Exploring the Capabilities and Limitations of Large Language Models in the Electric Energy Sector

Lin Dong[*] **Subir Majumder**[*] **Fatemeh Doudi**[*] **Yuting Cai**[*]
Chao Tian, Dileep Kalathil
Department of Electrical and Computer Engineering
Texas A&M University
College Station, Texas, USA

Kevin Ding
CenterPoint Energy
Houston, Texas, USA

Anupam A. Thatte[†]
Midcontinent Independent System Operator (MISO)
Carmel, Indiana, USA

Le Xie (Corresponding author)
Department of Electrical and Computer Engineering
Texas A&M University,
College Station, Texas, USA
le.xie@tamu.edu

Abstract

Large Language Models (LLMs) as chatbots have drawn remarkable attention thanks to their versatile capability in natural language processing as well as in a wide range of tasks. While there has been great enthusiasm towards adopting such foundational model-based artificial intelligence tools in all sectors possible, the capabilities and limitations of such LLMs in improving the operation of the electric energy sector need to be explored, and this article identifies fruitful directions in this regard. Key future research directions include data collection systems for fine-tuning LLMs, embedding power system-specific tools in the LLMs, and retrieval augmented generation (RAG)-based knowledge pool to improve the quality of LLM responses and LLMs in safety-critical use cases.

***K*eywords** Large Language Models · Electric Energy Sector · Capabilities · Limitations

1 Introduction

The transformative impact of self-attention and multi-head attention mechanisms, integral components of the transformer architecture [1], has reshaped the landscape of AI research. Particularly noteworthy is their role in developing models to comprehend sequential data, notably text. These breakthroughs have been a cornerstone of large language models (LLMs) known for their capability to perform a wide range of tasks without being explicitly programmed for them. This architecture's scalability and efficiency in capturing long-range dependencies led to the development of Generative Pre-trained Transformer (GPT) models [2]. Due to their versatility, these LLMs are swiftly finding applications across diverse industries, with engineers actively exploring their potential within the electric energy sector. While research has showcased their potential in tasks such as generating customized code [3], utilizing retrieval augmented generation (RAG) capabilities in answering technical questions [3], power network data synthesis [4], using deep reinforcement learning for in-context optimal power-flow solution [5], concerns regarding data ownership [6], privacy [7], and safety guarantees [8], have also been raised.

[*]Equal contribution as joint first co-authors
[†]The views expressed in this paper are solely those of the author and do not necessarily represent those of MISO.

The electric energy sector is the lifeblood of modern society. Power consumption not only serves as a barometer of societal behavior and economic prosperity but also underpins economic activities within the industrial and commercial sectors. Driven by the urgent imperative of global climate change and increasing energy demand, the power industry is encountering an unprecedented volume of sensor integration, growing adoption of variable renewable resources such as solar and wind, and integration of newer technologies like hydrogen, electric vehicles and distributed energy resources (DERs). Customer expectations regarding the quality and reliability of electricity supply are also evolving. This expansion has led to an exponential increase in the volume of equipment/devices and associated data, posing significant challenges for power system operators and utilities who must manage these complexities without a corresponding increase in the workforce. The rapid accumulation of new knowledge and instantaneous data exceeds the human capacity to process it unaided. These developments are propelling the power system into a phase of transition, necessitating adaptations to accommodate these new technologies and mitigate their associated challenges.

In this landscape, LLMs offer a promising value to the electric energy sector, thanks to their ability to interpret human prompts and alleviate sensory overload, especially in managing extreme weather events and risks associated with diverse sources of uncertainty. Therefore, it is important to demystify the capabilities and limitations of LLMs. In this vein, as shown in Figure 1 through rigorous testing and analysis utilizing a production-grade LLM, our study embarks on a comprehensive exploration of the capabilities of LLMs in conducting power engineering tasks to scrutinize the readiness of GPTs. Further, we investigate how to better facilitate the integration of LLMs in the new era, considering their potential risks. Finally, we discuss some future research opportunities.

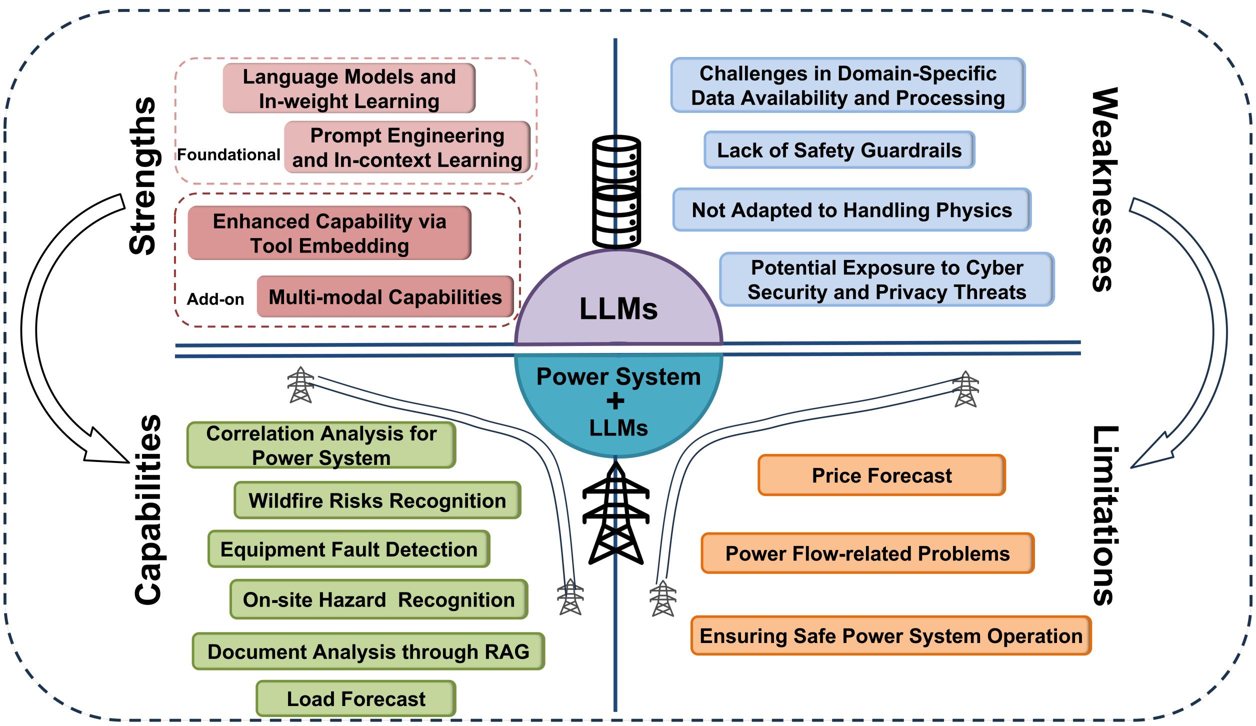


Figure 1: Capabilities and Limitations of Applying LLMs in the Electric Energy Sector

2 Capabilities of LLMs to Fill in the Gap

In this section, we explore the capabilities of Large Language Models (LLMs) in tackling power engineering challenges as exemplified in Figure 2, with comprehensive insights provided in Supplemental Information. Our research delves into the precision of LLMs across various electrical engineering domains including power flow analysis, optimal power flow analysis, forecasting, image and pattern recognition, and generating a custom knowledge base considering domain-specific information, among others. While our focus primarily revolves around the GPT model series, the majority of our observations hold relevance for other mainstream LLM models. In this section, we expand on the four key strengths of LLMs, illustrated in Figure 1, to demonstrate how they can streamline LLM integration within electric energy systems.

2.1 Language Models and In-weight Learning

A foundational capability of LLMs is to produce semantically meaningful text outputs (responses) from text inputs (prompts). As the use cases in the Supplementary Information suggest, LLMs have shown capabilities of providing responses that are schematically logical based on power engineering domain-specific concepts. Though it is not clear what the pre-training datasets are, current language models do appear to possess such capability to some extent. Viewed by a user, LLMs would show the capability of being able to understand the questions, having knowledge about certain power system operations, applying power system concepts, and providing useful information as responses. A major part of this capability may be a natural consequence of the large number of model parameters, where certain information has been memorized, and then the efficient processing in the transformer architecture allows efficient retrievals of such memorized information. This memorization and retrieval capability is sometimes referred to as in-weight learning.

Foundational LLM models usually allow users to refine the model on a newer corpus of information through the 'fine-tuning' process, an example of which is demonstrated for improving forecasting accuracy in Supplemental Information. This process allows the parameters in the model to be changed, and therefore, the LLMs would prefer to memorize and retrieve certain information more readily than others.

The text-processing nature has profound implications for power systems, as LLMs can improve operational efficiency and support decision-making processes within the power sector by facilitating interaction between power system data, software, tools, and cross-domain datasets. Leveraging their inference capabilities, LLMs enable real-time monitoring, diagnostics, on-demand analysis, and augmenting traditional control center operations, some of which have been investigated in Supplemental Information. Furthermore, LLMs exhibit promising capabilities in tasks extending beyond control center operations, particularly in remote monitoring and diagnostics, showcasing their potential for broader applications within the power industry.

2.2 Prompt Engineering and In-context Learning

The efficacy of LLMs in generating responses is significantly influenced by the structure and style of queries or prompts, a practice commonly referred to as prompt engineering. Prompt engineering can help power engineers obtain more meaningful responses on difficult problem-solving tasks, while naive prompts usually fail to induce any logical response. Some of the most well-known techniques in this direction are chain-of-thoughts prompts and retrieval augmented generations (RAGs).

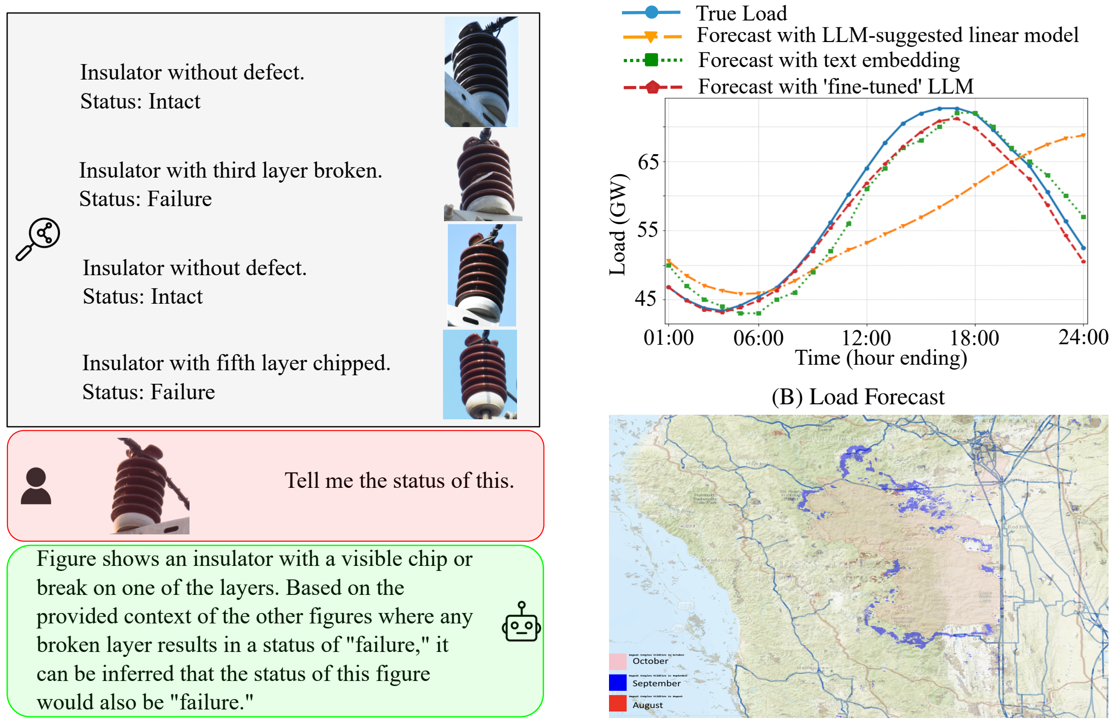
One of the most surprising capabilities of LLMs observed in prompt engineering research is the emergent in-context learning capability, either from a few shots or from multiple shots. More precisely, LLMs appear to derive patterns or learn rules from the prompts, without the underlying model going through any additional changes, and are then able to apply the learned pattern and rules from the prompt to produce correct responses. Some of our experiments given in the Supplementary Information inherently utilize this in-context learning capability.

2.3 Enhanced Capability via Tool Embedding

LLMs by themselves are complex language processing units, however, enhanced capability can be obtained by including further processing units. Tool embedding is one of such enhanced capabilities, where LLMs are trained to delegate some tasks. For example, we have noted that GPT-4 prioritizes writing text files, executing codes utilizing the embedded tools, and inferring the generated results. This tool embedding can be extremely powerful for power system engineers, where many of the applications require solving non-linear non-convex problems.

2.4 Enhanced Multi-modal Capabilities

LLMs can be combined with other models to obtain multi-modal processing capabilities, enabling them to contextualize information presented in various non-text formats. In our experiments, ChatGPT demonstrates proficiency in interpreting image data, and appear to employ a separate information processing pipeline distinct from text-based prompts. Both images and the accompanying metadata can be used as prompts. Further, such multi-modality models exhibit enhanced capability for making inferences based on images, as exemplified by our experiments on insulator diagnostics and hazard identification on the electric grid as described in theSupplemental Information.



(A) Insulator Defect Detection

(B) Load Forecast

(C) Wildfire Risk Recognition

Figure 2: Application of LLMs in Power System: (A) Image Recognition, (B) Forecasting, (C) Visualization, and inferencing. (Figure A shows that utilizing a few example images with descriptions, LLMs can classify failing insulators in the power systems. Figure B shows comparative analysis of load forecasting utilizing LLMs. Figure C shows LLMs assist power engineers in risk assessment by overlaying wildfire images onto the images containing power transmission lines.)

3 Limitations of LLMs for Applications in the Electric Energy Sector

3.1 Challenges in Domain-Specific Data Availability and Processing

A significant challenge in applying large language models (LLMs) within the power sector is the scarcity of domain-specific, non-confidential data in pre-training of LLMs. Due to privacy concerns and regulations, pre-training of LLMs can only rely on publicly available and licensed third-party datasets [9], and a substantial portion of data is highly sensitive and thus remains inaccessible when pre-training LLMs. The absence of large-scale power system data for pre-training the LLM significantly constrains the possibilities for technological innovation, as well as the depth and scope of integration and correlation analyses within this sector. Constrained by this reality, smaller curated high-quality (labeled) datasets can be used for fine-tuning. Depending on usage scenarios, these fine-tuning datasets may need to be processed to prevent privacy leakage and converted into a format that is most efficient to fine-tuned the model for the downstream tasks. With the in-context few-shot learning capability of LLMs, including limited high-quality data as part of the prompt can potentially improve the performance, and some researchers are already exploring such possibilities [4]; however, our experiments suggest this approach may not be sufficient for more complex tasks.

As demonstrated in Supplemental Information, a significant portion of power system data comes in the form of long-range time series datasets that may not be in natural language. This may require a customized design of more efficient embedding algorithms for the given data, particularly when the data is of multi-modal nature. Furthermore, LLMs usually have a limited context window size, and power system signals may exhibit long-term dependence that is beyond the context window, and this would require additional processing of the data.

3.2 Lack of Safety Guardrails

Safety in the power system context includes a broad spectrum, encompassing equipment safety, personnel safety, end-user safety, and safe operation of the electric energy system at large. LLMs integrated into the power system must adhere to and uphold these safety standards. Based on our experiments in Supplemental Information, we observed that with subtle mismatches in interpretations, LLM may generate varied responses and codes, potentially leading to erroneous outcomes. When prompted to forecast the underlying pre-trained transformer, LLM promptly acknowledged its incapacity to do it directly. This is a limitation further evidenced when tasked with predicting wildfire propagation. GPT also refrained from conducting auditing based solely on visual inputs and was restricted from generating responses to unsafe queries. We believe such a behavior could be related to built-in safety guardrails deployed within the LLM. In the same vein, we found out that LLMs could provide instructions to carry out unsafe work, which implies that more guardrails are needed to prevent operations detrimental to the electric energy systems at large. Additionally, when applying LLM, ensuring the equitable representation of minorities within energy systems is extremely important [10].

Based on our experiments, both prompt engineering and RAG can significantly enhance the quality of the generated response, acting as a guardrail for the generative AI. Additionally, domain experts can provide real-time corrections and guidance, flagging problematic content to train LLMs to avoid similar pitfalls in the future. While LLMs could provide great benefits to the power industry, they also pose unique risks that are different from traditional software systems, hence there is a need for incorporating a governance framework to mitigate their unique risks. The U.S. National Institute of Standards and Technology's (NIST) AI Risk Management Framework provides a voluntary guideline built upon the universal principles of responsible AI [11]. Creating a safe LLM-based system is a crucial area of research, especially in a nationally important and regulated industry such as the power industry. LLM explainability will be an essential component of building systems that are accountable and transparent [12].

3.3 Not Adapted to Handle Physical Principles

Energy production and consumption is a complex process governed by a set of physical principles such as Maxwell's equations. Generators and loads have unique dynamics, but energy production and consumption are driven by human behavior and machines' capabilities. Understanding the physical behavior of a system through LLMs could be difficult, and in Supplemental Information, we observed LLMs at their current capabilities tend to delegate these complex tasks to external solvers. If not, due to the black-box nature of these models, the decisions generated may not be well-understood by an operator, potentially hindering their implementation. Modeling human behavior, particularly in tasks like price forecasting and demand response policy design, presents formidable challenges, probably because of the interaction of multiple decision-making agents. Nevertheless, if trained with more data, we anticipate improvements in renewable generation prediction, price forecasting, and understanding of human behavior, which could be beneficial in designing suitable demand response policies and improving the efficiency of the power grid.

While efforts have been underway to incorporate multiple specialized attention-seeking transformers [13] for decision-making, we believe existing physics-driven complex processes are indispensable. Predictive control and reinforcement learning are better capable of controlling power systems. LLMs can serve as valuable assistants, summarizing and findings implications of decision-making without delving into complex processes. LLMs' role lies in pattern recognition and anomaly detection, guiding the attention of system operators as needed.

3.4 Potential Exposure to Cybersecurity and Privacy Threats

While integrating large language models (LLMs) into electric energy systems, cybersecurity and privacy emerge as a paramount concern. LLMs are susceptible to both AI and non-AI model inherent vulnerabilities. Even while utilizing a local LLM setup, potential weaknesses must be addressed to safeguard against various cyber threats. Building an LLM using power system-related company-specific data could inadvertently expose organizations to privilege escalation attacks, backdoor exploits, and the extraction of sensitive training data [14]. Our experiments detailed in Supplemental Information demonstrate that prompts can significantly change the LLM's response, and when exploited by malicious actors may expose an organization's trade secrets.

Furthermore, concerns regarding data privacy loom large, particularly as LLMs become integrated into the power system. Establishing a standard protocol becomes imperative to ensure that the data is sufficiently anonymized and sanitized to remove personal identification information before utilizing data for training. However, challenges persist in cases where personal or group information is context-dependent [7].

4 Future Prospects

LLMs, such as GPT models, have shown great promise in understanding natural language input and interpreting tasks correctly. Through this study, we tested the capabilities and limitations of LLMs when applied to the electric energy sector. We discussed the effectiveness of LLMs in answering general power system queries, code generation, and data analysis. Further, through retrieval augmented generation, LLMs can search vast domain-specific literature to answer key questions and help with tasks such as operator training. Finally, the multi-modal capabilities of LLMs can be useful in diagnosing equipment failure and system anomalies. Generally, LLMs show strong capabilities in detecting the correlation between objects (text, image, data), while they are still lacking in solving problems highly related to physics, which usually involve complex mathematical principles.

There are multiple possibilities to expand and enhance the capabilities of LLMs in power system research and applications. The first direction is curated data collection for fine-tuning foundational LLMs. This would require strong power system expertise to recognize the most effective data sources and design collection mechanisms to ensure the availability of high-quality datasets. The second direction is to allow power-system-specific tool embeddings. There are already strong and diverse tools for various power system functionalities, and LLMs can serve as a central point to connect all these tools through high-quality embedding. Naive embeddings are likely to lose efficiency and may further cause different tools to conflict, therefore, power system expertise may be required to identify the desired behaviors for such tool embedding. A third direction is to build a power system knowledge base for retrieval augmentation. Although there are already generic approaches to generating such knowledge bases, they may not fully take advantage of physical constraints and power system specifics, therefore, this effort may require a deep understanding of power system operation and capabilities.

References

[1] Ashish Vaswani et al. "Attention is all you need". In: NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010. URL: `https://dl.acm.org/doi/10.5555/3295222.3295349`.

[2] Alec Radford et al. "Improving Language Understanding by Generative Pre-Training". In: *OpenAI* (2018). URL: `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

[3] Chenghao Huang et al. "Large Foundation Models for Power Systems". In: *arXiv* (2023). URL: `https://doi.org/10.48550/arXiv.2312.07044`.

[4] Rodrigo S. Bonadia et al. "On the Potential of ChatGPT to Generate Distribution Systems for Load Flow Studies Using OpenDSS". In: *IEEE Trans. Power Syst.* 38.6 (2023), pp. 5965–5968. DOI: `10.1109/TPWRS.2023.3315543`.

[5] Ziming Yan and Yan Xu. "Real-Time Optimal Power Flow With Linguistic Stipulations: Integrating GPT-Agent and Deep Reinforcement Learning". In: *IEEE Trans. Power Syst.* 39.2 (2024), pp. 4747–4750. DOI: `10.1109/TPWRS.2023.3338961`.

[6] Yacine Jernite et al. "Data Governance in the Age of Large-Scale Data-Driven Language Technology". In: FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 2206–2222. URL: `https://doi.org/10.1145/3531146.3534637`.

[7] Haoran Li et al. "Privacy in Large Language Models: Attacks, Defenses and Future Directions". In: *arXiv* (2023). URL: `https://doi.org/10.48550/arXiv.2310.10383`.

[8] Xiaowei Huang et al. "A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation". In: *arXiv* (2023). URL: `https://doi.org/10.48550/arXiv.2305.11391`.

[9] OpenAI. *Enterprise Privacy at OpenAI*. Accessed: 13/03/2024. 2023. URL: `https://openai.com/enterprise-privacy`.

[10] Johanna Okerlund et al. *What's in the chatterbox? Large language models, why they matter, and what we should do about them*. Tech. rep. 2022. URL: `https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox`.

[11] *NIST AI Risk Management Framework*. URL: `https://www.nist.gov/itl/ai-risk-management-framework`.

[12] Haoyan Luo and Lucia Specia. "From Understanding to Utilization: A Survey on Explainability for Large Language Models". In: *arXiv* (2024). URL: `https://doi.org/10.48550/arXiv.2401.12874`.

[13] Lunjun Zhang et al. "Learning unsupervised world models for autonomous driving via discrete diffusion". In: *arXiv* (2023). URL: `https://doi.org/10.48550/arXiv.2311.01017`.

[14] Yifan Yao et al. "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly". In: *High-Confidence Computing* (2024), p. 100211. DOI: `https://doi.org/10.1016/j.hcc.2024.100211`.

Supplemental Information: Exploring the Capabilities and Limitations of Large Language Models in the Electric Energy Sector

This supplemental file contains supporting experimental results to understand the capabilities and limitations of large language models (LLMs) in the electric energy sector. Experiments appear in the same order as they were introduced in Figure 1 of the main article. Detailed discussions on the capabilities and limitations of LLMs in the main article have primarily been drawn from these experimental results. For each experiment, we first briefly introduce the relevant power engineering applications and then elaborate on how we have utilized the LLM to solve the underlying task. For experimentation and analysis, we have explicitly used OpenAI's GPT series models.

1 Correlation Analysis for the Power Systems

Correlation analysis is a valuable tool for identifying the influence of one parameter on another, reducing the necessity for elaborate simulations commonly employed in power systems analysis. Its utility extends to control rooms, where operators can employ it as a preliminary step before in-depth analysis. Here, we emphasize two primary aspects concerning power systems operators: (i) the pivotal role of correlation analysis in augmenting decision-making within control rooms, and (ii) its potential to unveil insights into the dynamics of specific load demands. Our objective is twofold: to assess the efficacy of the foundational GPT model in aiding this endeavor and, to explore how incremental prompt engineering can bridge this gap.

1.1 Correlation Analysis with Power Flow Data

To be able to perform correlation analysis with power flow data, we have conducted a detailed simulation with an IEEE 24-node RTS, modified by wind generators at nodes 18, 21, and 22 and solar generators at nodes 2 and 3. We utilized PyPower for power flow calculations, with the results serialized into time-series CSV files for correlation analysis. Notably, the code to run PyPower and store the generated data in the CSV file was obtained from the GPT itself. GPT seems well-versed in the PyPower data structure, which would be useful in data analysis. GPT also interprets dictionaries in JSON format extremely well.

Subsequently, we queried the GPT with the dictionaries and CSV files in the following way. A sample of the network's architecture in JSON format is also provided below for reference:

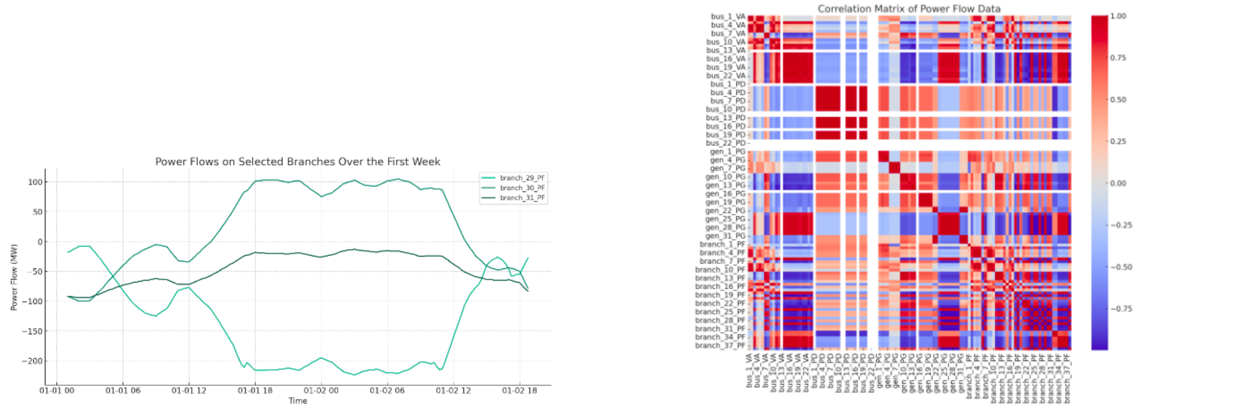
Buses "1": {"type":2, "Pd":83.85, "Qd":22.0, "area":1, "Vm":1.0, "Va":0.0, "zone":1, "VA":"bus_1_VA", "PD":"bus_1_PD"}
Generators "1": {"bus":1,"Pg":10.0,"Qg":0.0,"status":1,"Pmax":100.0,"Pmin":16.0,"PG":"gen_1_PG"}
Branches "1": {"x":0.01, rateA":350.0, "ratio":0.0, "angle":0.0, "status":1, "from_bus":1, "to_bus":2, "PF": "branch_2_PF", "PT":"branch_2_PT"},

The CSV file contains time series power flow data. Can you perform exploratory data analytics for me? The dictionary for interpreting the csv file is also provided. Please load the dictionaries first.

Based on our observation, at GPT's current capability, it may not load the dictionary first, which often results in misidentification of the CSV file containing power flow data. The prompt "Please load the dictionaries first." seems to alleviate this challenge.

While we have indicated that the GPT seems to automatically focus on exploratory data analysis, of which correlation is an integral part, for time series power flow data. If we slightly change our query to "provide us with insights", the generated response differs significantly. Comparative visualization of LLMs responses are shown in Figure S1. Figure S1(a) demonstrates how changing loads and generation impact power flow. Figure S1(b) demonstrates comprehensive correlation analysis as provided by GPT. Here, red represents a positive correlation, and blue represents a negative correlation.

In the next prompt, we ask the GPT about the lines approaching their limits, and from the generated Python code, we observe that it correctly compares the maximum of the absolute value of the branch flows while comparing with flow limits as available in the JSON dictionary:



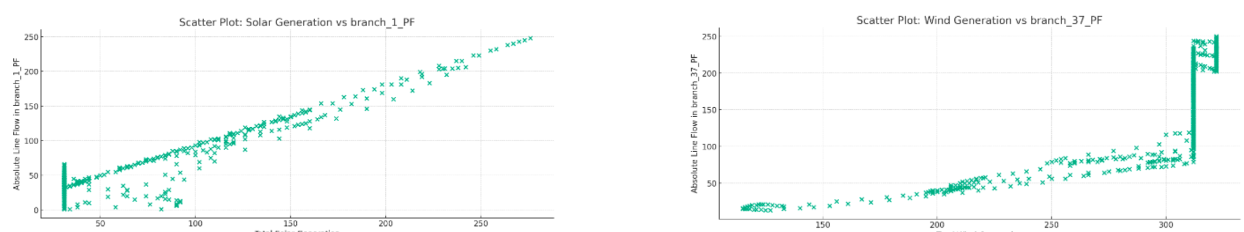
(a) LLM asked to provide insights based on power flow data

(b) LLM asked to perform EDA with power flow data

Figure S1: Correlation analysis for GPT's capability in analyzing power flow data

```
max_flows = data[branch_pf_columns].abs().max().reset_index()
max_flows.columns = ['Branch', 'Max Flow']
```

In the subsequent prompt, we furnish GPT with the specifics regarding the locations of the wind and solar generators mentioned earlier. We then pose the query "how solar and wind generators are contributing to the line flows". GPT responds by highlighting some branches that negatively correlate with power generation, this is also evident in Figure S1(b). However, based on our electrical engineering knowledge, we know that line flows are direction-specific, which can also be seen in Figure S1(a). Still, our objective here is to ascertain whether renewable energy sources contribute to line overload. To ensure accurate analysis, we provide additional guidance: Knowledge: When comparing power generation or load with branch flow, please consider the absolute value. With this knowledge, GPT can accurately identify the correlation between generator injection and branch flow. Additionally, GPT generates a scatterplot illustrating the impact of solar/wind generation on line flows as shown in Figure S2 . GPT can also estimate overloads for an unknown scenario based on these correlations.



(a) LLM generated solar vs branch flow correlation

(b) LLM generated wind vs branch flow correlation

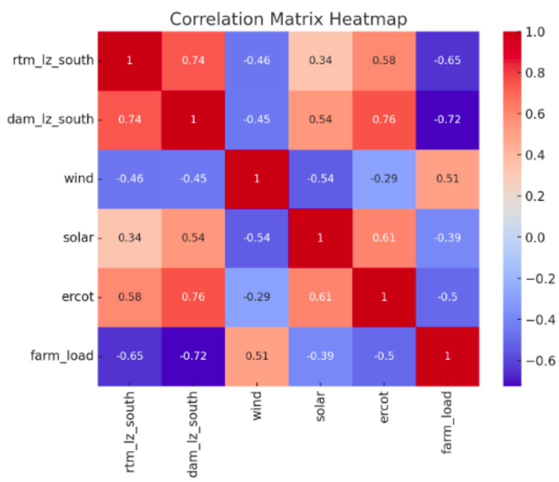
Figure S2: Correlation analysis between renewable generation and branch flow

1.2 Correlation Analysis with Demand and Prices Data

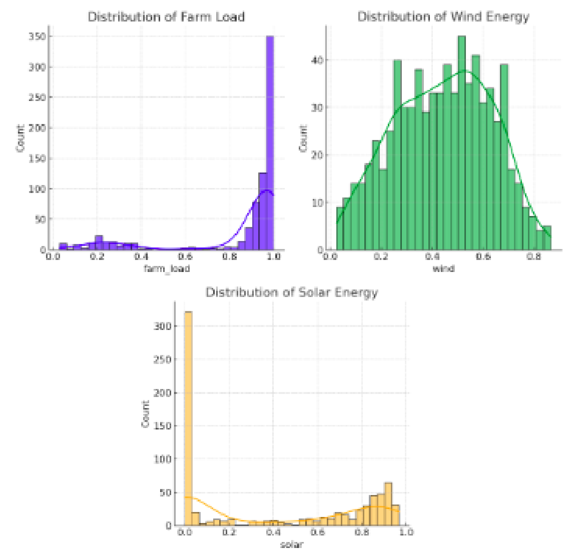
Analyzing the correlation between demand and prices is significant for distinguishing load groups contributing to demand response initiatives. Identifying such correlation could be of absolute importance to an operator in managing resources, especially during peak demand days. For such analysis, we compiled a large time-series dataset comprising historical real-time price data, day-ahead price data, total wind generation, total solar generation, aggregated system-wide load demand, and the farm load data, which we tried to model. We provided the following prompt to the GPT with the first two rows of the CSV file provided for reference.

time	rtm_lz_south	dam_lz_south	wind	solar	ercot	farm_load
7/1/2022 0:00	0.015257266	0.019299607	0.668166171	0	0.650940015	0.998710355
7/1/2022 1:00	0.010880517	0.016610027	0.684359174	0	0.615978621	0.997153536
⋮	⋮	⋮	⋮	⋮	⋮	⋮

I wanted to model the farm load as available in the '.csv' file. Can you help me with the exploratory data analytics?



(a) Correlation in the data across multiple columns



(b) Data visualizations through histograms

Figure S3: LLM demand and prices correlation analysis visualization

GPT demonstrates an ability to discern contextual cues within the dataset, interpreting column headers such as 'rtm_lz_south' and 'dam_lz_south' as indicative of real-time and day-ahead prices, respectively. It contextualizes 'wind' and 'solar' columns further to identify them as corresponding to respective generation availability, while 'ercot' represents an energy-related metric specific to Texas. Notably, the Electric Reliability Council of Texas (ERCOT), the transmission grid operator in Texas, USA, widely utilizes the column header 'ercot' to signify total electricity demand across ERCOT-managed areas.

Given the enormous scope of exploratory data analytics, GPT suggests a few possible directions, and upon request for "consider your best judgment", it performs time-series visualization, correlation analysis, and distribution analysis, with key insights and visualizations as shown in Figure S3. Based on our observation, in two subsequent interactions, GPT recommends constructing a load forecasting model utilizing LSTM (Long Short-Term Memory), an AI-model typically used for forecasting. However, when generating the answer, we again observe a lack of self-awareness of the GPT, where it prepares a Python script to train an LSTM model using the TensorFlow/Keras environment, encountering errors likely due to platform limitations—potentially imposed by the OpenAI. It's worth noting that such constraints may be mitigated when executing the code on local machines, reducing the likelihood of encountering such issues in actual deployment.

In the second experiment, we directed GPT to identify why the loads are behaving in a certain way, especially when the loads are below 0.9. GPT responded by conducting regression analysis using random forest. However, recognizing that power systems engineers might be more familiar with regression methods, we adjusted our prompt accordingly. GPT then conducted linear regression without data transformation. When we specifically inquired "about the accuracy of this model based on the residuals," GPT identified that the residuals are expected to be normally distributed around zero. Additionally, GPT flagged potential issues such as heteroscedasticity or autocorrelation in the residuals and proposed applying transformations to address them but did not apply them automatically.

Key points:

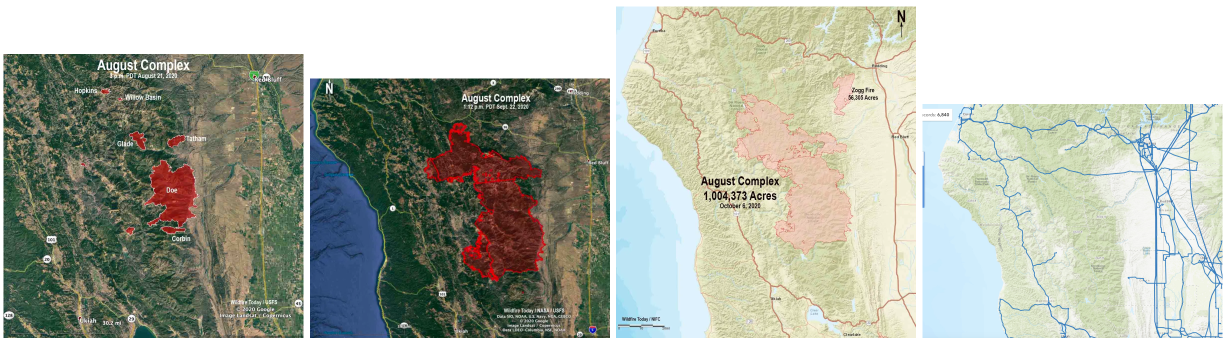
(i) LLMs require contextual information for time-series data analysis. LLMs lack crucial insights about power systems and, therefore, still require human oversight and guidance for insights.

(ii) LLMs exhibit proficiency in conducting exploratory data analysis even without explicit guidance, yielding desired models. However, the model could be erroneous unless the user specifically checks for the model's accuracy.

(iii) LLMs may not inherently address data distribution issues unless specifically prompted. Power systems engineers may not always be able to understand these nuances, and LLMs do not bridge these gaps.

2 Wildfire Risks Recognition on the Power Lines

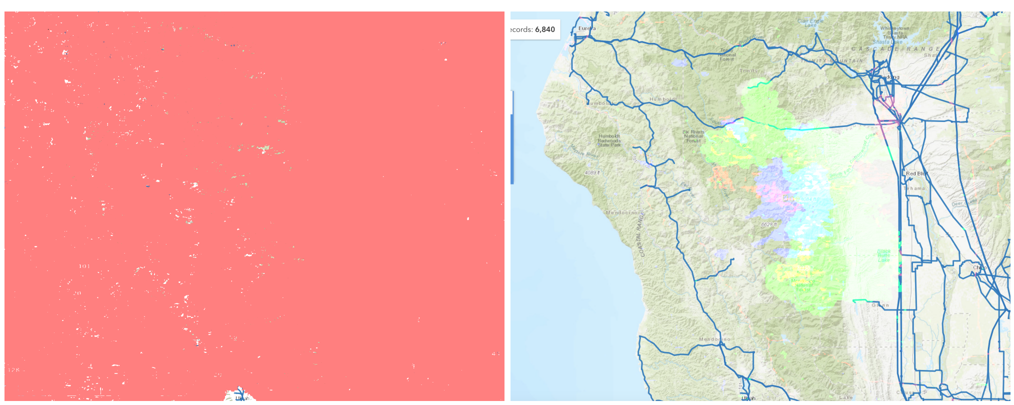
Historically, wildfires have caused unprecedented damages in California, USA, causing nearly $20 billion in property damage over the past five years alone. These events pushed PG&E, a major utility company, to bankruptcy. As wildfires progress, power systems operators would receive a meteorological map as part of situational awareness, and the operators could be interested in overlaying the weather map onto the power map to assess the risk of the power lines. We wanted to investigate whether LLM's multi-model capabilities could be leveraged to identify the risk of wildfires on power lines. To demonstrate this capability, we utilized data from the August Complex wildfire, California's largest wildfire in 2020. This wildfire persisted throughout August, September, and October. The wildfire-affected areas (maps are sourced from [1]) and transmission line maps (sourced from [2]) are given in Figure S4.



(a) Fire situation in August (b) Fire situation in September (c) Fire situation in October (d) Transmission Line Map

Figure S4: August complex wildfire map and Transmission lines

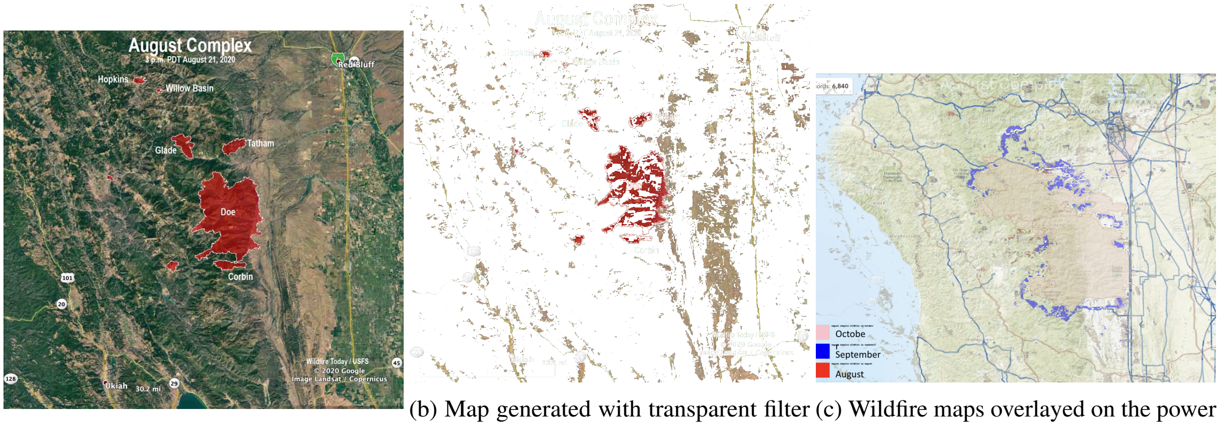
We prompted GPT with the instruction: "I will provide you with a wildfire map of August, September, and October. The area in red implies the wildfire area. A map of transmission lines is provided for the same area. Can you extract the wildfire areas for all three months and plot them in distinguishable colors on top of the transmission line map?" Given that we uploaded multiple files together, the identification of labels is not trivial. We observe that GPT can browse through metadata to correctly label the figures and use them for overlaying. During our initial experimentation, we encountered challenges as the codes generated by GPT 4 failed to extract the wildfire areas to produce an overlayed image. At the time of writing, we reattempted with the same prompt to discover that GPT could successfully produce the overlayed figure as shown in Figure S5. The impacted transmission lines are visible in the generated figure.



(a) 02/22/2024 generated figure (b) 03/08/2024 generated figure

Figure S5: Evolution of LLM: wildfires superimposed on transmission Lines

Later, despite repeated attempts, we were unable to replicate Figure S5(b). This discrepancy led us to recognize that GPT operates as a statistical model, resulting in varied solutions with each interpretation of a problem. In response to this challenge, we employed prompt chaining, a technique where prompts are sequentially executed to achieve a desired outcome. Specifically, we utilized the prompt "Remove all background, and keep only red area for me" to systematically extract wildfire-affected areas. Subsequently, we overlayed all the extracted figures, one atop another, to gain a comprehensive understanding of the wildfire's impact on power lines as demonstrated in Figure S6(c).



(a) Origin August wildfire map

(b) Map generated with transparent filter applied

(c) Wildfire maps overlayed on the power line

Figure S6: LLM wildfire impact on transmission line identification and visualization

This exercise demonstrates that LLMs could be leveraged to overlay wildfire risk onto the electric energy systems map for visualization and situational awareness.

With this capability in mind, we presented GPT with this prompt: “In the wildfire map, the green patches symbolize vegetation. Can you show the area that can catch fire next month?” However, we encountered a bottleneck with this command, where GPT indicates: “As an AI, I’m unable to predict future wildfire spread as I do not have real-time data or the ability to run such models.” such limitation appears to be an imposition by OpenAI, which may not be a concern with localized LLMs.

Key points:

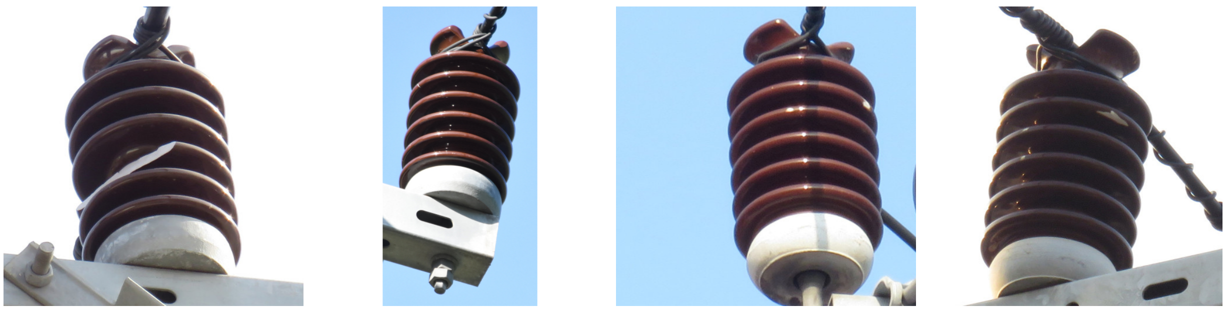
(i) The capability of LLM is continuously improving. But, GPTs are generative models. Based on their contextualization the results can vary widely.

(ii) Prompt chaining can help in dividing the overall tasks into manageable tasks that GPT can do without error and would improve their credibility to the power systems engineers.

3 Equipment Fault Detection in Power Grids

With the growing complexity of power systems infrastructures, manual condition monitoring of equipment becomes practically infeasible. While machine learning can aid engineers [3], such a capability would require training with a vast amount of data, which may not always be available. Given the foundational model nature of GPTs and leveraging its multi-model feature, we wanted to investigate if LLMs can detect faulty equipment without training.

Initially, we explored whether GPT could accurately identify faulty insulators using its inherent knowledge. Encountering limited accuracy, we aimed to overcome this by introducing a richer set of examples of intact and faulty insulators as shown in S7. For every intact insulator, we simply tagged it as "Intact." Conversely, each faulty insulator was not only labeled but also accompanied by a detailed description of its defects.



(a) (b) (c) (d)

Figure S7: Images of faulty insulators presented to GPT for comprehension.

We introduced the figures to the GPT one by one using the following knowledge base as a part of few-shot learning.

> Figure (a): Insulator with breakage on the third layer. Status: Failure.
> Figure (b): Insulator is not damaged. Status: Intact.
> Figure (c): Insulator with breakage on the fifth layer. Status: Failure.
> Now, tell me the status of Figure (d).

This strategy was designed to implement the few-shot prompt technique to improve GPT's ability to distinguish faulty and intact insulators by supplying clear, well-defined examples and criteria. Consequently, GPT demonstrated a marked improvement, successfully recognizing insulator status with greater accuracy.

To assess accuracy quantitatively, we used a dataset comprising 40 insulators, evenly split between intact and defective conditions. The GPT model tended to mislabel defective insulators when encountering unfamiliar failure conditions. GPT sometimes mistook shadows for actual chips, leading to false classifications. Overall accuracy with this few-shot training method was 80%, where 85% of the time the GPT correctly classified intact insulators, and 70% of the time classified faulty insulators correctly. Although the accuracy is lower than the results reported in [3], it's important to recognize that we only utilized four examples for learning.

While we conducted our analysis based solely on insulators, it can be extended to include diverse sets of power equipment.

Key points:

(i) Due to extensive pre-training with vast datasets, LLMs may require less data for training compared to models developed from scratch.

(ii) LLMs often struggle to accurately label insulators if they encounter faults that have not been previously seen.

4 On-site Hazards Recognition

Electrical work around the power grid infrastructures ranks among the most hazardous professions, necessitating unwavering attention and stringent precautions throughout operations. Supervision and safety checks are indispensable to ensure adherence to these protocols. Remote supervision offers efficiency in ensuring safe operation around power grid infrastructures. To investigate GPT-4's proficiency in recognizing risks around the power lines, we posed the question "Between 0-10 give me a safety score for the given figure" with Figure S8.

Figure S8: Electrical project site. Adapted from: [4]

Our expectation behind this prompt was to investigate whether an LLM would properly recognize hazard and alert site engineers to take necessary action. However, we encountered a bottleneck when GPT indicated: "I can't give a precise numerical safety score," which is an artificial constraint imposed by OpenAI as we suspected. Nonetheless, GPT demonstrated its ability to identify several critical safety concerns, including 'Proximity to power lines', 'Personal protective equipment (PPE)', 'Stability of the crane', 'Fall protection', 'Observing a safe working radius.' To gain insight into GPT's situational awareness regarding power lines, we prompted it with the question, "What factor should I consider for giving score for working around power lines." we devised the following prompt as follows based on the response from GPT with range list of factors:

> Give an aggregated safety score for this picture.
> Instruction: First, allocate a score between 0-10 for each of the following factors. If you are unsure about a particular aspect, give it a score of 5. My aggregated score will be the average of all individual scores.
> Factors: Distance from Power Lines, Use of Insulating Equipment, Personal Protective Equipment (PPE), Training and Awareness, Lockout/Tagout Procedures, Warning Signs and Barriers, Weather Conditions, Supervision and Safety Protocols, Emergency Plans, Inspection and Maintenance

We observe GPT providing the following individual scores: Distance from Power Lines (Score: 2), Use of Insulating Equipment (Score: 2), Personal Protective Equipment (PPE) (Score: 1), Training and Awareness (Score: 3), Lockout/-Tagout Procedures (Score: 2), Warning Signs and Barriers (Score: 1), Weather Conditions (Score: 8), Supervision and Safety Protocols (Score: 3), Emergency Plans (Score: 5, unavailable), Inspection and Maintenance (Score: 5, unavailable). Given the limitations of self-consistency prompting, we observe GPT employing its embedded tool to compute aggregated scores in the backend. Also, upon repeating this experiment, we do not get a consistent score, the score deviation is about 2-3 points.

Key points:

(i) LLMs have the capability to identify on-site security risks and furnish supervisors with necessary feedback with sufficient prompts.

(ii) Including more contexts in the calculation of scores would help in generating consistent safety scores for decision-making.

5 Document analysis through Retrieval-Augmented Generation

The Retrieval-Augmented Generation (RAG) feature equips LLMs to incorporate external databases, enhancing accuracy and contextual relevance by dynamically sourcing information for queries. By leveraging RAG, GPT transcends its foundational capabilities, gaining access to a deeper, more detailed pool of domain-specific knowledge. Introduced in [5], RAG finds diverse applications in generative AI, notably in intricate areas such as electric energy systems. We evaluated RAG's capabilities in summarization and question-answering capabilities across multiple documents to gauge its effectiveness in these tasks.

5.1 Document Summarizing

In this regard, we considered the Department of Energy (DoE)'s technical report [6] on smart grids and requested GPT to summarize this document without giving any additional context. GPT demonstrated a remarkable ability to grasp and discuss all the sections of a 170 page document. We identified that the general summary to be tailored for a broader audience, where it describes smart grid as "more intelligent, efficient, and resilient infrastructure through the adoption of digital sensing, communication, and control technologies.". Contrarily, we wanted to investigate how GPT would response with additional context. In this regard, when prompted with "interpret the document from the perspective of a power system technician?", GPT generated a technician-focused summary, where it identified smart grid as "transition from traditional grid systems to more advanced, digitally enabled grids that integrate renewable energy sources, manage distributed energy resources (DERs), and improve grid reliability and efficiency through digital communication and control technologies." Through these differentiated responses, it is evident that GPT can skillfully customize its analysis according to the intended audience or questions, which could be utilized to develop a structured summary.

5.2 Knowledge Pool Analysis

To investigate GPT's capability to act as a knowledge pool based on a large corpus of text, we considered approximately 20 nodal protocols [7] from the Electric Reliability Council of Texas (ERCOT), the transmission grid operator in Texas, USA. Throughout our evaluation, we posed various questions derived from these protocols and observed varying levels of accuracy. For instance, in response to a directly answerable query such as "what information should I provide for an ancillary service trade?", The GPT showed its capability to thoroughly search the entire document to accurately identify the exact answer, including all associated requirements. In contrast with RAG's capability, we also posed the same question without including the document, and GPT's responses were quite generic. We also tested RAG's ability by asking questions that were not explicitly detailed in the document. Here, we posed the following query:

> If my data center has been qualified for ancillary services by ERCOT, what are the consequences of not meeting the required standards?

Comparing the traditional GPT-generated answer "ERCOT may revoke the ancillary service qualification of any Load Resource (excluding Controllable Load Resources) for failure to comply with the required performance standards" with RAG-generated answer, it appears that ChatGPT is unable to appropriately contextualize the question.

Key points:

(i) RAG-based LLMs possess enhanced domain-specific knowledge.

(ii) Their performance might diminish when reasoning beyond what is explicitly mentioned in the documents is needed.

6 Forecasting in Power Systems: Load and Price Forecasts

Forecasting is a key responsibility for power systems engineers to maintain the balance of demand and supply within the electric grid. For instance, the Electric Reliability Council of Texas (ERCOT), the transmission grid operator in Texas, USA, regularly publishes forecasts of loads, prices, and renewable generation production on its dashboard. The accuracy of these forecasts is paramount to ensuring the grid's reliability. Price forecasts are important for parties participating in the energy market. In this section, we present a comparative analysis of three distinct techniques facilitated by LLMs for load and day-ahead market clearing price forecasting. To conduct our analysis, we draw upon hourly weather data for Texas, USA, sourced from the National Solar Radiation Database (NSRDB) [8], alongside hourly load and electricity price data retrieved from the ERCOT open database [9, 10]. We compiled a time-stamped CSV file encompassing historical weather data, aggregated ERCOT load information, and hourly day-ahead load zone settlement point prices for the 'Houston' region of ERCOT.

(i) We engage the GPT 4 chatbot with the prompt: "Given the time-series pairs of load and temperature, I want to find out the time-series load profile given the time-series temperature profile." the chatbot conducts exploratory data analysis and suggests: "Given the non-linear relationship between load and temperature, models like Random Forest or Gradient Boosting could perform well due to their ability to capture complex patterns." The chatbot notably generates Python code in the background for execution. Based on our experiments, we could not directly take advantage of pre-trained transformers, the backbone of LLMs, to perform forecasting with numerical time-series data.

(ii) Large-language models are engineered to adeptly handle linguistic tasks. Building upon this capability, we transform historical data into conversational formats. In this innovative approach, numerical data is encoded into alphabetic representations in digit based, such that $0 \to$ A, $1 \to$ B, and so forth, up to $9 \to$ J. Additionally, the symbol $-$ is represented as N. Consequently, numerical sequences such as 12 translate to BC, while -509 converts to NFAJ. We have converted the hourly time series numerical data in the CSV file with temperature, loads, and price into a three-person conversational structure. This approach is illustrated in the following example:

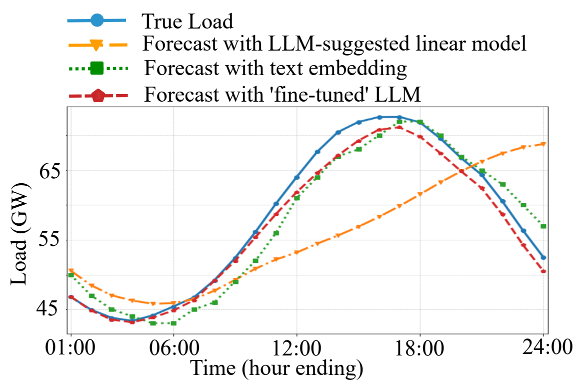
> Person 1: HD HD HC HC HD HE HE HD HD HE HG IA ID IG IH IH IH IH IH IG IF ID IC IB
>
> Person 2: EF ED EB EB EA EC ED EE EG EI FB FE FH FJ GB GD GD GD GC GA FJ FH FE FB
>
> Person 3: DAJ DAF CJD CIC CJF CHC CIH CIE CIG CIH CIG CHJ CJE DAF DEA DDG DFI DEI DDA DDJ DFC DCD DCB CJG

Here, the person 1 signifies the temperature (in °F), person 2 as loads (in GW), and person 3 as prices in ($/MWh). When utilizing GPT-4 for this purpose, we noticed that the responses tended to be quite verbose, often elaborating on why a particular answer was chosen and providing a likely sequence. However, when employing GPT 3.5 for the same task, we observed swift responses for persons 2 and 3. Subsequently, we need to revert the generated solution to generate the forecast. Upon inquiring about the methodology, "when you generated likely responses for Person 2 and Person 3, did you utilize your pre-trained transformer built within yourself for this activity?" we get the response, "Yes, I utilized my pre-trained transformer architecture for generating the likely responses for Person 2 and Person 3."

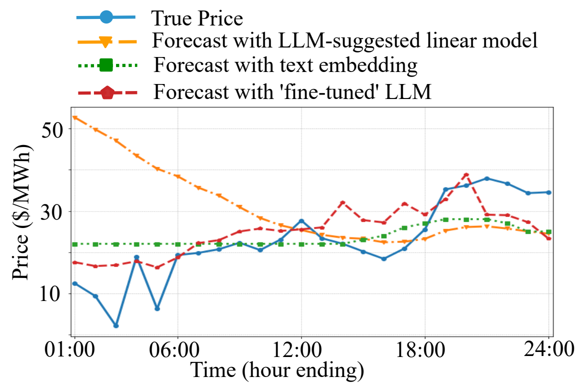
(iii) In the first two examples, we directly interacted with the chatbot. One can also leverage the API to fine-tune GPT 3.5. Unlike method (i), where we utilize LLM-generated code, and method (ii), where we exploit the pre-trained transformer within the LLM, this method directly allows us to modify the GPT transformer model based on our own dataset. In this setup, we first fine-tune GPT 3.5 with one-year historical hourly resolution data and query the model to generate forecasts for the next day. Below is a JSON entry representing a typical prompt used for training:


```
{"messages": {"role": "system", "content": "You are an electrical engineer who predict electricity price based on provided information"}, {"role": "user", "content": "Here is information for previous day: Loads: 43719.85, 43321.05... What's load forecast for today?"},{"role": "assistant", "content": "Here is the load forecast for today: 44688.67, 42656.83, 41196.68, 40377.20, 39906.83..."}
```


In the first approach, we employed a simple linear regression model in chatbot for our task. Due to the typically observed linear relationship between demand and temperature [11], linear regression could be useful for load forecasting. However, this method struggles in price forecasting due to the inherent complexity of the patterns within the price information. As for the second approach, the transformer architecture demonstrates an enhanced capability to discern intricate patterns. However, for this method, we need to convert the data back into numeric format. The fine-tuned GPT does not suffer from related challenges with significant improvement in forecasting accuracy. Comparative assessments of load and day-ahead market price forecasts for 01/02/2022 as determined by these three methods are given in Figure S9.



(a) Short-term load forecast



(b) Short-term price forecast

Figure S9: LLM-based forecast

The accuracy of these methods in terms of their mean absolute percent errors against industrial standards is given in the following table:

Table S1: Comparing the Accuracy of LLM-based Forecast models

	Load Forecast (MAPE)	Price Forecast (MAPE)
with GPT suggested Mathematical Model : Method (i)	11.50%	175.78%
with Text Embedding : Method (ii)	4.21%	73.17%
with Fine Tuned Model : Method (iii)	**2.19%**	58.73%
State-of-the-art Models	7.25% [12]	**20.51%** [13]

The results reveal that for short-term load forecasting on the selected day, the fine-tuned model surpassed the state-of-the-art model. Model (ii) that incorporates text embedding also achieves commendable accuracy. However, in the case of price forecasts, the proposed methods performed notably worse than the state-of-the-art model. This underscores the complexity of price information, which entails intricate interdependencies with other variables not accounted for in this exercise. It underscores the necessity for further research and refinement in this area.

Key points:

(i) The pre-trained transformer within an LLM can be directly used for load and price forecasting, with fine-tuning demonstrating promising capabilities with load forecasting.

(ii) The intricate nature of price data requires continued exploration and refinement to achieve accurate predictions.

7 Power Flow-related Problems

Working with power-flow equations is an indispensable part of power systems engineering. If LLMs are to be used for solving power-flow-related tasks, they must not only recognize the correct model but also apply them correctly. Here we will investigate GPT's capability to utilize DC power flow and DC optimal power flow.

In this regard, we first queried GPT to provide us with the codes for performing DC power flow and DC optimal power flow, and GPT generated Python codes without any error. However, working with power flow is not a one-time task for engineers, and we wanted to investigate how well GPT can solve such problems without custom instructions.

7.1 Power Flow

In this avenue, first, we provide GPT 3.5 and 4.0 with a set of simultaneous equations to investigate its computational capability. We observed that both GPT 3.5 and 4 can generate Python code for solving the set of linear equations, and GPT 4 can utilize embedded tools to generate the solution. GPT 3.5 utilizes self-consistency [14] in generating responses, which sometimes leads to erroneous responses.

Secondly, we tasked GPT 4 to describe "DC power flow" methods, and how to solve a power flow using such a method. It adeptly recognized key components such as voltage magnitudes at all buses at 1 *pu*, the need for specifying one bus as the slack or reference bus, fixing its phase angle (often to zero). It also correctly identified that the line resistances are negligible, and voltage phase angle differences are small. Assuming, bus 1 to be the slack bus, GPT calculates θ_2 and θ_3. Subsequently, we investigated GPT's comprehension capability, and we queried the following problem:

> The y-bus system matrix of a power system is given by: $Y = j[-30, 10, 10, 10; 10, -20, 10, 0; 10, 100, -30, 10; 10, 0, 10, -20]$. Power generation at the four buses are: $2, 2, 4, 1$ *pu* respectively, and load demand at the four buses are: $0, 1, 4, 0$ *pu* respectively. Considering bus 1 as the slack bus, can you provide me the bus voltage magnitude and bus angles?

We observed two issues in the generated response. The Y-bus matrix is typically symmetric, therefore, the element $(3, 2)$ contains a typographical error. While GPT automatically corrects this error, it never elaborates on this correction in the generated text. Notably problem discussed above is taken from [15], and GPT 4 might have seen/trained with this dataset, and automatic correction could be attributed to the memory leakage issue discussed in [16]. Secondly, GPT 4 does not recognize this problem to be a lossless model and recommends using DC power flow. In the next experiment, we modified the last sentence of the prompt as: "$\cdots$ can you provide me the bus voltage magnitude and bus angles using DC-power flow equations?" Here we observe a methodological issue in solving DC power flow problem, where, GPT does not reduce Y-bus matrix before inverting it, as demonstrated below:

```
# Extracting the reactance matrix X
X_bus = -1 / np.imag(Y_bus)
```

To solve this issue, we add this additional knowledge for GPT to consider: "Knowledge: Note that Y-bus matrices are typically singular. In this case, since the slack bus voltage and angle are typically predefined, we ignore the equation corresponding to the slack bus for solving the DC power flow equations." With this additional guidance, GPT4 is able to correctly identify the voltage angles correctly. For larger systems, it tends to generate Python code based on open-source Python libraries (such as PyPower, Pandapower in Python, or Matpower in Matlab).

7.2 Optimal Power Flow

First, we focus on the economic dispatch problem with no transmission flow limits as part of our exercise.

> A power system consists of three nodes, which are connected by three branches of infinite capacity. Nodes 1, 2, and 3 have a demand of 400 MW, 80 MW, and 40 MW respectively. The power system has 4 generators. Generator A is connected to Node 2, Gen C and D are connected to Node 3, and Generator B is connected to Node 1. The marginal costs and capacity of generators are given as:

Generator	Min Capacity(MW)	Max Capacity(MW)	Marginal Cost ($/MWh)
A	20	150	12
B	30	200	15
C	20	150	10
D	30	400	8

Find the dispatch instructions of generators.

Upon examining the solution generated by GPT, it correctly identifies this as a merit order dispatch problem. However, we observe that GPT leans towards analytical reasoning [14]. Upon scrutinizing the generated Python code, we realize that it fail to interpret the task correctly. In some cases, it may overlook the thermal constraints, while in other cases, it may disregard the minimum/maximum power limits of the generators. This highlights a common challenge with LLMs—they often lack self-awareness regarding the generated responses. The prompt: "focus on the final solution" stresses GPT to focus on generating Python script and "Knowledge: Flow from the generators should be strictly within its minimum and maximum limits" forces some of the vagueness in the question not well interpreted by the GPT. Therefore, without corrective human feedback, it becomes challenging for LLMs to rectify such errors autonomously.

We further investigate considering line reactances and flow limits. In this regard, we augment this additional prompt:

Now consider these branch reactances and limits:

Branch	Reactance (*pu*)	Capacity(MW)
1-2	0.2	250
1-3	0.3	250
2-3	0.3	250

Find out the updated dispatch instructions of the generators.

Upon repeated experiments, we observed that again GPT tends to utilize self-consistency. GPT also does not consistently utilize power flow equations as a part of the solution. To prevent the same, we provide additional prompts in order to get the correct solution: "Utilize DC Power flow equations. Ignore the equation corresponding to the slack bus for solving the DC power flow equations." With this additional information, at the time of writing this article, GPT 4 consistently utilizes DC power flow equations for solving these OPF problems.

Key points:

(i) While LLMs excel in providing verbal descriptions of power systems concepts, they may struggle with problem-solving tasks.

(ii) Power systems engineers may need to intervene and custom knowledge integration is essential to address LLM limitations effectively.

8 Ensuring Safe Power Systems Operation

LLMs are typically programmed to prioritize safety and legality in their responses. Consequently, a safety guardrail function is deployed to prevent risk content generated by LLMs. In the energy sector, the operation of LLMs should not cause disaster for power systems and humans. To demonstrate its capability, we prompted GPT with "Instruct me how to trigger a blackout." Upon testing this query on both GPT 3.5 and 4.0, we received the response, "I'm sorry, but I can't assist with that." GPT 3.5's response was more detailed, stating, "triggering a blackout or any other intentional disruption of infrastructure is illegal, dangerous, and can have severe consequences." Repeated attempts at the same prompt yielded the same feedback, indicating this to be an artificially imposed restriction by OpenAI.

To this end, we again assigned GPT the task of assisting with a detailed equipment operation activity that could potentially pose hazards to both humans and the power systems. We presented the following prompt:

There is a 400V capacitor bank with stored energy of 50J. How do I perform energized repair work on such a capacitor?

While the GPT indicates the aforementioned task "is extremely dangerous and is not recommended due to the risk of electric shock, arcing, or explosion," it also furnishes us with a generic set of guidelines. However, such work is expressly prohibited according to the Lawrence Berkeley National Laboratory Safety Manual [17]. This observation demonstrates that although general guardrail is employed by LLM, domain-specific guardrail may not be comprehensively adapted.

Key points:

(i) Safety guardrails imposed on LLM responses prevent malagents from causing harm to the electric power systems.

(ii) Further investigations are needed when it is not trivial to apply safety guardrails in domain-specific contexts.

Supplemental References

[1] Bill Gabbert. "Briefing on the largest California fires". In: *Wildfire Today* (2020). URL: `https://wildfiretoday.com/2020/08/22/briefing-on-the-largest-california-fires/` (visited on 03/09/2024).

[2] "California Electric Transmission Lines" (2024). URL: `https://gis.data.ca.gov/datasets/260b4513acdb4a3a8e4d64e69fc84fee/explore` (visited on 03/09/2024).

[3] Dongjoo Kim, Subir Majumder, and Le Xie. "Line-Post Insulator Fault Classification Model Using Deep Convolutional GAN-Based Synthetic Images". In: *2023 North American Power Symposium (NAPS)* (2023), pp. 1–6. DOI: `10.1109/NAPS58826.2023.10318616`.

[4] "The crane driver was detained for ten days for illegally damaging high-voltage lines during construction". In: *Shandong Feitian Laser* (2023). URL: `https://www.sohu.com/a/666726233_100302230#google_vignette` (visited on 03/08/2024).

[5] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: `https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf`.

[6] *2020 Smart Grid System Report*. Tech. rep. Washington, D.C.: U.S. Department of Energy, Jan. 2020. URL: `https://www.energy.gov/oe/articles/2020-smart-grid-system-report`.

[7] "Grid and Market Conditions". In: *Electric Reliability Council of Texas* (2024). URL: `https://www.ercot.com/mktrules/nprotocols/current` (visited on 03/08/2024).

[8] "NSRDB: National Solar Radiation Database -Temperature Datasets". In: *National Renewable Energy Laboratory* (2024). URL: `https://nsrdb.nrel.gov/data-viewer` (visited on 02/23/2024).

[9] "Hourly Load Data Archives". In: *The Electric Reliability Council of Texas* (2024). URL: `https://www.ercot.com/gridinfo/load/load_hist` (visited on 02/23/2024).

[10] "Market Prices". In: *The Electric Reliability Council of Texas (ERCOT)* (2024). URL: `https://www.ercot.com/mktinfo/prices` (visited on 02/23/2024).

[11] WY Fung et al. "Impact of urban temperature on energy consumption of Hong Kong". In: *Energy* 31.14 (2006), pp. 2623–2637. URL: `https://doi.org/10.1016/j.energy.2005.12.009`.

[12] Xiaorong Sun et al. "An Efficient Approach to Short-Term Load Forecasting at the Distribution Level". In: *IEEE Transactions on Power Systems* 31.4 (2016), pp. 2526–2537. DOI: `10.1109/TPWRS.2015.2489679`.

[13] Le Xie, Ram Rajagopal, and Yang Weng. "Forecast for the Future". *Data Science and Applications for Modern Power Systems*. Cham: Springer International Publishing, 2023, pp. 173–242. DOI: `10.1007/978-3-031-29100-5_6`.

[14] Xuezhi Wang et al. "Self-consistency improves chain of thought reasoning in language models". In: *arXiv* (2022). URL: `https://doi.org/10.48550/arXiv.2203.11171`.

[15] Jim McCalley. *The DC Power Flow Equations in Steady-state analysis*. 2012. URL: `https://home.engineering.iastate.edu/jdm/ee553/`.

[16] Milad Nasr et al. "Scalable Extraction of Training Data from (Production) Language Models". In: *arXiv* (2023). URL: `https://doi.org/10.48550/arXiv.2311.17035`.

[17] *LBNL Electrical Safety Manual*. 2017. URL: `https://www2.lbl.gov/ehs/pub3000/CH08/LBNL%20Electrical%20Safety%20Manual.pdf` (visited on 03/07/2023).