Explainable AI Framework for Wearable Signal–Driven Medical Decision-Making Using Large Language Model Agents and PPG Representations

Malcolm Koskinen
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
malcolmkoskinen@colostate.edu

Akshay R. Arora
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
akshay.arora261@unh.edu

Rendres May
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
rendresm@ku.edu

Abstract

The proliferation of wearable devices generating continuous photoplethysmography (PPG) signals has created unprecedented opportunities for personalized clinical decision-making outside traditional healthcare settings. However, translating raw biosignals into trustworthy, actionable medical insights demands architectures that simultaneously handle physiological complexity, ensure interpretability, and maintain robustness under distributional shifts and adversarial threats. This paper introduces an explainable artificial intelligence framework that couples PPG-specific foundation model representations with large language model (LLM) agents to produce context-aware, natural-language medical decision support. The framework employs a layered design in which self-supervised PPG encoders extract structured, semantically rich embeddings, a registry of LLM agents performs causal reasoning and evidence synthesis, and an integrated explainability layer generates counterfactual narratives, feature attributions, and uncertainty quantification. We analyze the system-level trade-offs between computational efficiency, latency, and explainability fidelity, and we discuss the essential role of human-in-the-loop governance in high-stakes environments. Adversarial robustness is addressed by incorporating input purification modules and agent-level security enhancements that mitigate prompt injection and representation manipulation. Furthermore, the paper examines cross-domain implications by comparing PPG-based decision systems with electrocardiogram and imaging paradigms, highlighting the distinct challenges imposed by inter-subject variability, motion artifacts, and consumer-grade sensor noise. We also reflect on regulatory readiness, fairness across demographic groups, and the sustainability of deploying large-scale models in resource-constrained edge environments. Rather than proposing a singular algorithm, this work contributes a comprehensive architectural blueprint and a critical discourse on the structural, ethical, and operational forces shaping the next generation of explainable, wearable-driven AI in medicine.

Keywords

Explainable AI, Large Language Model Agents, Photoplethysmography, Wearable Devices, Medical Decision-Making, Adversarial Robustness, Foundation Models.

1. Introduction

The shift toward continuous, out-of-clinic physiological monitoring is redefining the boundaries of modern healthcare delivery. Photoplethysmography, a low-power optical sensing technique that captures blood volume pulsations, has become the cornerstone modality in millions of wrist-worn and fingertip devices. Unlike episodic measurements taken in controlled environments, PPG streams contain rich, subclinical signatures of cardiovascular regulation, respiratory dynamics, and autonomic nervous system activity that can, in principle, support early detection and longitudinal management of chronic diseases. However, the very properties that make PPG attractive—ubiquity, non-invasiveness, and affordability—also introduce severe analytical challenges. Raw PPG waveforms are corrupted by motion artifacts, ambient light interference, skin tone-dependent signal attenuation, and enormous inter-individual morphological variability. These factors erode the reliability of black-box classifiers and deeply undermine the trust required for adoption in clinical workflows.

In parallel, the landscape of artificial intelligence has been reshaped by the emergence of large language models capable of context-sensitive reasoning, knowledge retrieval, and multi-turn dialog. When these models are instantiated as agents that can invoke tools, reflect on uncertainty, and articulate their reasoning in natural language, they offer a path toward decision-support systems that do not merely emit a risk score but engage clinicians in an interpretative dialogue about a patient's trajectory. Simultaneously, the field of biosignal foundation models has shown that self-supervised pre-training on vast corpora of unlabeled PPG recordings can produce compression capabilities that capture latent physiological structure with unprecedented fidelity, even under domain shifts. The frontier lies in the deliberate synthesis of these two lines of innovation into a coherent framework that is not an afterthought collection of components but a designed socio-technical infrastructure where representation learning, agentic reasoning, and explainability are mutually reinforcing. This paper formulates such a framework, placing equal emphasis on architectural decomposition, governance mechanisms, adversarial resilience, and deployment pragmatics.

2. Background and Related Work

PPG signal analysis has historically progressed through a sequence of engineered feature extraction pipelines, from fiducial point detection to time-frequency transformations, before being overtaken by end-to-end deep learning models capable of learning directly from raw waveforms [1]. The arrival of convolutional and recurrent architectures enabled robust feature learning for tasks such as atrial fibrillation screening, blood pressure estimation, and sleep apnea detection, yet these systems remained brittle when deployed on populations and hardwares diverging from training conditions [2]. To address this brittleness, recent work has introduced PPG-specific foundation models pre-trained with generative masking objectives that encode waveform structure into compact latent spaces. One such architecture, SIGMA-PPG, employs a statistical-prior informed masking strategy that improves representation quality under heavy motion corruption, yielding embeddings that transfer effectively across device types and demographic strata [3]. This class of pre-trained encoder forms the signal backbone of the framework we describe.

On the deliberative side, LLM agents have been studied extensively as knowledge engines in medicine. Models such as Med-PaLM and GPT-4 have demonstrated expert-level

performance on board-style question answering and have been prototyped as clinical summarization assistants [4]. While these efforts underscore the potential for natural language interfaces to lower cognitive load and enhance decision transparency, they also expose a critical gap: the LLM's reasoning is typically anchored to textual descriptions or structured electronic health record data, not to raw physiological time series. Bridging this gap by coupling an LLM agent directly with a PPG foundation model creates a semi-structured interface where the agent reasons over embeddings, metadata, and contextual prompts, explaining its interpretations in terms of both physiological priors and patient-specific patterns. Yet, as with any system touching patient safety, the need for interpretability goes beyond textual justifications to encompass formal guarantees about the fidelity and stability of the explanation relative to the underlying signal representation.

Explainability in medical AI has been framed through post-hoc attribution methods, concept bottlenecks, and counterfactual reasoning [5]. In the wearable domain, however, explainability must address not only model internals but also the provenance of the signal, the quality of the sensor, and the environmental context. Several governance frameworks, including the FDA's proposed regulatory approach for adaptive AI, now mandate that manufacturers provide traceability from sensor input to clinical recommendation. Consequently, the fusion of PPG foundation models and LLM agents necessitates an explainability fabric that operates at multiple semantic levels: signal-level saliency, latent-space traversal, and natural-language narrative generation.

A further dimension that has recently received attention is the adversarial robustness of LLM-based medical agents. Adversaries can manipulate input prompts, contaminate prompt history, or exploit model priors to induce harmful clinical advice. Research on security enhancement methods has proposed architectural defenses such as structured reasoning scaffolds, uncertainty-aware response filtering, and adversarially fine-tuned guard agents specifically designed for medical decision contexts [6]. These contributions highlight the need to treat robustness as a first-class design constraint, which our framework integrates through layered sanitization and representation-level verification.

3. PPG Signal Processing and Foundation Model Representations

The first major subsystem of the proposed framework is the signal ingest and representation engine. Raw PPG waveforms arrive from wearable devices at varying sampling rates, often between 20 and 100 Hz, accompanied by accelerometer streams and metadata that capture device type, wear position, and acquisition time. Preprocessing includes adaptive bandpass filtering, motion artifact reduction using parallel accelerometer channels, and quality assessment via signal quality indices that reject segments with low signal-to-noise ratio. Rather than feeding preprocessed waveforms directly into a task-specific classifier, we map them through a frozen or fine-tuned PPG foundation model that transforms variable-length segments into fixed-dimensional embeddings. These embeddings are designed to capture both short-term morphological features, such as systolic peak shape and diastolic decay, and longer-range modulations related to respiratory sinus arrhythmia and vascular tone.

The choice of which foundation model to adopt carries structural consequences for downstream interpretability, latency, and robustness. A large transformer-based encoder pre-trained on diverse datasets may yield embeddings that are more resistant to demographic bias if the pre-training corpus is sufficiently inclusive. However, computational demands may preclude on-device inference, forcing a split between edge and cloud components that introduces new attack surfaces and latency variability. In contrast, a smaller, distilled encoder

can reside entirely on the wearable or companion phone, ensuring data locality but potentially sacrificing representational richness. The framework is agnostic to the specific encoder architecture as long as it exposes a standardized embedding interface that includes uncertainty quantification and, crucially, a reconstruction pathway that allows approximate inversion from embedding space back to waveform space. This invertibility property is essential for generating counterfactual explanations and for verifying that the agent's reasoning is grounded in actual physiological features rather than spurious embedding dimensions.

The integration of a PPG foundation model also reshapes the data governance boundary. Patient-generated PPG streams can be encoded locally, and only privacy-preserving embeddings need to be transmitted to cloud-hosted reasoning agents. This architecture aligns with emerging regulatory principles on data minimization and allows differential privacy mechanisms to be applied in embedding space without destroying clinical utility. Nonetheless, the pipeline must contend with systematic domain gaps: a foundation model trained predominantly on resting-state recordings from healthy adults may underperform when applied to pediatric populations or patients with severe arrhythmias. Continuous monitoring of embedding drift and online calibration using federated learning techniques becomes an operational necessity that we address further in the deployment section.

4. Large Language Model Agents for Contextual Reasoning

The embedding vectors produced by the PPG foundation model do not constitute an intelligible clinical output in isolation. The second architectural tier consists of a suite of LLM-based agents that consume these embeddings along with structured patient context—age, medical history, medication list, recent activity logs—and perform reasoning cycles that culminate in a natural-language interpretation. The agent design follows a registry pattern: specialist agents are responsible for distinct clinical domains such as arrhythmia assessment, hemodynamic stability monitoring, and sleep quality evaluation. A coordinator agent, which maintains a structured memory of recent embeddings and agent outputs, dynamically selects which specialist to invoke and synthesizes their observations into a coherent summary or alert.

This decomposition enhances traceability because each specialist's reasoning chain can be logged and audited independently. Moreover, the coordinator can request that a specialist produce not only a conclusion but also an explanation grounded in physiological concepts. For example, when assessing a potential atrial fibrillation episode, the specialist may identify a sequence of irregular inter-beat intervals embedded in the latent representation, cross-reference this with the patient's prior episode frequency, and express its finding as "based on a 60-second PPG window, the pulse interval variability exceeds the 90th percentile of the patient's baseline, suggesting possible paroxysmal atrial fibrillation, confounded by moderate motion artifact between seconds 20 and 25." This style of reasoning, which blends quantitative latent evidence with qualitative narrative, is precisely the affordance that LLM agents bring when properly scaffolded.

However, the introduction of LLM agents into a medical decision context raises substantial challenges around hallucination, alignment, and the boundary of clinical discretion. Agents must be constrained by a policy layer that prohibits definitive diagnosis and always frames outputs as decision support. The framework enforces this by injecting a structured system prompt that delineates the agent's role, a memory of institutional guidelines, and a set of fallback procedures when uncertainty exceeds calibrated thresholds. Furthermore, the prompt itself can be dynamically assembled from a template that includes patient demographics and relevant clinical context, but this contextual injection must be carefully sanitized to prevent

indirect prompt injections that could subvert the agent's behavior. We therefore implement a dedicated prompt sanitizer that inspects all user-originated strings for embedded instructions or semantic deviations before they reach the LLM context window.

The computational footprint of deploying multiple LLM agents is non-trivial. Inference latency must remain compatible with the cadence of streaming PPG data, which typically delivers embeddings every few seconds. Batching strategies, speculative decoding, and model quantization can compress the latency to an acceptable range, but the trade-off between model scale—and hence reasoning depth—and responsiveness is a persistent structural tension. We argue that this tension can be productively managed by employing a cascade architecture where a lightweight small language model performs fast triage and only escalates ambiguous cases to a larger, more capable model. Such a cascade aligns with clinical escalation pathways and reduces the average cost per patient-hour, making the framework viable in low-resource settings.

5. Explainability and Interpretability Mechanisms

Explainability in the proposed framework is not an addendum to the decision pipeline but a vertically integrated property spanning from signal representations to natural-language narratives. The first layer of explainability operates within the PPG foundation model via saliency mapping. Because the encoder includes a reconstruction pathway, we can perturb input PPG windows and measure the impact on the embedding vector, enabling the identification of waveform regions that most influence the downstream agent's output. This saliency can be visualized as a waveform annotation that highlights, for instance, a diastolic notch anomaly that drove a specialist's concern about aortic valve function. Visual inspection by a clinician can then confirm or refute the machine's attention, turning the black box into a conversational partner.

The second layer involves latent-space explainability. Embedding dimensions are not independently interpretable, but by traversing the latent space along directions that correspond to known physiological variation—such as heart rate variability power in low-frequency and high-frequency bands—the framework can generate counterfactual embeddings and reconstruct the corresponding PPG waveforms. When a specialist agent flags a hemodynamic abnormality, the counterfactual module can answer "what would the signal need to look like for this abnormality to be absent?" and reconstruct that hypothetical signal. Presenting clinicians with a side-by-side comparison of the actual and counterfactual waveforms creates a powerful bridge between latent reasoning and perceptual intuition. This approach directly addresses the challenge of opaque foundation models and builds trust by rooting explanation in observable physiological evidence.

The third explanation layer resides within the LLM agents themselves. We employ chain-of-thought auditing, where each specialist is required to emit a step-by-step trace of its reasoning before producing a final recommendation. This trace is parsed by an independent explainer agent that checks for logical consistency, evidence grounding, and adherence to clinical guidelines. If a reasoning gap is detected, the explainer can prompt the specialist to reconsider or escalate to a human reviewer. Additionally, the framework supports structured uncertainty quantification by requiring agents to classify each assertion with an evidence grade—ranging from "directly observed in the PPG segment" to "extrapolated from population norms." This gradation helps clinicians gauge the trustworthiness of each element of the output and reduces the risk of automation bias.

Importantly, the multi-layered explainability design is not cost-free. The computational overhead of generating saliency maps, latent traversals, and chain-of-thought traces can double end-to-end latency. We argue that for non-urgent monitoring scenarios, this latency is an acceptable investment in safety and transparency. In time-critical applications such as real-time cardiac arrest prediction, a stripped-down fast pathway can deliver a high-precision alert with minimal explanation, while a post-hoc detailed report is generated asynchronously. This dual-mode operation exemplifies the kind of context-aware governance that health systems require to balance safety, efficiency, and interpretability.

6. Adversarial Robustness and Security Considerations

The integration of LLM agents into medical decision pipelines dramatically expands the attack surface relative to earlier monolithic classifiers. An adversary can target the PPG signal via physical-world perturbations—subtle wrist movements or optical interference—or can craft digital attacks on the embedding stream, the agent prompts, or the memory state. Addressing this requires a multi-tier defense that spans the signal, representation, and agent levels. At the signal level, the framework applies input purification filters that detect statistical anomalies in the raw waveform and accelerometer traces, rejecting segments that exhibit patterned deviations consistent with adversarial manipulation. This approach draws on anomaly detection techniques developed for sensor security, which monitor for deviations from expected noise distributions.

At the representation level, the PPG foundation model's embeddings are passed through a perturbation detector that compares incoming embeddings against a running statistical baseline. Embeddings that fall outside a Mahalanobis distance threshold trigger a conservative fallback mode where only the most robust clinical indicators are computed, and the LLM agent is restricted to a templated, low-risk response set. This mechanism prevents adversarially crafted embeddings from inducing high-confidence errors in the specialist agents, a failure mode documented in the literature on adversarial examples in medical imaging that extends analogously to physiological time series.

The most critical vulnerability lies within the LLM agent layer. Prompt injection, context manipulation, and data poisoning can cause the coordinator or specialist agents to override their safety constraints or fabricate clinical evidence. Recent work has specifically addressed adversarial robustness for LLM agents in medical decision-making tasks, proposing methods that combine structured deliberation scaffolds with runtime guard agents that monitor output consistency and veto responses that deviate from clinical plausibility [6]. Incorporating such guard agents into our framework creates a defensive boundary that operates independently of the reasoning specialists, enforcing a separation of concerns between clinical reasoning and safety enforcement. The guard agent maintains a whitelist of permitted semantic frames and can interrupt a specialist's response mid-generation if it detects dangerous divergence, replacing the unsafe output with a neutral escalation message.

Beyond technical defenses, adversarial robustness must be embedded in the organizational processes surrounding the framework. Regular red-teaming exercises that simulate sophisticated adversaries—including those with physical access to the wearable device—should feed back into model updates and policy refinements. Additionally, model cards and transparency reports should disclose the residual risk of adversarial manipulation under realistic threat models, enabling clinical users to calibrate their reliance on the system accordingly. The structural principle is that no single defense layer can be assumed impenetrable; rather, a system of overlapping, independently designed safeguards provides

defense in depth. This philosophy mirrors practices in mission-critical avionics and nuclear power control, where safety cases rely on multiple, diverse protective barriers.

7. Deployment, Governance, and Ethical Implications

Deploying an explainable PPG-LLM framework in real healthcare ecosystems involves navigating a complex interplay of regulatory, infrastructural, and sociotechnical forces. From a regulatory perspective, the framework qualifies as a software as a medical device in many jurisdictions, triggering design controls, risk management according to ISO 14971, and ongoing post-market surveillance. The framework's modularity—separating the PPG encoder, agent registry, and explainability fabric—facilitates incremental regulatory clearance, where the encoder can be validated as a signal processing component while the LLM agent undergoes iterative clinical evaluation. However, the very adaptiveness that makes LLM agents powerful also challenges a regulatory paradigm built on frozen, versioned software. We envision a lifecycle management system where agent behaviors are governed by configuration files that can be updated without full model retraining, yet each configuration change triggers a pre-specified clinical validation protocol proportional to its risk tier.

Infrastructure decisions deeply shape both clinical fidelity and equity. Running the entire stack on cloud infrastructure enables the use of large, state-of-the-art LLMs but introduces connectivity dependencies that disadvantage rural clinics and low-resource settings. A hybrid architecture where a quantized PPG encoder and a small language model reside on the edge device, with optional cloud escalation for complex cases, emerges as a pragmatic compromise. This hybrid model aligns with the principles of sustainable AI, reducing the carbon footprint of continuous inference by minimizing data transmission and cloud compute. The sustainability dimension extends to the maintenance costs of keeping foundation models and LLM agents up to date with evolving medical knowledge, a challenge that calls for community-driven, open-source model repositories analogous to those emerging for large language models.

Fairness across demographic groups is a persistent concern, especially given well-documented disparities in PPG signal quality due to skin pigmentation and peripheral perfusion differences. The framework must be audited for performance equity using stratified evaluations that disaggregate error rates by skin type, age, and comorbidity profiles. Pre-trained PPG foundation models must be examined for whether their pre-training data oversampled lighter-skinned individuals, and mitigation strategies such as targeted fine-tuning on underrepresented groups and embedding reweighting should be deployed when disparities are detected. The LLM agent introduces an additional fairness vector: language generation models may inadvertently produce explanations that are less coherent or less empathic for certain demographic cohorts, a form of linguistic bias that can erode trust and worsen health outcomes. Continuous monitoring of explanation quality, combined with human evaluation panels that reflect the diversity of the intended patient population, provides a governance mechanism to catch and correct such biases.

Finally, the sociotechnical dimension of human-AI teaming in this context cannot be overstated. The framework is designed not to replace clinicians but to augment their cognitive capacity, particularly in ambulatory monitoring where clinicians are overloaded with data. For augmentation to succeed, the system must be transparent about its confidence, must actively request human input in ambiguous cases, and must be designed such that overriding the AI's suggestion is both easy and normatively reinforced. This requires thoughtful user experience design and institutional policies that protect clinicians who disagree with the system.

Embedding these practices within a learning health system ethos—where every override is logged as a potential feedback signal for model improvement—creates a virtuous cycle of refinement that benefits all stakeholders.

8. Conclusion

This paper has articulated an explainable AI framework that weds PPG foundation model representations with large language model agents to produce transparent, robust, and context-aware medical decision support for wearable scenarios. By structuring the system into signal encoding, agentic reasoning, and multi-layer explainability, we have shown how physiological grounding and narrative interpretability can be jointly achieved without sacrificing either. The architectural discussion highlighted key trade-offs around edge-cloud partitioning, computational latency, and the depth of explanation, and we argued for a dual-mode design that adapts to clinical urgency. Adversarial robustness was presented as a cross-cutting property enforced by input purification, embedding perturbation detection, and dedicated guard agents operating at the LLM level. Crucially, the framework positions governance, fairness, and sustainability not as external constraints but as integral components of the architecture, shaping everything from model card design to dynamic escalation policies.

Realizing the vision described here demands interdisciplinary collaboration among signal processing engineers, machine learning researchers, clinical safety experts, and health system administrators. Future work should prioritize large-scale clinical validation studies that measure not only diagnostic accuracy but also clinician trust calibration, decision speed, and patient outcomes when the framework is deployed longitudinally in diverse clinical environments. Furthermore, as foundation models for biosignals and LLM agents continue to evolve, the framework must be periodically reassessed against emergent capabilities and failure modes. We hope that the structural principles advanced in this paper will inform the responsible development of next-generation wearable intelligence, shifting the discourse from raw performance metrics toward resilient, interpretable, and equitably deployed decision-support ecosystems.

References

1. Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3), R1–R39.

2. Pereira, T., Tran, N., Gadhoumi, K., Pelter, M. M., Do, D. H., Lee, R. J., … & Hu, X. (2020). Photoplethysmography based atrial fibrillation detection: a review. npj Digital Medicine, 3(1), 3.

3. Guo, Z., Chen, T., Jiao, Y., Pan, Y., Hu, X., & Ferrario, M. (2026). SIGMA-PPG: Statistical-prior Informed Generative Masking Architecture for PPG Foundation Model. arXiv preprint arXiv:2601.21031.

4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., … & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

5. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

6. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

7. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., … & Bakas, S. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

8. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (pp. 1597–1607). PMLR.

9. Dunn, J., Kidzinski, L., Runge, R., Witt, D., Hicks, J. L., Schüssler-Fiorenza Rose, S. M., … & Snyder, M. P. (2021). Wearable sensors enable personalized predictions of clinical laboratory measurements. Nature Medicine, 27, 1105–1112.

10. Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6449–6464). Association for Computational Linguistics.

11. Loh, H. W., Ooi, C. P., Tan, E., Ng, W. Y., Tan, R. S., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade. Computer Methods and Programs in Biomedicine, 215, 106620.

12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).

13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.

14. Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6(1), 26094.

15. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.

16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

17. Schulam, P., & Saria, S. (2017). Reliable decision support using counterfactual models. In Advances in Neural Information Processing Systems (pp. 1697–1708).

18. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

19. Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine, 383(9), 874–882.

20. Wang, Z., Chen, Q., & Wang, W. (2023). Prompt injection attack against LLM-integrated applications. arXiv preprint arXiv:2306.05499.

21. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.