# Optimizer-Driven Techniques for Retuning Prompts on LLM Backend with Migration across Models for Healthcare and Cyber Security Operations

Hassan Rehan[1,*], Zane Zimmerman[2], Jamshaid Iqbal Janjua[3]
[*] *Corresponding Author: hassan.rehan202@ieee.org*

[1] *Department of Computer & Information Technology*
*Purdue University*
West Lafayette, IN, United States

[2] *Mays Business School*
*Texas A&M University*
College Station, TX, United States

[3] *Al-Khawarizimi Institute of Computer Science (KICS)*
*University of Engineering & Technology*
Lahore, Pakistan

***Abstract*—The relocation of the workflows that are provided by one model to another is becoming more and more frequent, yet it remains particularly difficult in the healthcare and cyber security operations where the accuracy, safety, privacy, and cost factors are off the table. We present an optimizer-based design where migration is a black-box, constraint-sensitive, optimization problem, automatically re-optimizing existing instructions to a new LLM backend, and not based on internal model information. The method combines a meta-optimizer, transformation operators, and warm-starting using history evaluation traces to accelerate adaptation and minimize the cost of trials. We test the framework on popular LLM backends, such as OpenAI GPT-4, Anthropic Claude, and Google Gemini, on operationally motivated healthcare tasks and cyber security tasks, such as clinical note summarization and triage, compliance-oriented question answering, incident report analysis, alert triage, and threat intelligence extraction. The proposed approach is effective, better at adaptation and the cost of operations in various settings, such as compared to manual retuning, zero-shot transfer, and random search, and constrained to domains and safety. The findings indicate that optimizer-based migration is a viable and scalable approach to ensuring stable LLM performance when quickly changing the backend in production environments.**

***Keywords*—Prompt Migration, Optimizer-Driven Retuning, LLM Backends, Healthcare Operations, Cyber Security Operations**

## I. INTRODUCTION

Large language models (LLMs) are being integrated into pipelines of operation where they are used to summarize, triage, extract, and answer questions in areas that require great reliability. LLMs are applied in healthcare operations to organize clinical notes, aid intake and triage, and facilitate compliance-oriented question answering. They are utilized in the analysis of incident reports, triage alerts, and to derive threat intelligence artifacts like indicators of compromise and tactics in cyber security operations. The benefits of these deployments are high productivity, but these deployments come with high non-functional requirements: outputs are to be accurate, auditable, privacy-preserving, safe, and cost-bounded. Recent publications emphasize the increased use of LLMs in healthcare and the commensurate risks of hallucinations, information privacy, and critical errors in their operation and safety [1], and the rapidly expanding application of LLMs to cyber defense processes where dual-use safety margins and operational correctness are critical [2].

An actual difficulty that has become apparent with the adoption of production is the problem of backend migration: organizations often change between providers or model family of one LLM to another because of price changes, latency demands, policy limitations, data residency, availability, or perceived safety posture. Although the upstream application logic and task definition may be the same, the post-migration behavior of an LLM-driven workflow may vary dramatically. The slightest changes in teaching after, formatting compliance, refusal behavior, verbosity and risk sensitivity tend to lead to the failure of prompts that were previously stable. This may take the form of missing structured summaries, falsified clinical information, or data leakage in healthcare, or unstable refusal constraints, lost high-severity cases, or poor extraction fidelity in cyber security. Previous studies have highlighted the fact that, the safety, security and privacy threats of generative systems are not accidental and that deployment must be under systematic controls, not just ad hoc immediate amendments [3], [4].

Migration is now often dealt with by manual prompt retuning, trial-and-error, and patching. This method is time-consuming, expensive, and not easily auditing, particularly where there are many prompts in a workflow, and the budget to evaluate is small. In addition, manual tuning is less likely to optimize operational constraints including cost, latency and violation rates, and is more likely to emphasize utility in tasks it is applied to. This is especially

troublesome in fields like healthcare and cyber security, where any prompt that moves a accuracy statistic up, but causes a privacy leakage or unsafe advice, is operationally invalid. Simultaneously, the optimization community has devised principled frameworks to do black-box and constrained optimization, including those capable of dealing with costly evaluations and constraint violations, which are applicable since an LLM backend is essentially a black-box function between prompts and inputs to outputs and costs [5]. Nonetheless, these concepts are seldom modeled as a migration model based on the specific safety/compliance requirements of healthcare and cyber security processes.

In this paper, an optimizer-based prompt migration framework is proposed, which considers migration to the backend as a constraint-sensitive black-box optimization problem. Rather than using internal model information, the framework re-tunes an existing prompt package to a new backend based on measured results on evaluations of the target domain. It combines (i) a meta-optimizer that suggests candidate prompt variants with a fixed trial budget, (ii) a library of structured transformation operators that encode common repair moves and domain-specific constraints, and (iii) warm-starting with historical evaluation traces to speed up convergence and reduce cost of trials. The framework is also intended to be auditable and it explicitly maximizes task utility in concert with safety, privacy, and operational limits, in line with the demands highlighted in deliberations of reliable and safe AI implementation [4], [6].

We test the approach on various popular LLM backends on healthcare and cyber security tasks that represent real operational tasks: clinical note summarization and triage, compliance-focused question answering, incident report analysis, alert triage, and threat intelligence extraction. Findings indicate that migration based on optimizers is always effective compared to zero-shot transfer, random search in prompt variants, and manual retuning in both areas, where it is more effective and fewer violations during deployment block operations but also regulates operational cost.

## II. Literature Review

The use of large language models (LLMs) in the healthcare and cyber security process is becoming more popular in areas where accuracy, safety, and accountability are crucial. They aid in healthcare with summarization, question answering, and clinical documentation and in cyber security with threat intelligence extraction, alert triage and incident analysis. But they may produce different results on prompts, tasks and backends posing risks of hallucination, privacy leakage, unsafe recommendations, and weak reproducibility [7], [8].

The previous studies in the field of healthcare demonstrate that the use of LLM can lead to decreased documentation and enhanced access to information, yet the issues of reliability, bias, privacy, auditability, and regulatory compliance are still essential factors to be considered [7], [9]. In cyber security, LLMs have potential to summarize reports, analyze threats, and extract indicators, but also encounter adversarial prompting, unstable refusal behavior, and dual-use risks [8], [10]. On a broader scale, reliable deployment involves the optimization of utility, as well as confidentiality, robustness, refusal correctness, and policy adherence [11], [12].

Timely engineering experiments indicate that wording, structure, exemplars, and formatting have a strong impact on performance and that automated optimization can be more effective than conventional prompt design with small budget constraints [13], [14]. The requirement of systematic and dynamic retuning to system transitions is also supported by related research in the context of contextual understanding, resource-conscious system design, model sensitivity, forecasting optimization, and backend efficiency [15], [16]. This is supported by the fact that even minor design decisions can have a significant impact on behavior and performance [17], [18] and that backend-level efficiency is also significant in deployment contexts [19].

Therefore, the process of backend migration can be regarded as black-box, constraint-based optimization problem due to the fact that developers can usually only see outputs, costs, and violations but not the details of the internal model. The black-box and constrained optimization studies prove this perception, and are effective when budgets are small, assessments are costly, and hard constraints exist on safety [20]. In general, existing literature emphasizes the usefulness of LLMs, the dangers of high-stakes deployment, and prompt sensitivity to backend changes, but fails to provide a coherent structure to perform prompt migration across models in an efficient and safe manner.

## III. Methodology

The content of this paper is a practical reliability issue: a deployed and tested LLM workflow can fail on a new provider or model family, even with the same task and data, despite the fact that the workflow is already deployed and tested. This degradation can be particularly expensive in healthcare and cyber security operations since the outputs may be triaged, documented, complied or respond to incidents, where errors can cause patient harm, regulatory liability, or threats being missed. Our proposed approach as mentioned in Fig.1, does not view migration as a rewrite of instructions, but as a disciplined, quantitative retuning procedure, explicitly maximized to be effective, with

operational constraints imposed on safety, privacy, cost, and latency.

Let Ms be the source LLM backend and Mt be the target backend. For each operational task, we assume a dataset D containing N items, where each item is an input-output pair (xi, yi). The input xi may be a clinical note, discharge summary, symptom narrative, or compliance prompt in healthcare, and may be an incident report, SOC alert description, log excerpt, or threat intel paragraph in cyber security. The expected output yi can be a gold structured summary, a triage label, a set of extracted entities, or a reference answer (when available). A prompt p is treated as a complete instruction artifact, including system-level constraints, the task template, output schema requirements, optional exemplars, and refusal policy. For a decoding configuration θ (e.g., temperature, max tokens), the target model produces an output yhat = Mt(x; p, θ). Migration begins from an existing prompt p0 that performed acceptably on Ms, and it aims to produce a migrated prompt p* (and optionally θ*) that remains reliable on Mt.

Instead of assuming access to internal model details, we adopt a black-box setting: the only interface to Mt is query-and-response, and the only signals we can optimize are measured task metrics, cost/latency traces, and safety or privacy checks on model output. This premise applies to the majority of real migrations to OpenAI, Anthropic, and Google backends where no internals are available and where the enterprise governance tends to limit fine-tuning or proprietary tuning pipelines.

Our objective is to maximize one scalar score which is a combination of task effectiveness and penalties on violating operational constraints. To any candidate prompt and decoding pair (p, θ) we compute a score:

J(p, θ) = Ubar(p, θ) − penalty_cost − penalty_latency − penalty_safety − penalty_privacy.

Here Ubar(p, θ) is the average utility on an evaluation batch (for example ROUGE-like overlap for summaries, F1 for extraction, macro-F1 for triage). The punishments are calculated based on quantifiable amounts. Assuming Cbar is the average cost per item (e.g., token cost), and Bcost is the cost budget, then we can define penalty cost as λ1 1 max(0, Cbar Bcost). and penalty latency as λ2 1 max(0, Tbar Btime). Violation rates (e.g., Vsafe and Vprivacy) are proportional to the safety and privacy penalties, which are penalty_safety = 3 3 Vsafe and penalty_privacy = 4 Vprivacy. This minimal structure is purposeful, because it is backend-agnostic, simple to audit and tolerant of noisy black-box tests.

The terms of safety and privacy take precedence in the healthcare field due to the clinical risk and sensitive content treatment. In cyber security, safety is scoped to dual-use limits (e.g. denying exploit generation and yet permitting defensive advice), and privacy is congruent with redaction of sensitive identifiers like user names, internal hostnames or proprietary incident information. In both domains, the goal specifically discourages prompts that score well, but are operationally inadmissible, such as by leaking information or giving unsafe advice, being excessively verbose, or imposing overhead on resources.

The proposed framework iteratively improves prompts for Mt using three coupled mechanisms. First, a meta-optimizer selects which candidate prompt to try next under a strict evaluation budget. Each iteration proposes a candidate (pk, θk), runs Mt on a few representative items, calculates J(pk, 0k), and changes its selection policy. The noise in outcomes and the high cost of model calls are the motivations to make the optimizer budget-aware, only promising candidates are promoted to larger evaluations. This renders the process appropriate in actual migrations where cost management is a must. Second, the space of candidate searches is determined by the transformation operators on p0. The operators are designed, verifiable updates that respond to known cross-backend failure modes, but are specifically customized to healthcare and cyber security. In medical care, operators focus on schema fidelity (e.g., “Problem List, Medications, Allergies, Plan, Red Flags) abstinence where evidence is not available, and language that is privacy-sensitive and does not replicate identifiers.

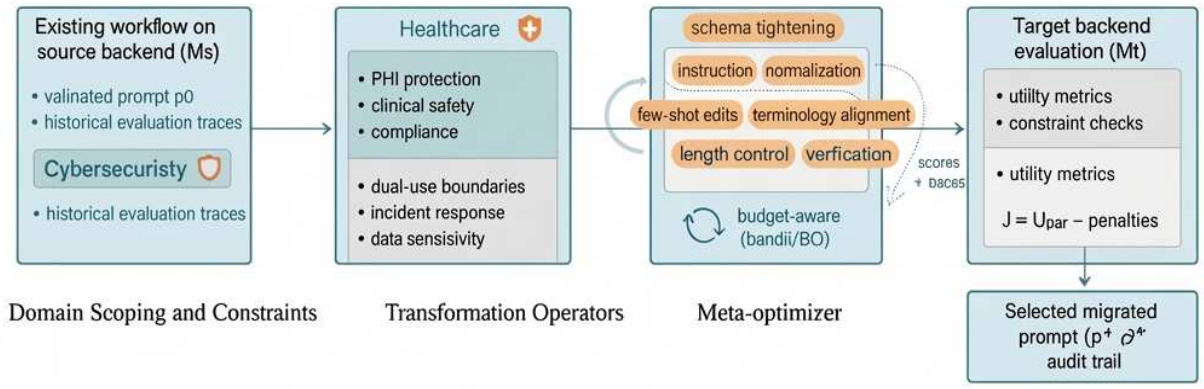


Fig.1 Optimizer-Driven Prompt Migration Framework for Healthcare and Cyber Security Specialization

Operators apply indicators of compromise extraction schema, tactics and techniques extraction schema, impact summary extraction schema, and recommended containment steps in cyber security, and incorporate rules against offensive content in incident response and produce terse and actionable incident response outputs. The other operators found in both fields are length control, formatting strictness and terminology normalization since the changes made at the backend often change the extent to which models adhere to formatting or the meaning of ambiguous instructions.

Third, warm-starting relies on historical evaluation traces recorded on Ms to speed up migration on Mt. The trace record contains the inputs that were hard cases, the common patterns of errors (lacking fields, hallucinating facts, refusal failures, schema violation), and which prompt variants have historically corrected them. Warm-starting is applied to (i) prioritize transformations aimed at the most frequent observed domain failure modes, and (ii) front-load evaluation batches with risky items, like clinically ambiguous notes (healthcare) or adversarial prompt injection and dual-use queries (cyber). This helps to minimize the number of trials to achieve a compliant and effective prompt on Mt.

Algorithm.I Optimizer driven prompt framework for Healthcare and Cybersecurity

```
Initialize:
  p_best = p0, J_best = J(p0)
  Budget = max_trials or max_total_cost
Repeat until Budget exhausted:
  1) Propose candidate (pk, θk)
    - choose healthcare- or cyber-specialized operator set
    - optionally warm-start from historical traces
  2) Small-batch test on Mt
    - run on risk-heavy items
     (PHI-adjacent notes or high-severity alerts)
  3) Apply domain gates
    Healthcare gates:
      - PHI leakage check
      - unsafe medical advice check
      - schema completeness check
    Cyber gates:
      - dual-use refusal check
      - extraction schema validity
      - sensitive identifier leakage check
  4) Score candidate
    - compute Ubar, Cbar, Tbar, Vsafe, Vprivacy
    - compute J(pk, θk) = Ubar - penalties
  5) Promote or discard
    - if gates fail: discard or heavily penalize
    - if promising: evaluate on larger batch (multi-fidelity)
  6) Update best feasible prompt
    - if J(pk, θk) > J_best and constraints satisfied:
       p_best = pk, J_best = J(pk)
Return:
  migrated prompt p* = p_best and decoding θ*
  plus audit trail: operators applied, scores, costs, violations
```

Multi-fidelity protocol in the Algorithm.I, is used to evaluate prompt candidates to trade between statistical confidence and cost. Initial pilot tests provide candidates with small batches that are domain-risk, e.g., PHI-proximate clinical text or high-severity SOC alerts, since failure on these items is more severe than marginal gains on easy items. Applicants who demonstrate early indications of breaching safety or privacy limits are eliminated or severely punished to prevent squandering funds. The feasible candidates are then promoted to larger batches to get more stable utility estimates and a final selection made on a held-out validation split with a final report on a test split that is never used in optimization

Healthcare assessment refers to factuality, coverage, and safe language. The quality of summarization is not only based on overlap scores but also the clinically relevant content checks, including the presence of key problems, medications, and follow-up actions when they are supported by the input. The classification metrics used to evaluate triage tasks are macro-F1, and in case confidence scores are generated, the calibration error is also evaluated to prevent overconfident unsafe results. Privacy checks confirm that the prompts make the model repeat sensitive identifiers or produce faked patient information.

Cyber security assessment focuses on extraction fidelity, and decision usefulness. The extraction tasks are graded using the precision, recall, and F1 on various fields such as IPs, domains, hashes, affected hosts, time ranges, and suggested actions. Triage is tested on macro-F1 over severity classes and error analysis which is particularly punitive of missed cases of high severity. Safety checks consider the correctness of refusal to offensive requests and still provide defensive advice (e.g., hardening measures, safe form of a detection query). Privacy checks are concerned with the redaction of internal identifiers and timely disclosing sensitive operational information.

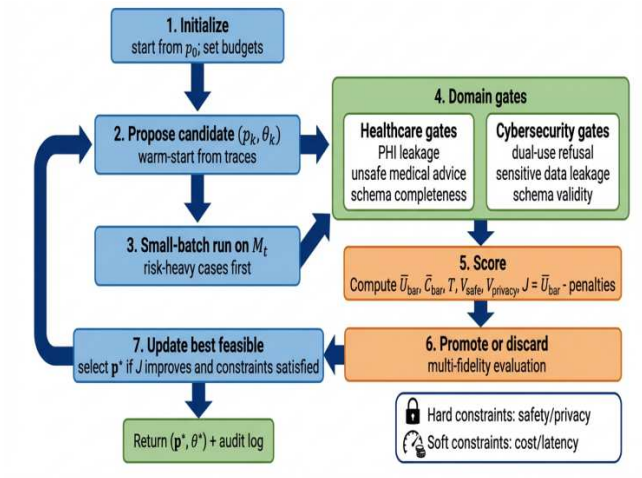


Fig.2 Budgeted optimization loop with domain-specific safety and privacy gates

In order to measure benefits, we compare the suggested framework with a set of potential options that are practical and frequently used by teams in the context of migration as shown in Fig.2. Zero-shot transfer uses p0 with no changes Mt. Manual aims to retune with human edits based on common prompt-engineering heuristics. Random search samples do not use an adaptive selection strategy to transformations. We also perform ablations on (i) warm-starting, (ii) constraint-aware scoring (eliminating penalties and gates), and (iii) domain-specific sets of operators (only using generic transformations). This is done by comparison at the same evaluation budgets so that gains can be viewed as a change in sample efficacy and cost of operation, rather than just an increment in the volume of the trials.

The technique is meant to be audited in high-stakes regulated situations. Applications obtain all the prompt text, transformation sequence, decoding parameters, evaluation batch identifiers, results, metric breakdowns, cost, and latency, and safety or privacy flags. The last artifact contains the migrated prompt p, and 0, as well as a domain-specific compliance summary that justifies why the chosen prompt is suitable for healthcare, & cyber security.

## IV. Results

The reports of the results are presented in two categories: healthcare and cyber security due to the different deployment success criteria in the two fields. In healthcare a migrated prompt can only be considered acceptable when utility is also maximized and PHI leakage and unsafe advice are kept near zero as these are deployment-blocking violations. In cyber security, success demands better extraction and triage performance and stabilization of dual-use safety behavior, i.e. greater refusal correctness and less over-refusal to allow analysts to still get legitimate defensive advice. The healthcare outcomes are summarized in Fig.3 and the cyber security in Fig.4.

The retuning proposed (optimizer-driven) yields better clinical utility (summarization quality, schema completeness, triage Macro-F1, and abstention accuracy) and significantly reduces deployment-blocking violations (PHI leakage and unsafe advice) compared to zero-shot transfer, manual retuning and random search. The healthcare outcomes show that there is no single quality measure that is being improved. The framework enhances schema completeness, necessary to downstream clinical review, and risk metrics that have a propensity to regress during backend swaps. This confirms the argument that migration needs to be treated as constraint-optimal in controlled clinical processes. The offered approach enhances the operational utility (incident analysis and threat intel extraction F1, alert triage Macro-F1, and high-severity recall) and stabilizes dual-use safety behavior, boosting refusal correctness and decreasing over-refusal to allow defensive workflows to continue after the backend migration.

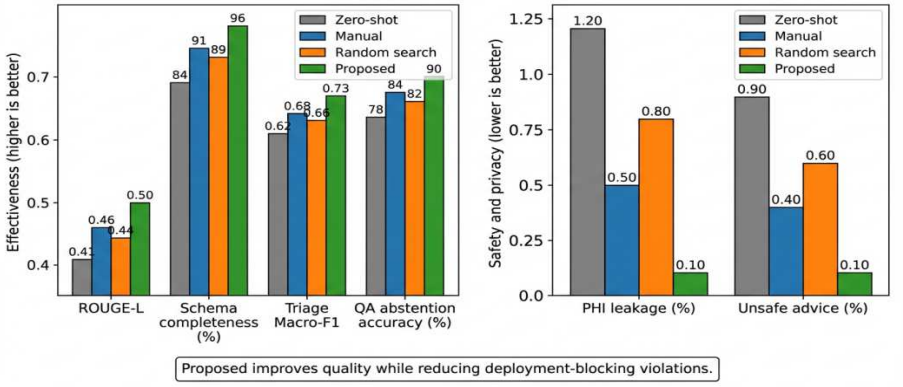


Fig.3 Healthcare Migration Results

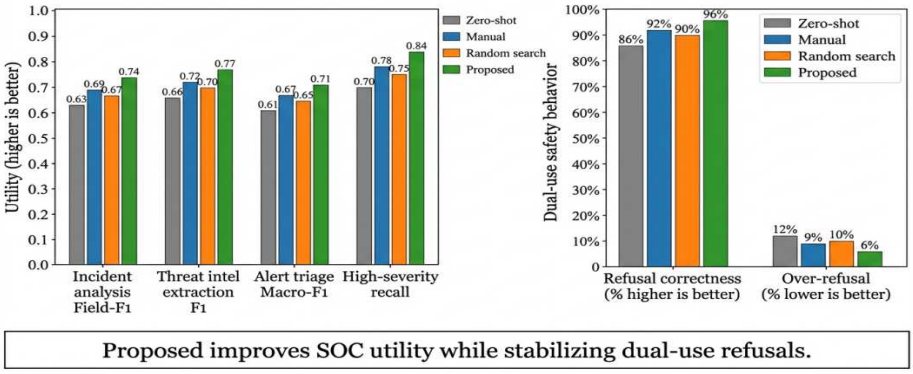


Fig.4 Cyber Security Migration Results

The outcomes of the cyber security reveal that the framework enhances performance that is analyst-relevant especially high-severity recall, and at the same time, it enhances refusal correctness. This combination is hard to achieve by manual retuning alone since changes to the backend may alter refusal thresholds and formatting behavior in a manner inaccessible to ad hoc editing, & instead requiring systematic, measured retuning

## V. Conclusion

The current paper introduced an optimizer-based model of timely migration between LLM backends of healthcare and cyber security workflows. The proposed approach can retune existing prompt packages in a systematic manner without accessing internal models by formulating migration as a black-box, constraint-aware optimization problem. The model integrates a meta-optimizer, designed transformation operators and warm-starting with historical traces to enhance adaptation effectiveness with small evaluation budgets. The experimental findings in various backend migration conditions revealed that the suggested method is more effective than zero-shot transfer, hand retuning, and random search. It enhances the effectiveness of the tasks in both healthcare and cyber security work, minimizes violations of safety, privacy, and deployment blocks, and it does it more cost-effectively. These results suggest that the problem of backend migration cannot be handled as a prompt transfer challenge, but as an optimized process, which takes into account utility and operational bottlenecks jointly. All in all, the suggested framework provides a feasible and scaleable approach towards the preservation of trustworthy workflows powered by LLM in the face of regular back-end modifications. This can be further applied to continuous migration environments, expanded domain coverage, and more closely coupled with automated compliance detection in future work.

## References


[1]. S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche and M. S. Hossain, “Large Language Models in Healthcare: Applications, Opportunities, and Challenges,” *IEEE Access*, vol. 12, pp. 1-19, 2024.

[2]. M. A. Ferrag, O. Friha, L. Maglaras and H. Janicke, “Large Language Models for Cyber Security: Opportunities, Challenges, and Future Directions,” *IEEE Access*, vol. 12, pp. 1-24, 2024.

[3]. Y. K. Dwivedi et al., “So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *IEEE Access*, vol. 11, pp. 1-34, 2023.

[4]. V. Chang, P. A. K. Acharya, M. A. Ferrag and N. Kumar, “Secure and Trustworthy Generative Artificial Intelligence for Critical Applications: Challenges and Research Directions,” *IEEE Access*, vol. 12, pp. 1-18, 2024.

[5]. B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, “Taking the Human Out of the Loop: A Review of Bayesian Optimization,” *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148-175, Jan. 2016, doi: 10.1109/JPROC.2015.2494218.

[6]. D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne and H. V. Poor, “Privacy-Preserving Generative AI for Intelligent Systems: Opportunities, Challenges, and Future Directions,” *IEEE Network*, vol. 38, no. 2, pp. 1-9, 2024.

[7]. S. U. Amin, M. Alsulaiman, G. Muhammad, M. A. Mekhtiche and M. S. Hossain, “Large language models in healthcare: Applications, opportunities, and challenges,” *IEEE Access*, vol. 12, pp. 1-19, 2024.

[8]. M. A. Ferrag, O. Friha, L. Maglaras and H. Janicke, “Large language models for cyber security: Opportunities, challenges, and future directions,” *IEEE Access*, vol. 12, pp. 1-24, 2024.

[9]. A. Sallam, “The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations,” *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 1-14, 2024.

[10]. M. A. Ferrag, L. Maglaras, S. Moschoyiannis and H. Janicke, “Deep learning and large language models for cyber security: A review of applications and research challenges,” *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 312-345, 2024.

[11]. V. Chang, M. A. Ferrag and N. Kumar, “Secure and trustworthy generative artificial intelligence for critical applications: Challenges and future research directions,” *IEEE Access*, vol. 12, pp. 1-18, 2024.

[12]. P. Rana, S. Garg and G. Kaddoum, “Privacy and safety challenges in generative AI systems: A survey of threats, controls, and deployment considerations,” *IEEE Access*, vol. 12, pp. 1-22, 2024.

[13]. C. Pryzant, D. Iter, J. Li and Y. Yang, “Automatic prompt optimization with ‘gradient descent’ and beam search,” in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, 2023, pp. 1-8.

[14]. S. Mishra, D. Alon, A. Fabbri, C. Li, Y. Chen and J. M. R. Kiro, “Prompt sensitivity and automatic prompt engineering for large language models,” in Proc. IEEE Conf. on Artificial Intelligence, 2024, pp. 1-7.

[15]. J. I. Janjua, M. Irfan, T. Abbas, A. Ihsan and B. Ali, "Enhancing Contextual Understanding in Chatbots and NLP," 2024 International Conference on TVET Excellence & Development (ICTeD), Melaka, Malaysia, 2024, pp. 244-249, doi: 10.1109/ICTeD62334.2024.10844601.

[16]. T. A. Khan, M. S. Khan, S. Abbas, J. I., S. S. Muhammad and M. Asif, "Topology-Aware Load Balancing in Datacenter Networks," 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob), Bandung, Indonesia, 2021, pp. 220-225, doi: 10.1109/APWiMob51111.2021.9435218.

[17]. J. I., S. Zulfiqar, T. A. and S. A. Ramay, "Activation Function Conundrums in the Modern Machine Learning Paradigm," 2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2023, pp. 1-8, doi: 10.1109/ICCA59364.2023.10401760.

[18]. A. Ahamed, N. Ahmed, J. I. Janjua, Z. Hossain, E. Hasan and T. Abbas, "Advances and Evaluation of Intelligent Techniques in Short-Term Load Forecasting," 2024 International Conference on Computer and Applications (ICCA), Cairo, Egypt, 2024, pp. 1-9, doi: 10.1109/ICCA62237.2024.10927804.

[19]. J. I. Janjua, T. A. Khan, S. Zulfiqar and M. Q. Usman, "An Architecture of MySQL Storage Engines to Increase the Resource Utilization," 2022 International Balkan Conference on Communications and Networking (BalkanCom), Sarajevo, Bosnia and Herzegovina, 2022, pp. 68-72, doi: 10.1109/BalkanCom55633.2022.9900616.

[20]. B. Letham, R. Calandra, A. Rai and E. Bakshy, “Constrained Bayesian optimization with noisy experiments,” IEEE Transactions on Pattern Analysis *and Machine Intelligence*, vol. 43, no. 9, pp. 1-13, 2021.