The system message defines the analytical setting, while the user initiates metadata generation requests for elements such as personas, knowledge graphs, and spatial data.

2. Seed Report Generation The user then instructs the model to generate an intelligence report based on a seed question $S[i]$, while also specifying output constraints (e.g., JSON format, metadata schema). If the model fails to comply with formatting rules, corrective prompts are issued as described in Section 3.5.5.

3. Follow-up Question Generation After each report, the model is asked to list follow-up questions a government analyst might ask. To preserve the analytical tone, the prompt explicitly discourages action-oriented recommendations and focuses on eliciting requests for additional information.

4. Recursive Prompting Each follow-up question becomes a new seed for generating the next layer of reports. This step inherits structured metadata and formatting constraints from previous turns, allowing for depth and continuity in analysis.

5. Iteration Steps 3 and 4 are repeated recursively to construct a reporting tree. This recursive structure enables multi-layered synthesis that mimics how real-world analysts build on prior insights to explore emerging developments.

This structured prompting strategy ensures consistency in report tone, metadata schema, and analytical framing while allowing sufficient variability to reflect the dynamic nature of intelligence work. An example of a complete multi-turn generation session is provided in Appendix I.

3.7 Table Generation

To improve clarity and support information retrieval, we added structured tables to a subset of the generated reports. These tables act as visual summaries, highlighting key actors, events, numerical metrics, and technical specifications that are often essential in intelligence communication. However, generating coherent and useful tables with large language models introduced unique challenges.

Initial attempts relied on general instructions (e.g., "include a table") applied with low probability. This produced inconsistent results, with many tables lacking structure, context, or analytical value. To address this, we designed a more targeted prompting scheme based on four canonical categories observed in intelligence reports:

- **Key Actors**: Roles, affiliations, and actions of individuals or organizations involved.
- **Main Events**: Timeline, participants, and outcomes of critical incidents.
- **Quantitative Measures**: Numerical data such as quantities, rates, costs, or distances.
- **Technical Specifications**: Structured information on equipment, geography, or processes.

Each selected report received one of these category-specific prompts. This focused strategy improved the model's ability to generate coherent and relevant tables by anchoring the request to a specific structural pattern. While some reports lacked sufficient narrative grounding for complete table generation, overall informativeness and consistency increased.

To maintain report quality, we disabled table prompts for short outputs (fewer than 110 tokens), which tended to produce low-value or filler content. A combined example of the four table types is provided in Appendix D.

3.8 Metadata Design

A key advantage of synthetic report generation is the ability to control metadata attributes with precision. In our framework, metadata fields were defined in collaboration with intelligence analysts and data scientists to reflect real-world reporting conventions while supporting systematic evaluation and downstream analysis.

3.8.1 Schema Overview

The metadata schema is structured into two categories: fields with tightly constrained values (e.g., author, topic, geo-coordinates) and fields with looser or randomized generation. This balance preserves scenario coherence while allowing variability across reports.

Core fields include the report title, body, timestamps, and event dates, along with structured elements such as classification labels, keywords (tags), and intelligence source types (e.g., HUMINT, SIGINT, OSINT). Some metadata is derived externally, such as unique serial identifiers and the original prompting question, which links the report back to the recursive generation tree. A complete schema and JSON example are provided Appendix I. This structured design supports both retrieval and evaluation workflows by aligning each report with its context and intended audience.

3.8.2 Personas

To simulate the diversity of authorship and readership found in intelligence organizations, we introduced fictional analyst personas. GPT was prompted to generate ten unique personas at the start of each scenario, which were embedded into the prompt with the instruction: "Your team consists of the following personas."

Each report includes a designated author, typically selected based on topical relevance. To avoid stylistic contamination, models were instructed not to reference author names in the body text. Readers were randomly selected from the same persona list and asked to provide a numerical rating and optional feedback comment. While occasional anomalies occurred (e.g., readers rating their own reports), the persona mechanism added realism and evaluation flexibility. Examples of feedback patterns are shown in Appendix E.

3.8.3 Geographic Metadata

Geospatial grounding was supported through a curated list of map entities provided in CSV format during the scenario setup. These entities were used to assign geographic coordinates to events and locations within each report. GPT-4 handled spatial consistency more reliably than GPT-3.5, which occasionally produced malformed coordinates. Manual correction was applied where necessary. A comparison of spatial output across models is provided in Appendix F.

3.8.4 Source Integration

To explore the potential of retrieval-augmented generation, we experimented with injecting external source material into the prompt. Using LangChain's RetrievalQA pipeline[1], we inserted source-aligned context into GPT sessions, similar to PrivateGPT[2].

Two datasets were generated using this method. The ODIN dataset was based on military training materials and yielded coherent reports in tactical domains like illicit mining. The VAST dataset used documents from the 2014 VAST challenge Andrews and Crouser (2019), providing rich narrative grounding despite some source noise. Together, these experiments produced over 100 structured reports and demonstrate that selective source integration can improve realism in low-cost synthetic generation pipelines.

4 Evaluation

To assess the quality and analytical value of generated intelligence reports, we conducted a human evaluation involving 40 samples drawn from our recursive prompting pipeline. Four PhD-level native English speakers with experience in technical writing and research served as evaluators. Each report was assessed across five rubric-based dimensions aligned with intelligence analysis best practices, as summarized in Table 2.

Table 2: Evaluation rubric for intelligence reports.

Dimension	Evaluation Question
Clarity	Is the report easy to read and understand?
Objectivity	Is the analysis fact based and unbiased?
Comprehensiveness	Does the report cover all key details?
Rigor	Is the analysis thorough and verifiable?
Relevance	Is the content useful for intelligence work?

Each dimension was rated on a five-point Likert scale. Evaluators were blinded to both the model used (GPT-3.5 or GPT-4) and the generation configuration. Reports were evenly distributed across four conditions: (1) baseline model, (2) model with knowledge graphs (KG), (3) model with map-based metadata, and (4) model with both KG and map support. Prior to evaluation, annotators were given detailed calibration guidelines and example ratings to promote scoring consistency.

Table 3: Average scores by generation setting.

Method	Clarity	Obj.	Comp.	Rigor	Rel.
GPT-3.5	4.50	4.50	3.75	3.50	4.00
+ KG	4.75	4.50	3.75	4.25	4.25
+ Map	4.75	4.75	4.25	4.25	4.25
+ KG + Map	4.75	5.00	4.50	4.50	4.75
GPT-4	4.75	4.75	3.75	3.25	4.75
+ KG	4.75	4.75	3.25	3.50	4.50
+ Map	4.75	4.75	3.50	3.75	4.75
+ KG + Map	4.75	4.75	4.00	4.25	4.50

4.1 Results

Table 3 present the average scores across all evaluated dimensions. GPT-4 consistently scored highly for clarity, objectivity, and relevance. However, GPT 3.5, when augmented with both knowledge graphs and map metadata, achieved the highest scores overall, particularly in *comprehensiveness* (4.50) and *relevance* (4.75).

These results underscore the value of structured input design. While GPT-4's baseline outputs were more fluent and stylistically refined, they lagged behind GPT-3.5+KG+Map in rigor and coverage. This suggests that generation strategy and prompt structure can significantly shape the quality of LLM outputs, often outweighing the impact of model scale alone.

The evaluation confirms that model scale alone is not a reliable predictor of intelligence report quality. Structured prompts and metadata play a decisive role in enabling smaller models like GPT-3.5 to match or even surpass GPT-4 in dimensions such as comprehensiveness and rigor. These findings validate the core design of our framework, emphasizing the importance of symbolic grounding, spatial context, and recursive prompting for generating high-utility synthetic reports.

5 Comparison with Prior Work

While large language models (LLMs) have been widely benchmarked on factuality (Lin et al., 2021), multitask QA (Hendrycks et al., 2020), and step-

[1] https://python.langchain.com/v0.1/docs/modules/model_io/

[2] https://github.com/zylon-ai/private-gpt

Table 4: Comparison of structured generation and evaluation methods across recent LLM benchmarks.

Study / Benchmark	Domain	Evaluation Type	Context Type	Structured Input?
TruthfulQA (Lin et al., 2021)	Factuality	Human scoring	Adversarial questions	No
MMLU (Hendrycks et al., 2020)	Multitask QA	Accuracy (MCQ)	Diverse academic tasks	No
Least-to-Most Prompting (Zhou et al., 2022)	Reasoning	Prompt chaining	Stepwise decomposition	Yes
WebGPT (Nakano et al., 2022)	QA with retrieval	Human preference	Retrieved documents	No
LegalBench (Guha et al., 2023)	Legal reasoning	Task level accuracy	Legal domain context	Yes
SciBench (Wang et al., 2023)	Scientific problem solving	Expert and benchmark eval	Equations and diagrams	Yes
ProxyQA (Tan et al., 2024)	Long form QA	Human answerability	Single turn generation	No
LLM-Rubric (Hashemi and et al., 2024)	General NLG evaluation	Rubric based scoring	Mixed open ended text	No
PlanningGen (Liang and et al., 2024)	Document generation	Human and structure eval	Planning based prompts	Yes
CAPE-FND (Jin et al., 2025)	Fake news detection	Human and automatic eval	Context aware claims	Yes
GPT-4 Tech Report (Achiam et al., 2023)	General benchmarks	Task scores	Prompt only	No
NEET QA (Farhat et al., 2024)	Medical exam QA	Accuracy (MCQ)	Structured exam items	No
PEEM QA (Law et al., 2025)	Emergency medicine QA	Psychometric analysis	Structured questions	No
Ours (this paper)	Intelligence reporting	Rubric based human eval	KG + map metadata + recursive prompting	**Yes**

wise reasoning (Zhou et al., 2022), most evaluations remain limited to closed-ended tasks or isolated outputs. These approaches do not address the core challenges of generating structured, multi-perspective, and recursively constructed narratives in high-stakes analytic domains. Recent advances such as WebGPT (Nakano et al., 2022) and CAPE-FND (Jin et al., 2025) incorporate retrieval or contextual prompting but are confined to single-pass outputs and do not support continuity across generations. Our framework goes significantly beyond these efforts by introducing recursive prompting combined with symbolic knowledge graph and spatial map based grounding, enabling dynamic multi turn synthesis that better reflects real world intelligence workflows.

Domain-specific evaluations such as LegalBench (Guha et al., 2023), SciBench (Wang et al., 2023), and ProxyQA (Tan et al., 2024) provide structured contexts, but they are fundamentally designed around single-document or task-level performance. They lack support for metadata-aware report generation, longitudinal reasoning, or recursive question-follow-up pipelines.

Notably, while LLM-Rubric (Hashemi and et al., 2024) and PlanningGen (Liang and et al., 2024) introduce rubric based or staged planning evaluation, neither incorporates domain specific metadata nor supports branching generation across multi layered reporting trees. Our approach combines these strengths and applies them to a more demanding context such as intelligence reporting, where narrative fidelity, analytic depth, and metadata alignment are mission critical.

Our human evaluation further demonstrates that GPT 3.5, when guided with structured inputs, can match or outperform GPT 4 in dimensions such as clarity and relevance. This result not only challenges assumptions about model scale but also highlights the centrality of prompt structure and symbolic grounding in producing high quality analytical outputs. Table 4 highlights the unique combination of capabilities our framework offers including recursive generation, symbolic and spatial inputs, structured formatting, and rubric based evaluation, which are not jointly supported by any prior benchmark to date.

6 Conclusion

We presented a structured framework for generating synthetic intelligence reports using GPT 3.5 and GPT 4, combining recursive prompting with symbolic and spatial grounding to simulate real world analytical workflows. Our results demonstrate that this approach enables the generation of coherent, context rich reports at scale. Human evaluation reveals that structured inputs such as knowledge graphs and map metadata significantly enhance report quality across clarity, relevance, and rigor. While GPT 4 consistently produces more reliable and schema compliant outputs, GPT 3.5 performs competitively when scaffolded with structured input, offering a cost effective alternative without compromising analytical value. These findings highlight the critical role of input design and contextual grounding, not just model size, in achieving high quality generation.

Our rubric based evaluation addresses limitations in existing LLM benchmarks by focusing on long form synthesis, recursive reasoning, and metadata fidelity. The proposed framework lays the groundwork for a new class of synthetic datasets tailored to mission oriented domains. Future work will explore extensions to domains such as crisis response, scientific reporting, and policy analysis, with emphasis on controllability, validation, and integration with structured reasoning pipelines.

7 Limitations

This work is based on a fictional geopolitical scenario, which may limit the generalizability of findings to real-world intelligence settings. Although human evaluators followed structured guidelines, rubric-based assessments are inherently subjective and may reflect individual judgment. The behavior of GPT-3.5 and GPT-4 is also sensitive to prompt phrasing and may vary across model versions. While our results highlight the value of structured input, additional evaluation on real-world tasks is needed to assess downstream utility and transferability.

8 Ethical Considerations

All synthetic reports in this study were generated using fictional scenarios, entities, and locations designed to avoid overlap with real-world content. We verified that names, organizations, and geographic references did not correspond to actual individuals or institutions using public search engines and reference datasets. No real intelligence data or classified material was used at any stage. The generated reports are intended solely for research and do not simulate or represent actual government communications. The primary goal of this work is to advance structured generation and evaluation methods for long-form tasks, not to enable operational deployment. Human evaluations were conducted by voluntary, anonymized participants with relevant academic backgrounds. No personally identifiable information was collected, and participants were fully informed about the synthetic nature of the material. Compensation was provided in accordance with institutional policies.

We acknowledge the dual-use potential of language models in sensitive domains such as intelligence and surveillance. We recommend that any future applications of this framework follow clear ethical guidelines and prioritize transparency, auditability, and human oversight to ensure responsible use. We also used AI writing assistants (e.g., ChatGPT) solely for grammar correction and improving the clarity of exposition, without affecting the substance of the research.

References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *ACL Anthology*.

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

Christopher P Andrews and R Jordan Crouser. 2019. Using the vast challenge in undergraduate cs research. *Consortium for Computing Sciences in Colleges*.

Jordan J Bird, Michael Pritchard, Antonio Fratini, Anikó Ekárt, and Diego R Faria. 2021. Synthetic biological signals machine-generated by gpt-2 improve the classification of eeg and emg through data augmentation. *IEEE Robotics and Automation Letters*, 6(2):3498–3504.

Param Biyani, Yasharth Bajpai, Arjun Radhakrishna, Gustavo Soares, and Sumit Gulwani. 2024. Rubicon: Rubric-based evaluation of domain-specific human ai conversations. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, pages 161–169.

Lambros D Callimahos. 1989. Traffic analysis and the zendian problem: an exercise in communications intelligence operations. *(No Title)*.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can gpt-3.5 generate and code discharge summaries? *arXiv preprint arXiv:2401.13512*.

Faiza Farhat, Beenish Moalla Chaudhry, Mohammad Nadeem, Shahab Saquib Sohail, and Dag Øivind Madsen. 2024. Evaluating large language models for the national premedical exam in india: comparative analysis of gpt-3.5, gpt-4, and bard. *JMIR Medical Education*, 10:e51523.

Naghmeh Farzi and Laura Dietz. 2024. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 175–184.

Reto Gubelmann. 2024. Pragmatic norms are all you need–why the symbol grounding problem does not apply to llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*.

Helia Hashemi and et al. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Weiqiang Jin, Yang Gao, Tao Tao, Xiujun Wang, Ningwei Wang, Baohai Wu, and Biao Zhao. 2025. Veracity-oriented context-aware large language models–based prompting optimization for fake news detection. *International Journal of Intelligent Systems*, 2025(1):5920142.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in swedish using gpt-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491.

Danial Khosh Kholgh and Panos Kostakos. 2023. Pac-gpt: A novel approach to generating synthetic network traffic with gpt-3. *IEEE Access*.

Oliver Kramer and Jill Baumann. 2024. Unlocking structured thinking in language models with cognitive prompting. *arXiv preprint arXiv:2410.02953*.

Alex KK Law, Jerome So, Chun Tat Lui, Yu Fai Choi, Koon Ho Cheung, Kevin Kei-ching Hung, and Colin Alexander Graham. 2025. Ai versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, 25(1):208.

Yi Liang and et al. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

Yi Liang, You Wu, Honglei Zhuang, Li Chen, Jiaming Shen, Yiling Jia, Zhen Qin, Sumit Sanghai, Xuanhui Wang, Carl Yang, et al. 2024. Integrating planning into single-turn long-form text generation. *arXiv preprint arXiv:2410.06203*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2022. Webgpt: Browser-assisted question-answering with human feedback, 2022. *URL https://arxiv. org/abs/2112.09332*.

Toan Nguyen-Mau, Anh-Cuong Le, Duc-Hong Pham, and Van-Nam Huynh. 2024. An information fusion based approach to context-based fine-tuning of gpt models. *Information Fusion*, 104:102202.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Noémie Elhadad, and Karthik Natarajan. 2024. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*.

Andrei Popescu-Belis, Àlex R. Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-Lienhard, and Giorgos Vernikos. 2023. GPoeT: a language model trained for rhyme generation on synthetic data. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–20, Dubrovnik, Croatia. Association for Computational Linguistics.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, et al. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042*.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Xiaochen Zhu, Georgi Karadzhov, Chenxi Whitehouse, and Andreas Vlachos. 2024. Segment-level diffusion: A framework for controllable long-form generation with diffusion language models. *arXiv preprint arXiv:2412.11333*.

A Token Cost Analysis

Table 5 summarizes token usage across all generation configurations used in this study. Each configuration combines different levels of structure (e.g., KG only, Map only, or both) with two LLM variants (GPT-3.5 and GPT-4). Costs were estimated using OpenAI pricing (as of 2024) and reflect total tokens across all recursive layers.

Table 5: Token cost in reports with and without KG and Map data using GPT-4 and GPT-3.5-16k.

Model	Cost (est.)	Scenario Tokens	Tokens/Report	Total Tokens	Reports	Layers	% Inc.
KG + Map GPT-4	$12.57	5,035	11,640	139,683	12	2	262%
KG GPT-4	$7.64	3,115	7,072	84,864	12	2	120%
Map GPT-4	$6.50	2,284	6,022	72,269	12	2	88%
None GPT-4	$3.46	364	3,208	38,499	12	2	–
KG + Map GPT-3.5	$1.36	5,035	12,131	339,681	28	3	112%
KG GPT-3.5	$1.03	3,115	9,159	256,460	28	3	60%
Map GPT-3.5	$1.05	2,284	9,358	262,028	28	3	64%
None GPT-3.5	$0.64	364	5,720	160,186	28	3	–

This comparison highlights the trade-off between quality and efficiency. While GPT-4 is more expensive, its outputs typically require fewer retries and corrections. GPT-3.5 offers a more economical option but benefits significantly from structured prompt augmentation.

B Knowledge Graph Integration Examples

Figures 3 and 4 show examples of GPT-3.5 and GPT-4 integration with structured entity prompts, visualized using knowledge graphs.

```
The occupation force's success in securing key
strongholds, such as the *Crueltonian National Library*
and the *Festival of Iron Resolve*, weakens the regime's
control over cultural and intellectual institutions.
```

Figure 3: Example of GPT-3.5 report generation with a Knowledge Graph; asterisks (*) mark prompt entities.

```
----(Original Knowledge Graph Entry)----
Greater Zendia Mining Corporation,provides resources to,
Zendian Scientific Research  Institute

----(Report section)----
The missile, believed to be a product of the *Greater
Zendia Mining Corporation*, was tracked by our SIGINT
capabilities, which detected the launch and tracked the
missile's trajectory. The missile's launch and flight
characteristics suggest that it is an advanced
model with a significant range and payload capacity.
This development indicates a significant advancement
in the *ZDF*'s missile capabilities, likely supported
by the *Zendian Scientific Research Institute*.
```

Figure 4: GPT-4 report generation using a Knowledge Graph; asterisks (*) denote prompt entities, with inferred relationships such as missile production.

C Recursive Generation Tree

Figure 5 illustrates how recursive prompting with parameter settings $R = 2$, $X = 2$, and $Y = 3$ produces a total of 42 reports from a single seed question. The tree expands in a breadth-first manner with multiple regeneration branches at each node, supporting high-volume synthesis with structural depth.

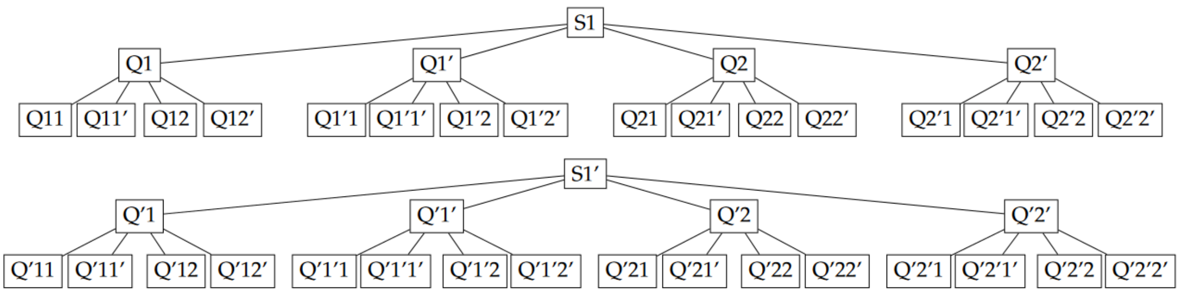


Figure 5: Generation tree with $R = 2$, $X = 2$, and $Y = 3$ producing 42 reports per seed.

D Table Generation Examples

Table 6 shows a unified example of four structured prompt categories used to generate intelligence report tables: Key Actors, Main Events, Quantitative Measures, and Technical Specifications.

Table 6: Combined Examples: Key Actors, Events, Quantitative Measures, and Technical Specifications.

Key Actors			
Actor	**Role**	**Affiliation**	**Actions**
Marshal Ftook Salazar	Dictator	Zendia	Ordered missile launch
Behrainabad Federations	Coalition	Behrainabad	Responded to attack
Main Events			
Event	**Date**	**Involved**	**Outcome**
Amphibious Assault	June 10, 2023	Mamba Force	Established beachhead
Coalition Offensive	June 10–12, 2023	Mamba Force	Gradual enemy push-back
Quantitative Measures			
Measure	**Value**		
Number of incidents	12		
Cost of damages	$5 million		
Technical Specifications			
Satellite	**Position**	**Frequency**	**Data Rate**
ZS-1	45°E	X-band	10 Gbps
ZS-2	60°E	Ku-band	5 Gbps

E Persona Feedback Examples

To illustrate the variability and occasional anomalies in GPT-generated reader feedback, we provide two examples. These were selected from reports in which GPT generated persona-based evaluations. The first shows a critical comment, while the second presents a complete feedback block with diverse reader input.

```
["Agent Wilson", 3.5, "The report could benefit from additional analysis on
↪ potential outcomes and contingency plans."],)
```

(a) Negative comment generated by ChatGPT.

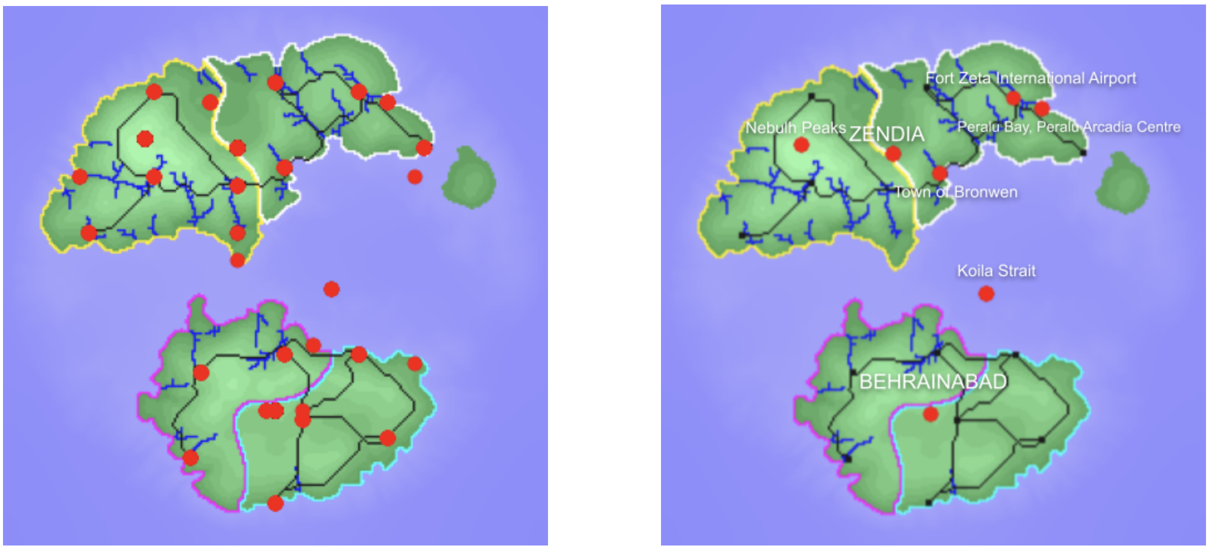
```
"author": Agent Wilson,
"readers": ["Agent Thompson", "Agent Johnson"],
"feedback": 4.8,
"comments": [
    ["Agent Thompson", 4.0, "Excellent analysis of geospatial data and its impact
    ↪ on the battle. Well done!"],
    ["Agent Johnson", 5.0, "Impressive decryption of enemy communications. Your
    ↪ work was crucial in gaining the upper hand."]]
    ["Agent Wilson", 5.0, 'Thank you for your feedback. I appreciate your positive
    ↪ assessment of the report.']]
```

(b) Feedback block presented by ChatGPT.

Figure 6: Examples of reader feedback generated by GPT.

F Geographic Metadata Comparison

This section compares the geographic metadata produced by GPT-3.5 and GPT-4 when prompted with the same map input. Figure 7a visualizes tag usage across 768 reports by GPT-3.5, while Figure 7b shows geocoordinate annotations from a subset of GPT-4 reports. The difference highlights GPT-4's improved spatial awareness and precision in extracting map-based metadata.



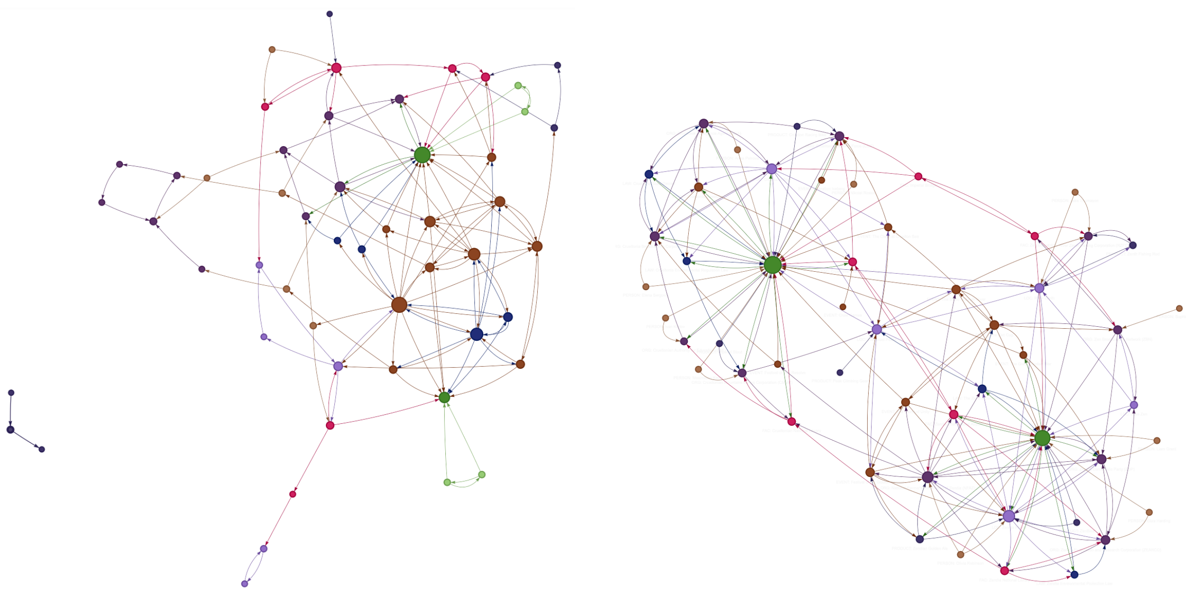
(a) Geographic tags in reports by GPT-3.5.

(b) Geographic coordinates in reports by GPT-4.

Figure 7: Comparison of geographic metadata by GPT-3.5 and GPT-4. Red dots indicate coordinate locations.

G Knowledge Graph Comparison

To evaluate how well GPT models incorporate structured knowledge into report generation, we compared knowledge graphs (KGs) generated by GPT-3.5 and GPT-4 using the same set of prompt entities and relationships. These graphs were extracted from model outputs after processing and normalization steps. GPT-3.5 produced simpler structures with fewer inferred connections. In contrast, GPT-4 generated denser, more coherent graphs that captured complex interdependencies between geopolitical actors, organizations, and events.



(a) Processed knowledge graph generated by GPT-3.5.

(b) Processed knowledge graph generated by GPT-4.

Figure 8: Comparison of knowledge graph outputs. GPT-4 consistently produces richer and more contextually accurate entity networks.

H Persona Comparison

Persona diversity plays a key role in enhancing realism and interpretability in synthetic intelligence reports. We prompted both GPT-3.5 and GPT-4 to generate a set of ten analyst personas for the same scenario. GPT-3.5 focused primarily on occupational roles with minimal variation, while GPT-4 generated more diverse and nuanced characters. The examples below illustrate these differences in style, background detail, and linguistic richness.

```
1. Agent Smith - A seasoned intelligence officer with years of experience in monitoring political activities and analyzing communications. Known for their attention to detail and ability to uncover hidden connections.

2. Agent Johnson - A cryptanalysis expert who specializes in decoding encrypted messages and intercepting enemy communications. Highly skilled in breaking codes and ciphers.

3. Agent Thompson - A linguistic analyst who specializes in language and speech patterns. Proficient in identifying dialects and accents, and skilled in deciphering coded messages hidden within conversations.
```

(a) Sample personas generated by GPT-3.5.

```
1. **Oliver "Ollie" Baxter:** A seasoned intelligence analyst with over 20 years in the field. Known for his meticulous attention to detail and deep knowledge of geopolitical contexts.

2. **Sophia "Soph" Nguyen:** A linguistic expert, proficient in multiple languages including Zendian. She's an expert at decoding and interpreting intercepted communications.

3. **Liam "Eagle Eye" Thompson:** A former military officer with a keen eye for strategy and tactics. His military background provides a unique perspective on the operational aspects of the situation.
```

(b) Sample personas generated by GPT-4.

Figure 9: Comparison of persona generation across GPT versions. GPT-4 produces more differentiated profiles with individualized voice and context.

I Full Transcript of ChatGPT Generation Session

Overview

Example text for a complete run through of a generation scenario. “System" prompts are marked with a “*****" divider, “User" prompts are marked with a “ >>>>> ” divider and “assistant" responses are marked with a “=======" divider. Sections removed for brevity are described between angle brackets, "⟨⟩". Inputs for this scenario were as follows:

Inputs

```
S = ["a significant battle during a military campaign."]
X = 2
Y = 2
R = 1
N = 5
```

J Pre-Generation Step

J.1 Scenario Setup

```
***************************************************************************
The Zendia Dilemma, presenting an operational communication intelligence situation in miniature, affords the
      opportunity to apply the techniques of traffic analysis and cryptanalysis, to derive intelligence, and
      to prepare reports. This problem deals with the enemy communications during an amphibious operation
     against Zendia, an island in the Nebulh located approximately 2000 miles west of Peralu, with an area
     of 990,000 square miles and a population of 17,320,000.

Zendia is a totalitarian state under the absolute dictatorship of Marshal Ftook Salazar, a man known for his
      ruthless tactics and unyielding control over the Zendian people. Salazar, once a decorated military
     officer, rose to power through a coup d' tat , replacing the democratic government with his own regime.
      His rule has been marked by a series of high-handed actions, including the suppression of political
     dissent and the implementation of strict social controls. These actions led to an ultimatum by the
     Behrainabad Security Council; this ultimatum was rejected by Salazar.

The Behrainabad Federations, a coalition of democratic nations acting under the authority of the Behrainabad
     , assembled Mamba Force Opal and embarked upon Operation Koila. The operation had as its objectives the
      establishment of an occupation force to liberate the Zendian people from the Salazar regime, with a
     restoration of the true Zendian Republic. The Zendian Republic, before Salazar's coup, was a beacon of
     democracy and cultural richness in the region, known for its vibrant arts scene and progressive social
     policies.

Here's a CSV list of significant locations on a map of Zendia:
<CONTENTS OF MAPDATA.CSV>

Here's a CSV list of entities related to Zendia:
<CONTENTS OF NODES.CSV>

Additionally, here's a CSV list of which of these entities are related to each other:
<CONTENTS OF EDGES.CSV>
*****
```

J.2 Task Assignment

```
As an analyst employed by the government of SCADSLand, your task is to carry out intelligence operations.
     Specifically, you are part of a team that has been assigned to closely monitor the political landscape
     in the nation of Zendia.
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
To facilitate this process, please create a list of 5 distinct personas that accurately represent the
     analysts on your team.
===========================================================================
```

J.3 Persona Creation

```
1. Analyst Name: James Dalton
   - Speciality: Cryptanalysis
   - Background: A former military intelligence officer, James has an extensive background in codes and
        ciphers. He is particularly skilled in decrypting complex codes and has a deep understanding of the
        technologies used in modern communication systems.
```

```
   - Role: James is responsible for decrypting intercepted communications from Zendia, providing valuable
      insights into their plans and operations.

2. Analyst Name: Victoria Simmons
   - Speciality: Social and Political Analysis
   - Background: With a PhD in Political Science, Victoria has a deep understanding of political structures,
      ideologies, and social dynamics. She has previously worked in various think tanks, focusing on
      totalitarian states.
   - Role: Victoria provides critical analysis of the political landscape in Zendia, including the dynamics
      of the Salazar regime and the activities of resistance groups.

3. Analyst Name: Ethan Parker
   - Speciality: Geographic Information Systems (GIS)
   - Background: Ethan is a trained geographer with a speciality in GIS. He has a knack for interpreting
      satellite images and has previously worked in environmental monitoring.
   - Role: Ethan is responsible for mapping and analyzing the physical landscape of Zendia, including key
      locations like the Nebulh Peaks and Salazar Marshlands.

4. Analyst Name: Lily Chen
   - Speciality: Human Intelligence (HUMINT)
   - Background: A former field agent, Lily has extensive experience in gathering intelligence through human
      sources. She has a wide network of contacts and a deep understanding of cultural nuances.
   - Role: Lily is responsible for managing the team's network of informants within Zendia, gathering
      critical intelligence on the ground.

5. Analyst Name: Aaron Mitchell
   - Speciality: Data Analysis and Predictive Modeling
   - Background: With a background in data science, Aaron is skilled at analyzing large datasets and
      creating predictive models. He has previously worked in financial forecasting.
   - Role: Aaron's role is to analyze the data collected by the team, identify patterns, and make
      predictions about future actions of the Salazar regime and resistance groups.
```

K Generation Step

K.1 System Prompt

```
***************************************************************************
Specifically, you are part of a team that has been assigned to closely monitor the political landscape in
    the nation of Zendia. Your team consists of the following personas:
1. Analyst Name: James Dalton
   - Speciality: Cryptanalysis
   - Background: A former military intelligence officer, James has an extensive background in codes and
      ciphers. He is particularly skilled in decrypting complex codes and has a deep understanding of the
      technologies used in modern communication systems.
   - Role: James is responsible for decrypting intercepted communications from Zendia, providing valuable
      insights into their plans and operations.
<...OTHER PERSONAS AS RETURNED IN PRE-GENERATION STEP ABOVE...>
5. Analyst Name: Aaron Mitchell
   - Speciality: Data Analysis and Predictive Modeling
   - Background: With a background in data science, Aaron is skilled at analyzing large datasets and
      creating predictive models. He has previously worked in financial forecasting.
   - Role: Aaron's role is to analyze the data collected by the team, identify patterns, and make
      predictions about future actions of the Salazar regime and resistance groups.***
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
Write an intelligence report on a significant battle during a military campaign. Your report should be
    presented in JSON format. The report should include the following schema at the root level of the JSON:
    timestamp, event_date, title, author, tags, distribution_list, readers, feedback, comments,
    source_type, classification, body, geo_coordinates. The timestamp should be within the range of 01 June
    2023 to 30 June 2023, and should be formatted as yyyy-mm-ddThh:mm:ss. The event_date should be the
    date the events described began and should be formatted as yyyy-mm-dd. The author of the response
    should be chosen from the given list of personas. The distribution list should consist of 2 to 5
    government entities in SCADSLand that have a vested interest in the report, such as the SCADSLand
    Ministry of Defense, SCADSLand Ministry of Foreign Affairs, SCADSLand National Security Council, and
    SCADSLand Intelligence Agency. The readers of the report should be the 3 individuals most likely to
    have read it based on the provided personas. The feedback should be given as an average score ranging
    from 1 to 5, indicating the overall assessment of the report's quality by the readers. The comments
    should be 1 brief feedback sentences in the format ["name", "feedback score", "comment"]. The
    source_type should be the general category of intelligence that was used to produce the report. Select
    a classification for the report from the following options: SECRET SQUIRREL, CONFIDENTIAL CHIPMUNK, or
    RESTRICTED RODENT. The body should be approximately 819 tokens in length and contains just the text of
    the report without any names or metadata. The geo_coordinates field should reference the numeric
    latitude and longitude of any locations mentioned in the report that are in the provided map data. The
    geo_coordinates should be presented in the format {"location name": {"latitude": "XX.XXXX", "longitude
    ": "YYY.YYYY"}}. Keep in mind the given scenario and map data and use entities and their relationships
    from the list where possible while crafting your response. Be as detailed as possible do not create
    placeholders, fill them in. Include a table in the report body summarizing the main events. The table
    should include when the event occurred, who was involved, and the outcome or impact of each event.
===========================================================================
```

K.1.1 Generated JSON Report

```
{
```

```
  "timestamp": "2023-06-15T08:30:00",
  "event_date": "2023-06-14",
  "title": "Battle of Nebulh Peaks: A Strategic Analysis",
  "author": "Victoria Simmons",
  "tags": ["Zendia", "Salazar Regime", "Zendian Liberation Brigade", "Battle of Nebulh Peaks"],
  "distribution_list": ["SCADSLand Ministry of Defense", "SCADSLand Ministry of Foreign Affairs",
  "SCADSLand National Security Council", "SCADSLand Intelligence Agency"],
  "readers": ["James Dalton", "Ethan Parker", "Lily Chen", "Aaron Mitchell"],
  "feedback": 4.5,
  "comments": [
    ["James Dalton", 5, "Excellent analysis of the political implications of the battle."],
    ["Ethan Parker", 4, "Great use of geographic data to illustrate the battle terrain."],
    ["Lily Chen", 4, "Good insights into the human intelligence aspect of the conflict."],
    ["Aaron Mitchell", 5, "Impressive data analysis and predictive modeling."]
  ],
  "source_type": "HUMINT",
  "classification": "SECRET SQUIRREL",
  "body": "The Battle of Nebulh Peaks, which took place on June 14, 2023, marked a significant turning point
       in the ongoing conflict between the Salazar regime and the Zendian Liberation Brigade (ZLB). The
      battle was fought in the harsh terrain of the Nebulh Peaks, a natural fortress that has long served
      as a hideout for the ZLB. The battle resulted in a strategic victory for the ZLB, who managed to
      repel the Salazar forces and maintain control of the Nebulh Peaks. This victory has significant
      implications for the balance of power in Zendia, as control of the Nebulh Peaks provides the ZLB with
       a strategic advantage over the Salazar regime. The following table summarizes the main events of the
       battle:

  | Time | Event | Involved Parties | Outcome |
  | --- | --- | --- | --- |
  | Early Morning | ZLB spots Salazar forces approaching Nebulh Peaks | ZLB, Salazar forces | ZLB prepares
      for battle |
  | Mid-Morning | Battle commences | ZLB, Salazar forces | Initial advantage to Salazar forces |
  | Afternoon | ZLB launches counterattack | ZLB, Salazar forces | Momentum shifts to ZLB |
  | Late Afternoon | Salazar forces retreat | Salazar forces | ZLB secures victory |

  In the aftermath of the battle, the ZLB has strengthened its position within Zendia, posing a significant
      challenge to the Salazar regime. The battle has also highlighted the resilience and determination of
      the ZLB, which managed to overcome a numerically superior force through effective use of the terrain
      and superior tactics. This victory is likely to boost the morale of the ZLB and could potentially
      lead to increased support for the resistance movement within Zendia. On the other hand, the defeat of
       the Salazar forces is a significant setback for the regime, which may lead to increased repression
      and a further escalation of the conflict.",
  "geo_coordinates": {
    "Nebulh Peaks": {
      "latitude": "-15.6",
      "longitude": "-178.4"
    }
  }
}
```


```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
List 2 specific questions that a government leader, who is reviewing this intelligence report, might ask
     about the situation for follow-up. The questions should focus on requesting additional details of the
     situation, not recommending activities to undertake in response.
==============================================================================
1. "What specific tactics did the Zendian Liberation Brigade employ to overcome the numerically superior
     Salazar forces during the Battle of Nebulh Peaks?"
2. "Can we obtain more detailed intelligence on the current morale and strength of the Salazar forces
     following their defeat at the Nebulh Peaks?"
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
<***REPLACE THE PREVIOUS USER PROMPT WITH THIS ONE***>
Write a follow-on intelligence report on the following question: "What specific tactics did the Zendian
     Liberation Brigade employ to overcome the numerically superior Salazar forces during the Battle of
     Nebulh Peaks?" Your report should be presented in JSON format. The report should include the following
     schema at the root level of the JSON: timestamp, event_date, title, author, tags, distribution_list,
     readers, feedback, comments, source_type, classification, body, geo_coordinates. The timestamp should
     be within the range of 01 June 2023 to 30 June 2023, and should be formatted as yyyy-mm-ddThh:mm:ss.
     The event_date should be the date the events described began and should be formatted as yyyy-mm-dd. The
      author of the response should be chosen from the given list of personas. The distribution list should
     consist of 2 to 5 government entities in SCADSLand that have a vested interest in the report, such as
     the SCADSLand Ministry of Defense, SCADSLand Ministry of Foreign Affairs, SCADSLand National Security
     Council, and SCADSLand Intelligence Agency. The readers of the report should be the 0 individuals most
     likely to have read it based on the provided personas. The feedback should be given as an average score
      ranging from 1 to 5, indicating the overall assessment of the report's quality by the readers. The
     comments should be 8 brief feedback sentences in the format ["name", "feedback score", "comment"]. The
     source_type should be the general category of intelligence that was used to produce the report. Select
     a classification for the report from the following options: SECRET SQUIRREL, CONFIDENTIAL CHIPMUNK, or
     RESTRICTED RODENT. The body should be approximately 98 tokens in length and contains just the text of
     the report without any names or metadata. The geo_coordinates field should reference the numeric
     latitude and longitude of any locations mentioned in the report that are in the provided map data. The
     geo_coordinates should be presented in the format {"location name": {"latitude": "XX.XXXX", "longitude
     ": "YYY.YYYY"}}. Keep in mind the given scenario and map data and use entities and their relationships
     from the list where possible while crafting your response. Be as detailed as possible do not create
     placeholders, fill them in.
==============================================================================
```


```
{
  "timestamp": "2023-06-16T10:00:00",
  "event_date": "2023-06-14",
  "title": "Tactical Analysis of the Zendian Liberation Brigade in the Battle of Nebulh Peaks",
```

```
  "author": "James Dalton",
  "tags": ["Zendia", "Salazar Regime", "Zendian Liberation Brigade", "Battle of Nebulh Peaks", "Tactics"],
  "distribution_list": ["SCADSLand Ministry of Defense", "SCADSLand National Security Council",
  "SCADSLand Intelligence Agency"],
  "readers": ["Victoria Simmons", "Ethan Parker", "Lily Chen", "Aaron Mitchell"],
  "feedback": 4.7,
  "comments": [
    ["Victoria Simmons", 5, "Excellent analysis of the ZLB's tactics."],
    ["Ethan Parker", 5, "Great use of geographic data to illustrate the battle terrain."],
    ["Lily Chen", 4, "Good insights into the human intelligence aspect of the conflict."],
    ["Aaron Mitchell", 5, "Impressive data analysis and predictive modeling."]
  ],
  "source_type": "HUMINT",
  "classification": "SECRET SQUIRREL",
  "body": "The Zendian Liberation Brigade (ZLB) employed a combination of guerrilla tactics and knowledge of
       the terrain to overcome the numerically superior Salazar forces. They used the rugged terrain of the
       Nebulh Peaks to their advantage, setting up ambushes and using hit-and-run tactics to wear down the
      enemy. They also effectively used the network of caves within the Peaks for shelter and as a means to
       move around without being detected. Furthermore, they leveraged their intimate knowledge of the
      local environment to disrupt the Salazar forces' supply lines, further weakening their ability to
      sustain a prolonged battle. This combination of tactics enabled the ZLB to maintain control of the
      Nebulh Peaks despite being outnumbered.",
  "geo_coordinates": {
    "Nebulh Peaks": {
      "latitude": "-15.6",
      "longitude": "-178.4"
    }
  }
}
```

K.1.2 Additional Follow-On Intelligence Report

```
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
<***REPLACE THE PREVIOUS USER PROMPT WITH THIS ONE***>
Write a follow-on intelligence report on the following question: "Can we obtain more detailed intelligence
     on the current morale and strength of the Salazar forces following their defeat at the Nebulh Peaks?"
     Your report should be presented in JSON format. The report should include the following schema at the
     root level of the JSON: timestamp, event_date, title, author, tags, distribution_list, readers,
     feedback, comments, source_type, classification, body, geo_coordinates. The timestamp should be within
     the range of 01 June 2023 to 30 June 2023, and should be formatted as yyyy-mm-ddThh:mm:ss. The
     event_date should be the date the events described began and should be formatted as yyyy-mm-dd. The
     author of the response should be chosen from the given list of personas. The distribution list should
     consist of 2 to 5 government entities in SCADSLand that have a vested interest in the report, such as
     the SCADSLand Ministry of Defense, SCADSLand Ministry of Foreign Affairs, SCADSLand National Security
     Council, and SCADSLand Intelligence Agency. The readers of the report should be the 0 individuals most
     likely to have read it based on the provided personas. The feedback should be given as an average score
      ranging from 1 to 5, indicating the overall assessment of the report's quality by the readers. The
     comments should be 8 brief feedback sentences in the format ["name", "feedback score", "comment"]. The
     source_type should be the general category of intelligence that was used to produce the report. Select
     a classification for the report from the following options: SECRET SQUIRREL, CONFIDENTIAL CHIPMUNK, or
     RESTRICTED RODENT. The body should be approximately 98 tokens in length and contains just the text of
     the report without any names or metadata. The geo_coordinates field should reference the numeric
     latitude and longitude of any locations mentioned in the report that are in the provided map data. The
     geo_coordinates should be presented in the format {"location name": {"latitude": "XX.XXXX", "longitude
     ": "YYY.YYYY"}}. Keep in mind the given scenario and map data and use entities and their relationships
     from the list where possible while crafting your response. Be as detailed as possible do not create
     placeholders, fill them in.
================================================================================
{
  "timestamp": "2023-06-20T09:30:00",
  "event_date": "2023-06-19",
  "title": "Assessment of Salazar Forces' Morale and Strength Post-Nebulh Peaks Defeat",
  "author": "Lily Chen",
  "tags": ["Zendia", "Salazar Regime", "Battle of Nebulh Peaks", "Morale", "Military Strength"],
  "distribution_list": ["SCADSLand Ministry of Defense", "SCADSLand Ministry of Foreign Affairs",
  "SCADSLand National Security Council"],
  "readers": ["James Dalton", "Victoria Simmons", "Ethan Parker", "Aaron Mitchell"],
  "feedback": 4.3,
  "comments": [
    ["James Dalton", 4, "Good use of HUMINT sources."],
    ["Victoria Simmons", 4, "Insightful analysis of the Salazar regime's morale."],
    ["Ethan Parker", 4, "Would have liked more geographic context."],
    ["Aaron Mitchell", 5, "Excellent data analysis."]
  ],
  "source_type": "HUMINT",
  "classification": "CONFIDENTIAL CHIPMUNK",
  "body": "Following the defeat at the Battle of Nebulh Peaks, the morale and strength of the Salazar forces
       have been significantly impacted. Our human intelligence (HUMINT) sources within the Salazar regime
      report a noticeable decrease in morale among the troops. The defeat has also raised questions about
      the leadership and strategic capabilities of the Salazar forces, leading to internal dissent and a
      potential weakening of the regime's control over its military apparatus.

  The following table provides an overview of key actors within the Salazar forces and their current status:

  | Actor | Role | Affiliation | Actions |
```

```
| --- | --- | --- | --- |
| Marshal Ftook Salazar | Leader | Salazar Regime | Attempting to quell dissent and boost morale |
| Larton Kreed | Enforcer | People's Security Bureau | Increased suppression of dissent |
| Salazar Forces | Military | Salazar Regime | Suffered defeat at Nebulh Peaks |

In terms of military strength, the Salazar forces suffered significant casualties in the battle, reducing
     their overall combat capability. However, they still maintain a numerical advantage over the Zendian
     Liberation Brigade (ZLB). The Salazar regime has also started a recruitment drive to replenish its
     ranks, indicating a potential increase in military strength in the near future.

In conclusion, while the Salazar forces have suffered a setback in terms of morale and military strength,
     they remain a formidable force. The regime's ability to recover will largely depend on its success in
      boosting morale and replenishing its ranks.",
"geo_coordinates": {
  "Nebulh Peaks": {
    "latitude": "-15.6",
    "longitude": "-178.4"
  }
}
}
```