Can LLMs Model the Environmental Impact on SSD?

Mayur Akewar[1], Gang Quan[1], Sandeep Madireddy[2], Janki Bhimani[1]
[1]Florida International University, Miami, Florida, USA
[2]Argonne National Laboratory, Lemont, Illinois, USA
{makew001,gaquan,jbhimani}@fiu.edu,smadireddy@anl.gov

ABSTRACT

Environmental stressors such as temperature, humidity, vibration, and radiation can severely impact the performance and reliability of SSDs, particularly in edge, automotive, aerospace, and datacenter deployments. Capturing sensor data in the field and conducting accelerated lab experiments are challenging, as they are time-consuming, resource-intensive, and often destructive to hardware. Specialized setups, such as thermal chambers or vibration rigs, are also required, which is why few studies explore this area, and current storage management techniques like RAID, tiering, and deduplication do not consider environmental factors. Models to capture these impacts would open new research opportunities across various fields. However, accurately modeling these effects remains challenging due to, ① the limited availability of experimental data, ② the complex, domino-like impact of historical exposure, ③ the interrelated nature of environmental factors, such as temperature and humidity, which exhibit correlation, ④ different response of each type of NAND flash memory TLC, MLC, and SLC to environmental factors, and ⑤ the difficulty that analytical and simple machine learning models face in generalizing across devices, environments, and unseen combinations of stressors. We believe that LLMs may offer a transformative alternative to this complex problem, with embedded domain knowledge and reasoning capabilities, to facilitate prompt-based natural language interaction. We propose a hybrid framework that combines Chain-of-Thought prompting and Retrieval-Augmented Generation to guide LLMs using physical principles and prior experiments. It enables interpretable "what-if" analysis of SSD behavior under environmental changes. Our results show that the LLM can effectively model the impact of temperature, humidity, and vibration on SSD performance, producing tail latency and bandwidth predictions with minimal error. The code and data are available on GitHub at https://github.com/Damrl-lab/SSD_LLM.

HotStorage '25, July 10–11, 2025, Boston, MA, USA

ACM ISBN 979-8-4007-1947-9/2025/07
https://doi.org/10.1145/3736548.3737835

CCS CONCEPTS

• **Computer systems organization** → **Reliability**; • **Information systems** → **Flash memory**.

KEYWORDS

Solid State Drive, Large Language Model, Prompt Engineering, Retrieval Augmented Generation

ACM Reference Format:
Mayur Akewar[1], Gang Quan[1], Sandeep Madireddy[2], Janki Bhimani[1]. 2025. Can LLMs Model the Environmental Impact on SSD?. In *17th ACM Workshop on Hot Topics in Storage and File Systems (HotStorage '25), July 10–11, 2025, Boston, MA, USA.* ACM, Boston, MA, USA, 7 pages. https://doi.org/10.1145/3736548.3737835

1 INTRODUCTION

Solid-State Drives (SSDs) deployed in harsh environments, from edge IoT devices and automotive systems to aerospace and modern datacenters, often face environmental stressors such as extreme temperature, humidity [15], vibration [5], and radiation [23]. These stress factors can dramatically degrade SSD performance and reliability [3, 5, 9, 10, 12, 15, 24, 25, 27]. For example, high temperatures accelerate charge leakage in flash memory cells (speeding up wear-out), and high humidity can deteriorate capacitors and interconnects in the SSD controller [15]. Likewise, mechanical shocks or constant vibration have been shown to increase I/O latencies and reduce throughput in SSDs [5]. As SSDs increasingly operate outside climate-controlled settings, especially for edge applications such as automotive and aerospace applications, understanding, modeling the impact of these environmental conditions becomes a critical factor for preventing failures and ensuring stable performance for these applications.

To keep track of SSD performance by capturing sensor data that can reflect the SSD performance on the fly in such environments can be challenging. To conduct accelerated lab experiments to simulate these environmental conditions are

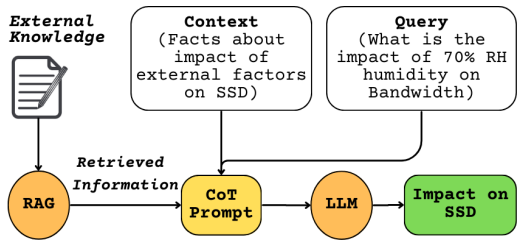


Figure 1: Environmental Modeling Process using LLM.

also difficult, as they requires specialized, resource-intensive setups, such as thermal chambers, vibration rigs, and radiation labs. In addition, these experiments can be time-consuming and often involve hardware stress that may damage the components being studied. Hence, there is very little literature [5, 15] on understanding such impacts. However, current works [5, 15] indicate that environmental factors can cause up to a 30% performance degradation. *Despite such a drastic impact, it is interesting to note that all of our existing storage management techniques, such as RAID, tiering, and deduplication, do not take the performance impacts of environmental factors into account.* Developing models that capture the influence of these stressors would not only enhance the resilience of SSDs but also open new research opportunities in adaptive storage solutions, predictive maintenance, and the development of higher-performing and more reliable storage systems for use in extreme environments.

However, accurately modeling environmental effects on SSDs remains extremely challenging. One fundamental obstacle is a lack of comprehensive data that are collected from well-controlled experiments that can capture different combinations of stressors are prohibitively expensive and time-consuming. Even a well-resourced study that tested over 100 new SSDs under various temperature-humidity scenarios could not reuse drives or exhaustively cover the space of conditions due to the high cost and effort required [5, 15]. Moreover, the impact of stress exposure is cumulative: a brief excursion to high temperature or humidity can cause latent damage (e.g., accelerated flash wear-out) that permanently alters subsequent device behavior [4, 6]. Environmental factors are also interrelated rather than independent: for instance, temperature and humidity swings often occur together in real deployments, with complex joint effects on SSD electronics [15]. SLC, MLC, and TLC NAND flash memory respond differently to environmental factors due to their distinct architectures. SLC, with its single bit per cell, is the most resilient to temperature, humidity, vibration, and radiation, while MLC and TLC, storing more bits per cell and are more vulnerable. Analytical formulas and machine learning models [4, 6, 11] built on field data or SMART [2] logs have limited ability to generalize across different devices and unforeseen combinations of stressors due to the complex nature of the problem. Straightforward models calibrated to one setting often fail in another, and they rarely provide insight into why a failure might occur [5, 15]. These limitations leave a gap between the insights gained from small-scale experiments and the need to predict SSD behavior under real-world environmental variability.

We believe that Large Language Models (LLMs) [8, 16, 20] offer a transformative alternative for modeling SSD behavior under environmental stress. Rather than relying on exhaustive sensor datasets and supervised learning that maps input features x to predict y by learning a function $y = f(x, \theta)$ parameterized by θ, *we treat this as a reasoning problem to be solved with knowledge and logic.* LLMs encode extensive domain information from technical literature and can synthesize cause-effect relationships described in natural language. By prompting an LLM with scenario-based questions (e.g., "How would a 15°C rise in ambient temperature affect the read latency of an SSD?"), we can obtain plausible predictions grounded in prior knowledge.

While LLMs offer strong capabilities, care needs to be exercised in their usage to overcome sensitivities and challenges due to the language interface. For instance, **Prompt dependence**: Prediction accuracy varies with prompt quality and clarity. **Lack of context**: Without detailed device or environment info (e.g., NAND type), the model may generalize incorrectly. **Domain bias**: LLMs may apply knowledge from mismatched scenarios if not guided properly. **Hallucinations**: In the absence of supporting data, LLMs can generate confident but false outputs, such as assuming temperature improves write endurance without evidence. To address above challenges, we design critical peripheral components of LLM to instruction-tune it and guide it to model the behavior of a domain expert, recalling experimental results (e.g. how humidity affected error rates in past studies) and known device physics (e.g. thermal effects on electron leakage) before drawing conclusions.

Key components of our framework are Chain-of-Thought (CoT) reasoning [22] to have the LLM work through intermediate steps, and Retrieval-Augmented Generation (RAG) [26] to inject relevant domain information and empirical data into the prompt (see Figure 1). For complex, domain-specific tasks like modeling environmental effects on SSDs, providing contextually relevant information is essential to improve both accuracy and trustworthiness. Our LLM-based framework enables interpretable 'what if' analyses of performance and likelihood of failure under various conditions, including previously unseen scenarios, and explains each prediction through step-by-step reasoning. This level of explainability is critical, especially when predictions guide high-stakes decisions, such as replacing SSDs in data centers due to potential slowdowns caused by prolonged low-temperature exposure during intense GPU workloads. In such cases, understanding

root causes is as important as the prediction itself. To the best of our knowledge, this is the first work to apply LLMs to simulate environmental stress effects on SSDs, filling a key gap by offering a proactive and explainable method for storage health assessment.

2 MOTIVATION

Why model environmental effects on SSDs? Modern applications demand dependable storage performance across a wide range of operating conditions. In edge computing and autonomous systems, SSDs might face extreme temperatures or continuous jostling; in data centers, airflow or cooling failures can cause unexpected thermal stress; and in aerospace or high-altitude deployments, radiation can bit-flip memory cells. Unmodeled environmental impacts can lead to sudden performance drops or device failures, catching operators off guard. For example, if an SSD's throughput silently falls by 30% under high vibration, a self-driving car's logging unit might lag, or a server might miss latency targets. Predictive modeling allows us to foresee these issues and take preventive action (like provisioning additional redundancy or cooling) before performance is lost or hardware is damaged.

Challenges of empirical testing. Understanding the environmental impact on SSDs through direct experimentation is resource-intensive and complex. Vibration studies [5] showed that even within vendor-specified limits, SSDs suffer significant I/O degradation, with tail latencies increasing by over 10%, and long-term exposure causing 30–45% performance loss and occasional failures. These studies required dozens of SSDs and over 120 hours of testing, with many drives rendered unusable [5, 15]. Temperature and humidity testing demands climate chambers and finely controlled sequences to observe shifts in latency and bandwidth [5, 15]. Radiation experiments are even costlier, requiring proton or gamma irradiation facilities [23], and have shown that both consumer and industrial SSDs eventually fail after reaching critical radiation exposure thresholds.

Opportunity with LLMs. LLMs (such as GPT-style [16] models) have demonstrated surprising abilities to perform reasoning tasks when given the right prompts. Their success in optimizing LSM Key-Value Stores [20] and analyzing HPC I/O traces [8] highlights their potential for tuning configurations and diagnosing storage system bottlenecks. They can combine factual recall with logical inference in a way that classic models have not been able to. For instance, an LLM can process a question like: "If an SSD with TLC NAND is running at 60°C and 80% relative humidity, how will its tail latency compare to 25°C and 50% humidity?", and it can attempt an answer by recalling relevant principles (e.g., "high temp lowers NAND read latency until thermal throttling or error rates increase; high humidity may affect electrical components, possibly increasing latency"). While an LLM might not natively know precise numbers, we can supply it with references or anchor points (i.e., ground truth of the known data from prior work) to guide it.

The LLM's strength is its ability to weave together multiple factors: it can consider "temperature increases electron mobility which might speed up reads, but humidity can increase capacitance and slow down writes, net effect might be...," all expressed in natural language reasoning. This natural language reasoning approach is compositional and interpretable. Instead of a cryptic formula, the model's intermediate chain-of-thought can articulate why it believes performance will change a certain way. This not only yields a prediction, but also a natural language explanation. For the systems community, this is valuable because it provides insight (almost like consulting an expert who explains their thinking) and helps debug or refine understanding of SSD behavior.

3 MODELING USING LLM

We treat predicting SSD behavior under different environmental conditions as a reasoning task for an LLM. This section explains how, by carefully designing prompts, providing reasoning feedback, and integrating external knowledge, we produce a model that is both accurate and interpretable.

3.1 Prompt Engineering

At the core of our framework is a prompt fed to LLM that encodes the scenario of interest. Prompt engineering is a critical step for model accuracy. The prompt consists of a structured combination of a question, few-shot examples, and the required answer format, ensuring accurate, context-sensitive predictions. We developed a unique prompt template that includes: (a) a description of the SSD's baseline state (e.g., type of NAND, internal structure of flash storage), (b) the environmental condition or change being applied (e.g., "temperature increasing from 30°C to 60°C while humidity held at 50%"), (c) the workload context (read-heavy, write-heavy, sequential or random I/O, etc.), and finally (d) an instruction to the LLM to predict the outcome in terms of performance and reliability impact. However, crafting effective prompts is challenging due to the complexity of modeling environmental impact and the sensitivity of LLMs to prompt structure. Poorly designed prompts may lead to irrelevant or incomplete outputs. To address these challenges, we leverage a key technique: **Chain-of-Thought (CoT) Prompting** [22] to get the LLM to systematically reason through the problem. Rather than directly outputting an answer, the model is instructed (either implicitly by the phrase "Think step by step" or by few-shot examples) to produce intermediate reasoning steps. We refined the prompt in an iterative loop: after each edit, we ran the model on the training set and computed the error (discussed in §4) between predicted and ground-truth latency and bandwidth. Only revisions that reduced

this error were kept, and we repeated the cycle until the mean-squared error fell below ten percent, which occurred after roughly twenty iterations. The resulting CoT template is therefore the minimal form that consistently meets our accuracy target on the training data. A snapshot of the CoT prompt is shown below.

An SSD with TLC NAND runs a workload while the temperature rises from room temperature to $50°C$*, with relative humidity fixed at 50%.*
Think step by step:
① How does increased temperature affect TLC NAND (e.g., electron mobility, latency)? ② Does $50°C$ risk thermal throttling or higher bit error rates in the controller? ③ Does 50% RH impact performance via circuit capacitance or signal delay? ④ Estimate the change in:
(a) Overall performance, (b) Tail Latency (90th–99.99th), (c) Bandwidth, (d) Failure likelihood.

Finally, the model would consolidate this into a direct answer: e.g., "Expected outcome: The 99th percentile write latency may increase by around 5–10% during the high-temp, high-humidity hour, owing to slowed controller response and more error-correction overhead. The SSD's error rate might increase (minor ECC corrections needed), but likely no immediate failure. Performance should recover once conditions normalize, though if humidity exposure is prolonged, some lingering high latency could persist." This chain-of-thought approach forces the LLM to consider each factor in turn and their interplay, much like how an expert would reason qualitatively. By splitting reasoning into steps, the model is less likely to skip over important interactions (like the combination of heat and humidity) and can draw on different pieces of knowledge at each step. [22] showed that CoT prompting can elicit better logical reasoning in LLMs, which we indeed find beneficial for this multi-factor analysis.

3.2 Retrieval-Augmented Generation (RAG)

To enhance the LLM's accuracy and factual grounding, we implement a RAG mechanism that dynamically supplies the model with relevant domain knowledge. This ensures the model's predictions are guided by real experimental results and specifications, reducing hallucinations and improving reliability, especially in complex or unfamiliar scenarios.

Our retrieval corpus is not limited to published experimental studies but also includes SSD datasheets, JEDEC reliability specs, technical reports, academic papers, internal engineering logs, and handbooks. Additionally, we encode foundational flash memory physics—such as temperature-induced threshold voltage drift, humidity-driven capacitance effects, and vibration-induced microcracks in PCBs. These entries are embedded and indexed into a vector store using dense representations. This allows efficient similarity-based retrieval [1, 7, 26] of relevant content.

At runtime, when a user poses a query (e.g., "rising humidity during workload execution"), the system retrieves semantically similar examples from the embedding store. These factual snippets are then added to the LLM's prompt before prediction. For example, if the temperature exceeds $55°C$ and ECC errors spike, the retrieved context might include prior observations like "TLC NAND showed 51% bandwidth gain at $50°C$" or "Perpendicular vibration causes ~15% tail latency increase in MLC SSDs." This grounding step enables online SSD health analysis by equipping the LLM to reason based on known behaviors, not just pretraining memory.

Why RAG Matters for Environmental Impact Modeling. RAG grounds LLM predictions in factual evidence that an potentially reduce hallucinations and improve reliability. Instead of relying solely on pretraining, it retrieves relevant specs, logs, historical information of the device, and experimental results, critical in multifactor scenarios where traditional models [13, 14, 17–19, 21] often fail. This enables the LLM to use more relevant real-world data without retraining or large labeled data.

4 EVALUATION

We evaluate our model by prompting with scenarios from published studies on temperature, humidity, and vibration, then comparing its predictions to ground truth for validation.

The dataset includes varying environmental conditions, capturing both immediate and residual impacts (after the SSD has been under stress for a period of time), and is constructed from a diverse set of scenarios based on real experimental setups from prior studies, including thermal-humidity transitions [15] and long-duration vibration exposure [5]. The dataset is divided into three parts, 20% of data goes for training, to prevent test contamination, 20% of scenarios was reserved exclusively as retrieval examples, while the remaining majority (60%) were used for testing. For instance, one prompt simulated a temperature rise from $22°C$ to $60°C$ at constant 50% RH to evaluate changes in tail latency. The scenarios covered both single-factor and multi-factor conditions, with variations in workload, temperature, humidity, and stress duration. We kept a limited amount of data in the training set to reduce the risk of memorization and overfitting. Contamination is further avoided because the contextual material given to the LLM using RAG contains only broad domain principles, for example, qualitative links between temperature and charge leakage or between vibration and latency, and very few examples (20% cases.). Since the answers required for each test query are absent from the prompt, the model must combine these generic hints with its own chain of thought to produce a prediction.

The final prompt is generated by first retrieving context through RAG based on the input query and then combining the retrieved context with the CoT prompt and the query

itself. At inference time we fed each test scenario to GPT-4o [16] using the final prompt template, and we recorded both the model's prediction and its reasoning. Every scenario was queried 10 times to measure consistency across runs.

4.1 Results

We evaluate our model by analyzing Tail Latency and Bandwidth under varying environmental conditions. We test our approach under two sets of conditions: *first, by varying Temperature and Humidity levels; and second, by applying two types of vibrations, parallel (horizontal) and vertical (perpendicular) vibrations.* Figure 2 presents the LLM's tail-latency and bandwidth predictions for a range of temperature and humidity settings, whereas Figure 3 shows the corresponding predictions under different vibration levels. We compare the predicted tail-latency percentiles (90, 95, 99, 99.9 and 99.99) and the predicted bandwidth changes with ground-truth measurements using the root mean square error (RMSE), defined as RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where y_i and $\hat{y}_i$ are the observed and predicted values. Lower RMSE values indicate closer agreement with reality and demonstrate that the LLM captures performance trends across diverse environmental conditions. Our model evaluates both runtime and post-runtime impacts of execution, specifically investigating the effects of environmental changes on SSD performance using LLM-generated insights.

The results, as shown in Figure 2, provide valuable observations on how variations in these environmental factors affect SSD resilience and responsiveness. • *With increased temperature and humidity levels, tail latency improves during running impact* (2a-I). • *When humidity decreases and temperature remains constant at a high level, tail latency improves during running impact* (2a-II). • *Increased humidity, while keeping the temperature constant at a high level, degrades tail latency during running impact* (2a-III). • *Varying humidity while maintaining a constant high temperature leads to tail latency degradation during post impact* (2a-IV). • *Varying temperature while keeping humidity constant at a high level results in tail latency degradation during post impact* (2a-V). • *SSDs exhibit higher bandwidth at elevated temperatures during running impact* (2b). • *A decrease in temperature leads to bandwidth degradation during running impact* (2c). Our results show low RMSE values under the various temperature and humidity settings: about 9% for tail-latency predictions and about 23% for bandwidth predictions.

Figure 3, illustrate the impact of vibrations on SSD performance, focusing on Tail Latency, under two vibration conditions and durations(running and post impact) providing following key insights: • *Tail latency degrades under the impact of parallel (3a) - (3c) and vertical (3d) - (3f) vibrations during short-term running (I), short-term post (II), and long-term (III) scenarios.* Figure (3g) illustrates the impact of parallel and vertical vibrations on bandwidth, focusing on both read and write operation performance. The key insights regarding bandwidth degradation under these conditions are as follows: • *Bandwidth degrades under the impact of both parallel and vertical vibrations for both read and write operations.* • *Bandwidth degradation is more significant under parallel vibrations compared to vertical vibrations.* Our result shows small RMSE values for the vibration scenarios: roughly 4.4 % for tail-latency under parallel vibration, 5.5 % under perpendicular vibration, and about 1.9 % for bandwidth.

The above results show that our model, with RAG and CoT prompts using LLM, is capable of capturing the complex performance effects of environmental factors, considering the impacts of historical exposures, interrelated factors, the internal type of flash memory, and both the duration and extent of exposure.

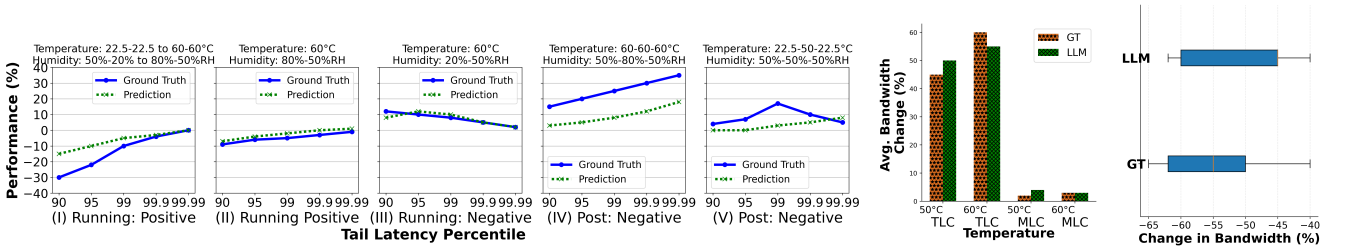
5 DISCUSSION

Our work introduces the first LLM-based model that connects AI reasoning with low-level storage performance under environmental stress, highlighting key implications, limitations, and future directions.

Generality vs. Specificity. LLMs demonstrate strong predictive capabilities for widely-used SSD types like TLC and MLC under standard conditions, proving their value in real-world modeling. However, for newer technologies like QLC/PLC or extreme scenarios (e.g., low pressure or high g-forces), accuracy may decrease due to limited prior data. Despite this, LLMs' reasoning ability enables zero-shot generalization, allowing knowledge from one hardware type to be applied to another without direct data, offering a significant advantage over traditional ML models, which require labeled data for each new configuration. By continuously enriching the retrieval corpus with emerging data, our LLM+RAG+CoT framework can evolve over time without the need for immensely expensive LLM retraining or finetuning.

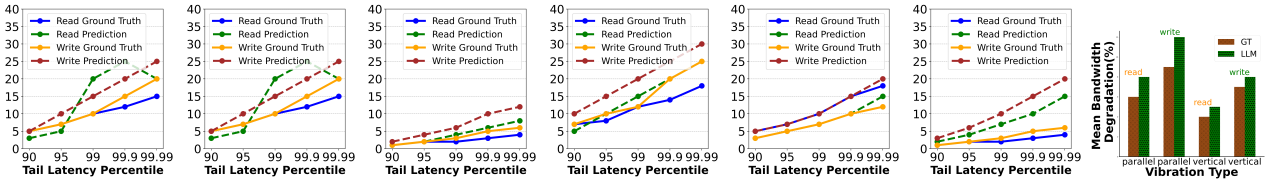
Component-level Modeling. Currently, the LLM treats the SSD as a single unit, aggregating behavior across its subsystems. However, SSDs consist of multiple components, such as NAND chips, controllers, DRAM, and connectors, each impacted differently by environmental stress. A more granular prompting strategy could break down the problem by having the LLM assess each component individually and then combine the results. Our early trials indicate that the model can reason in a modular fashion when explicitly guided. Formalizing this into a component-wise analysis framework could improve diagnostic accuracy and help pinpoint specific failure points. We plan to pursue it next.

Incorporating Domain Constraints. The flexibility of LLMs is a major strength, enabling them to reason across complex scenarios without rigid formulas, but this same flexibility can occasionally lead to outputs that defy physical



(a) Running and Post Impact (Positive or Negative) (b) Running: Positive (c) Running: Negative

Figure 2: Examining how varying temperature and humidity conditions affect Tail Latency and Bandwidth. (a) Impact on Tail Latency (b) and (c) Impact on Bandwidth



(a) I (b) II (c) III (d) I (e) II (f) III (g)

Figure 3: Impact of Parallel (a–c) and Vertical (d–f) Vibrations on tail latency. (g) shows bandwidth degradation.

or engineering limits. Unlike traditional models hard-coded with physical laws, LLMs benefit from guided reasoning. By embedding domain constraints directly into prompts or applying simple post-processing checks (e.g., capping throughput at interface limits), we can ensure outputs remain realistic. Importantly, the LLM's own chain-of-thought can be leveraged to catch logical inconsistencies, allowing the model to self-correct or flag questionable predictions. This combination of open-ended reasoning and structured grounding makes LLMs both powerful and trustworthy tools for modeling system behavior under environmental stress.

Trust and Verification. To address occasional hallucinations, we plan to generate test vectors for runtime checks and implement ensemble prompting, where multiple questions or models assess result consistency. If results converge, confidence in predictions will increase, enhancing the model's trustworthiness in practical deployments.

System-Level Impact of Accurate Predictions. Accurate forecasts of latency and bandwidth under varying temperature, humidity, and vibration help operators act early to meet service goals. For instance, if the model predicts a fifteen percent drop in bandwidth due to vibration, tasks can shift to stable nodes to maintain performance. Early warnings about rising latency also enable targeted cooling or redundancy, reducing energy use and hardware costs without relying on worst case planning.

Limitations. Despite promising results, some limitations remain. The system still depends on well-structured prompts, making full automation non-trivial, although it is scriptable. The model may miss edge-case failures or unknown interactions, especially without prior data in the knowledge base.

Future Work. To improve generalizability and robustness, we plan deeper architectural analysis and comparative evaluation. Future work includes ablation studies to better understand and optimize the interaction between RAG and CoT reasoning across diverse workloads and conditions (e.g., radiation, pressure, aging). We also aim to benchmark against fine-tuned open-source LLMs to explore trade-offs between generality, efficiency, and adaptability for real-world deployment.

6 CONCLUSION

This paper begins by posing the question: "Can LLMs model the environmental impact on SSDs?" We conclude by demonstrating for the first time that, with a properly designed peripheral framework using RAG and CoT, a single model can accurately predict the complex performance impacts of various environmental factors. The model generalizes across diverse scenarios and provides interpretable, natural language explanations for its predictions. This approach paves the way for AI-assisted storage system design, enabling proactive decisions over current reactive approaches.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their valuable feedback. This work was supported by the following NSF grants: CSR-2402328, CAREER-2338457, CSR-2406069, CSR-2323100, HRD-2225201, CCF-2119184 and CNS-2402327.

REFERENCES

[1] 2017. Faiss: A library for efficient similarity search. [Online]. Available: https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/.

[2] 2023. Self-Monitoring, Analysis and Reporting Technology (SMART). [Online]. Available: https://en.wikipedia.org/wiki/Self-Monitoring,_Analysis_and_Reporting_Technology.

[3] Saba Ahmadian, Farhad Taheri, Mehrshad Lotfi, Maryam Karimi, and Hossein Asadi. 2018. Investigating Power Outage Effects on Reliability of Solid-State Drives. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. [Online]. Available: https://doi.org/10.23919/DATE.2018.8342004.

[4] Jacob Alter, Ji Xue, Alma Dimnaku, and Evgenia Smirni. 2019. SSD Failures in the Field: Symptoms, Causes, and Prediction Models. In *SC '19: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. Article 75, 14 pages. [Online]. Available: https://doi.org/10.1145/3295500.3356172.

[5] Janki Bhimani, Tirthak Patel, Ningfang Mi, and Devesh Tiwari. 2019. What Does Vibration Do to Your SSD?. In *Proceedings of the 56th Annual Design Automation Conference 2019 (DAC '19)*. [Online]. Available: https://doi.org/10.1145/3316781.331793.

[6] Chandranil Chakraborttii and Heiner Litz. 2020. Improving the Accuracy, Adaptability, and Interpretability of SSD Failure Prediction Models. In *SoCC '20: Proceedings of the 11th ACM Symposium on Cloud Computing*. 120–133. [Online]. Available: https://doi.org/10.1145/3419111.3421300.

[7] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retriev*. [Online]. Available: https://doi.org/10.1145/3626772.3657834.

[8] Chris Egersdoerfer, Arnav Sareen, Jean Luca Bez, Suren Byna, and Dong Dai. 2024. ION: Navigating the HPC I/O Optimization Journey Using Large Language Models. In *HotStorage '24: Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems*. 86–92. [Online]. Available: https://doi.org/10.1145/3655038.3665950.

[9] Jun He, Sudarsun Kannan, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. 2017. The Unwritten Contract of Solid State Drives. In *Proceedings of the Twelfth European Conference on Computer Systems (EuroSys '17)*. 127–141. [Online]. Available: https://doi.org/10.1145/3064176.306418.

[10] Ying He, Shuwen Liang, and Song Fu. 2024. Impact of Environmental Factors on Flash Storage Performance in Autonomous Vehicles: An Empirical and Analytical Study. In *2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*. [Online]. Available: https://doi.org/10.1109/MOST60774.2024.00029.

[11] Jun Li, Bowen Huang, Zhibing Sha, Zhigang Cai, Jianwei Liao, Balazs Gerofi, and Yutaka Ishikawa. 2020. Mitigating Negative Impacts of Read Disturb in SSDs. *ACM Transactions on Design Automation of Electronic Systems (TODAES)* 26, 1, Article 3 (2020), 24 pages. [Online]. Available: https://doi.org/10.1145/3410313.

[12] Qiang Li, Hui Li, and Kai Zhang. 2019. A Survey of SSD Lifecycle Prediction. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. [Online]. Available: https://doi.org/10.1109/ICSESS47205.2019.9040759.

[13] Shuwen Liang, Zhi Qiao, Jacob Hochstetler, Song Huang, Song Fu, Weisong Shi, Devesh Tiwari, Hsing-Bung Chen, Bradley Settlemyer, and David Montoya. 2018. Reliability Characterization of Solid State Drives in a Scalable Production Datacenter. In *2018 IEEE International Conference on Big Data (Big Data)*. [Online]. Available: https://doi.org/10.1109/BigData.2018.8622643.

[14] Stathis Maneas, Kaveh Mahdaviani, Tim Emami, and Bianca Schroeder. 2020. A Study of SSD Reliability in Large Scale Enterprise Storage Deployments. In *Proceedings of the 18th USENIX Conference on File and Storage Technologies (FAST '20)*. [Online]. Available: https://www.usenix.org/conference/fast20/presentation/maneas.

[15] Adnan Maruf, Sashri Brahmakshatriya, Baolin Li, Devesh Tiwari, Gang Quan, and Janki Bhimani. 2022. Do Temperature and Humidity Exposures Hurt or Benefit Your SSDs?. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. [Online]. Available: https://doi.org/10.23919/DATE54114.2022.9774582.

[16] OpenAI. 2023. *GPT-4 Technical Report*. Technical Report. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf.

[17] Jay Sarkar and Feng-Bin Frank Sun. 2015. Reliability Characterization and Modeling of Solid-State Drives. In *2015 Annual Reliability and Maintainability Symposium (RAMS)*. [Online]. Available: https://doi.org/10.1109/RAMS.2015.7105166.

[18] Bianca Schroeder, Raghav Lagisetty, and Arif Merchant. 2016. Flash Reliability in Production: The Expected and the Unexpected. In *Proceedings of the 14th Usenix Conference on File and Storage Technologies (FAST '16)*. 67–80. [Online]. Available: https://www.usenix.org/conference/fast16/technical-sessions/presentation/schroeder.

[19] Swamit Tannu and Prashant J. Nair. 2023. The Dirty Secret of SSDs: Embodied Carbon. *ACM SIGENERGY Energy Informatics Review* 3, 3 (2023), 4–9. [Online]. Available: https://doi.org/10.1145/3630614.3630616.

[20] Viraj Thakkar, Madhumitha Sukumar, Jiaxin Dai, Kaushiki Singh, and Zhichao Cai. 2024. Can Modern LLMs Tune and Configure LSM-Based Key-Value Stores?. In *HotStorage '24: Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems*. 116–123. [Online]. Available: https://doi.org/10.1145/3655038.3665954.

[21] Yufei Wang, Xiaoshe Dong, Xingjun Zhang, and Longxiang Wang. 2019. Measurement and Analysis of SSD Reliability Data Based on Accelerated Endurance Test. *Electronics* 8, 11 (2019), 1357. [Online]. Available: https://doi.org/10.3390/electronics8111357.

[22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*.

[23] Edward P Wilcox, Adia Wood, Gregory Allen, Martin Carts, and Megan Casey. 2024. Solid State Drive Radiation Assurance With Active Testing. In *IEEE Nuclear and Space Radiation Effects Conference (NSREC)*. [Online]. Available: https://ntrs.nasa.gov/citations/20240008528.

[24] Kan Wu, Andrea Arpaci-Dusseau, and Remzi Arpaci-Dusseau. 2019. Towards an Unwritten Contract of Intel Optane SSD. In *Proceedings of the 11th USENIX Conference on Hot Topics in Storage and File Systems (HotStorage '19)*. 3. [Online]. Available: https://www.usenix.org/system/files/hotstorage19-paper-wu-kan.pdf.

[25] Gala Yadgar, Moshe Gabel, Shehbaz Jaffer, and Bianca Schroeder. 2021. SSD-Based Workload Characteristics and Their Performance Implications. *ACM Transactions on Storage (TOS)* 17, 1, Article 8 (2021), 26 pages. [Online]. Available: https://doi.org/10.1145/3423137.

[26] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. In *arXiv*. [Online]. Available: https://arxiv.org/html/2409.14924v1.

[27] Aviad Zuck, Philipp Gühring, Tao Zhang, Donald E. Porter, and Dan Tsafrir. 2019. Why and How to Increase SSD Performance Transparency. In *HotOS '19: Proceedings of the Workshop on Hot Topics in Operating Systems*. 192–200.