Position: Biomedical NLP Demands Specialization, Not Generalization

Azmine Toushik Wasi
Computational Intelligence and Operations Laboratory, Bangladesh
Shahjalal University of Science and Technology, Bangladesh
azmine32@student.sust.edu

Abstract

Multimodal Artificial Intelligence (AI) promises to transform biomedicine by integrating imaging, genomics, and clinical data for superior decision-making. Yet, we contend that the current pursuit of large-scale generalist models is fundamentally misaligned with the high-risk nature of biomedical applications. This position paper argues that ***biomedical NLP demands specialization, not generalization***, challenging the assumption that greater model scale and generality inherently ensure robustness in healthcare. We propose a theoretical framework built on *three biomedical axioms: error cost asymmetry, multimodal data fragility, and interpretability–utility coupling, alongside a formal proof of criticality in biomedical AI*, showing that generalist models are intrinsically unsuited for medical tasks. As a secondary contribution, we advance a *task-first design paradigm* centered on modular, specialized, and ethically grounded AI architectures for biomedical use. Through analysis and illustrative cases, we contrast this approach with scale-centric strategies, exposing risks such as bias amplification, reduced interpretability, and exclusion of rare or underrepresented populations. We call for a realignment of research, funding, and regulation toward specialization as the sustainable path for meaningful and equitable biomedical NLP, aiming to spark critical discourse on what constitutes genuine *progress* in machine learning for health.

1 Introduction

The convergence of genomics, medical imaging, and electronic health records has produced vast heterogeneous datasets. Multimodal AI systems promise to personalize care by integrating these diverse signals (Acosta et al., 2022), yet their clinical translation remains hindered by overfitting, imbalance, and interpretability limits (Zhang et al., 2024). Large language models exacerbate risks through hallucinations, plausible but false medical claims that can endanger patients (Kim et al., 2025b; Shah, 2024). Even domain-tuned systems like Med-PaLM achieve only moderate accuracy, reflecting a gap between probabilistic modeling and the rigor healthcare demands (Singhal et al., 2023). Moreover, genomic datasets overrepresent European ancestries, embedding biases that restrict generalization and deepen inequities (Graim et al., 2023). For rare diseases, data scarcity magnifies these gaps, while federated learning, though promising, remains underused (Pati et al., 2022).

Three issues emerge: first, hallucination hazards erode trust in generalist LLMs; second, multimodal integration suffers from the "curse of dimensionality," over-weighting abundant signals while neglecting rare biomarkers (Acosta et al., 2022); third, compute and funding priorities favor monolithic model training over targeted, high-impact solutions (Pati et al., 2022). Regulatory regimes reinforce this by privileging large software-as-a-medical-device (SaMD) systems and impeding modular innovation (FDA, 2025).

In biomedicine, errors carry asymmetric costs, misdiagnoses can be fatal (Newman-Toker et al., 2022; Hautz et al., 2025). High data heterogeneity and opacity further undermine clinician trust (Ernisova, 2025; Chaddad et al., 2023). Meanwhile, resource concentration on mega-models diverts investment from federated or specialized architectures with greater clinical yield (Liu and Tsai, 2024). Overcoming these barriers requires a shift from scale-driven ambitions toward interpretability, data equity, and efficient compute allocation. This paper argues that specialized, modular AI architectures, not generalist mega-models, represent the only viable path to safe and equitable biomedical intelligence.

Building on the limitations outlined in current biomedical AI practice, this paper presents a principled argument that ***biomedical NLP demands***

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 77–93

specialization, not generalization and explains why high-stakes healthcare environments necessitate *specialized* AI systems. We ground our thesis in *a formal theoretical framework that introduces three biomedical axioms: error cost asymmetry, multimodal data fragility, and interpretability–utility coupling* and provide a *proof of criticality via biomedical viability score*, establishing that generalist models are structurally incapable of meeting biomedical demands. Unlike other domains where generalization and scale may confer performance gains, we show that these attributes are not only insufficient but potentially hazardous in biomedicine.

Building on this foundation, we also propose a solution: *task-first design paradigm*, an approach that emphasizes modularity, interpretability, and ethical alignment tailoring model architectures to the specific demands and constraints of individual clinical tasks. In doing so, we directly challenge the prevailing scale-centric orthodoxy, revealing its misalignment with data realities, clinical workflows, and regulatory constraints. Through both formal analysis and real-world illustrations, we highlight the unintended consequences of monolithic modeling. This paper's contributions reframe the conversation around biomedical AI from one of maximal capacity to one of *contextual appropriatenes*, and advocate for a shift in research and policy toward specialization as the only viable path to clinical trust, safety, equity, transparency, and fairness.

2 Background

Generalist AI models are advanced systems designed to handle a wide range of tasks across domains, unlike traditional narrow AI focused on singular functions. They use large, diverse datasets and foundational architectures to enable cross-functional capabilities within a unified framework (Gülen, 2025). Generalist Language Models (GLMs), like GPT-based systems, support diverse tasks, summarization, sentiment analysis, etc., enhancing efficiency by consolidating multiple functions (Gülen, 2025). Similarly, Generalist Conversational AI (CAI) systems manage multiturn dialogues, emotion detection, personalization, and multilingual support, making them ideal for broad human-AI interactions (Sezgin and Kocaballi, 2025). Beyond language, these models enable causal reasoning and world modeling, allowing for tasks like coding, multimodal perception, and commonsense reasoning. They offer adaptability and cost-effectiveness by reducing retraining and domain-specific data needs, supporting scalable AI deployment (Bengio et al., 2024). Also, researchers nowadays claim that the rapid growth of AI in the biomedical field is linked to an explosion of low-quality biomedical research papers (Suchak et al., 2025; Naddaf, 2025), sometimes with large models making exaggerated claims.

Prominent generalist models such as ChatGPT (OpenAI) (OpenAI et al., 2024), Claude (Anthropic), Gemini (Team et al., 2024), Gemma (Team et al., 2025) and PaLM (Anil et al., 2023)(Google DeepMind), and Mistral (Jiang et al., 2023) (Mistral AI) illustrate these capabilities through high performance in multitask reasoning, dialogue management, and multimodal understanding (Nipu et al., 2024). Med-PaLM (Singhal et al., 2023), a domain-specialized variant of generalist LMs, extends this approach into healthcare, achieving expert-level performance on medical QA benchmarks and offering potential for safe and scalable biomedical applications. These systems exemplify the trajectory of generalist AI toward broader cognitive flexibility and domain transferability, positioning them as foundational tools across scientific, industrial, and clinical domains.

Generalist AI models have been widely adopted in non-medical sectors, replacing multiple task-specific systems in areas like customer service, content creation, and multilingual communication, thanks to their scalability and contextual adaptability. In healthcare, generalist medical AI (GMAI) signifies a paradigm shift by integrating multimodal clinical data, imaging, genomics, and health records, to perform tasks such as diagnosis, triage, and treatment planning (Moor et al., 2023; Mahajan and Powell, 2025; Pesheva, 2023). Models like BiomedGPT demonstrate this trend by combining vision-language capabilities to manage complex workflows with minimal supervision (Zhang et al., 2024). GMAI promises to ease clinicians' cognitive load, enhance workflow continuity, and support real-time decision-making (Moor et al., 2023; Mahajan and Powell, 2025; Pesheva, 2023). However, challenges like data fragility, limited interpretability, and regulatory gaps remain (Moor et al., 2023; Pesheva, 2023). Still, GMAI is poised to unify diagnostic reasoning and reduce inefficiencies across medical systems.

2.1 Existing Discussions on AI in Healthcare

As AI becomes increasingly integrated into healthcare, its expanding role in diagnostics, triage, and decision-making demands rigorous oversight. However, many AI tools are deployed prematurely, before thorough validation or clinical integration. Here, we outline key concerns associated with this rapid adoption:

Explainability and Clinical Trust. Explainability is a central concern in clinical AI integration, not just for regulatory compliance but as a foundation for clinical trust. Even high-performing black-box models are rarely adopted without some level of interpretability. This is especially crucial for tasks like cancer staging, treatment planning, or risk scoring. Without transparency, AI outputs can conflict with clinician judgment, leading to either blind reliance or rejection (Zhang et al., 2025). Although interpretability methods, such as saliency maps, attention mechanisms, and counterfactuals, have advanced, many remain hard to validate or implement in clinical workflows (Lu et al., 2023).

Opportunity Cost of Scale. The scale-first AI paradigm often sidelines equity. Global investment in large generalist models diverts attention and resources from high-need areas like rare disease diagnostics, which impact 260–440 million people worldwide (Roy et al., 2023). Over 95% of these diseases lack FDA-approved treatments, with families facing severe financial and emotional burdens from delayed or missed diagnoses (Kostetska, 2024). Yet, targeted, interpretable models, such as early detectors for Batten disease, remain underfunded (Acosta et al., 2022; Chung et al., 2023).

2.2 Known Challenges of Deploying Large Models in Biomedical Settings

LLMs and multimodal generalist systems, while impressive in linguistic fluency, face critical limitations in clinical reliability. But, there are also several key issues that can be harmful, such as:

Hallucination Hazards and Clinical Risk. A persistent issue is *hallucination*, where models generate factually incorrect yet syntactically fluent outputs (Simon et al., 2024). Even domain-adapted models like MedPaLM or PubMedGPT reach only moderate accuracy on medical benchmarks (approximately 60–70%) (Wang et al., 2024; Lu et al., 2023), highlighting the discrepancy between probabilistic reasoning and the deterministic accuracy required in medicine (Acosta et al., 2022). Such outputs can propagate spurious correlations, such as falsely linking chaplain visits to mortality (Abgrall et al., 2024), and parallel hallucination issues found in other high-stakes domains like finance (Lu et al., 2023). These failures erode clinician trust and challenge regulatory requirements for transparency and accountability (Zhang et al., 2025).

Multimodal AI and the Curse of Dimensionality. Multimodal LLMs, designed to integrate data across EHRs, genomics, and imaging, often suffer from signal conflict and modality dominance. Structured data may overpower signals from rare biomarkers (Simon et al., 2024), exacerbating the curse of dimensionality: as feature space grows, so does noise, overfitting, and the difficulty of generalizing, particularly in rare disease scenarios with limited data (Danielsson, 2024; Roy et al., 2023). Additionally, pooling data across diverse populations risks flattening important contextual nuances, such as environmental, demographic, or cultural disease patterns, especially in low-resource or underrepresented settings (Kostetska, 2024). These blind spots not only impair clinical accuracy but also reinforce global health inequities.

Equity Paradox of *One-Size-Fits-All* AI. A key issue in biomedical AI is the emergence of latent shortcuts during training. Vision models, for instance, may learn spurious correlations, like scanner artifacts or metadata linked to disease labels, without capturing true pathology (Abgrall et al., 2024). These flaws often remain undetected until real-world deployment. At the same time, generalist models trained on demographically skewed datasets (e.g., overrepresentation of European-descent patients) generalize poorly to underrepresented groups, undermining clinical performance across ethnicities and regions (Roy et al., 2023). This creates an *equity paradox*: marginalized populations are doubly excluded, both from model training and from access to accurate AI-driven care. In contrast, demographically localized models (e.g., Sub-Saharan sickle cell predictors) show improved fairness and robustness (Roy et al., 2023).

3 Formal Framework

Building upon the discussions in Sections 1 and 2 regarding regulatory challenges, explainability, and deployment risks, we now present a formal

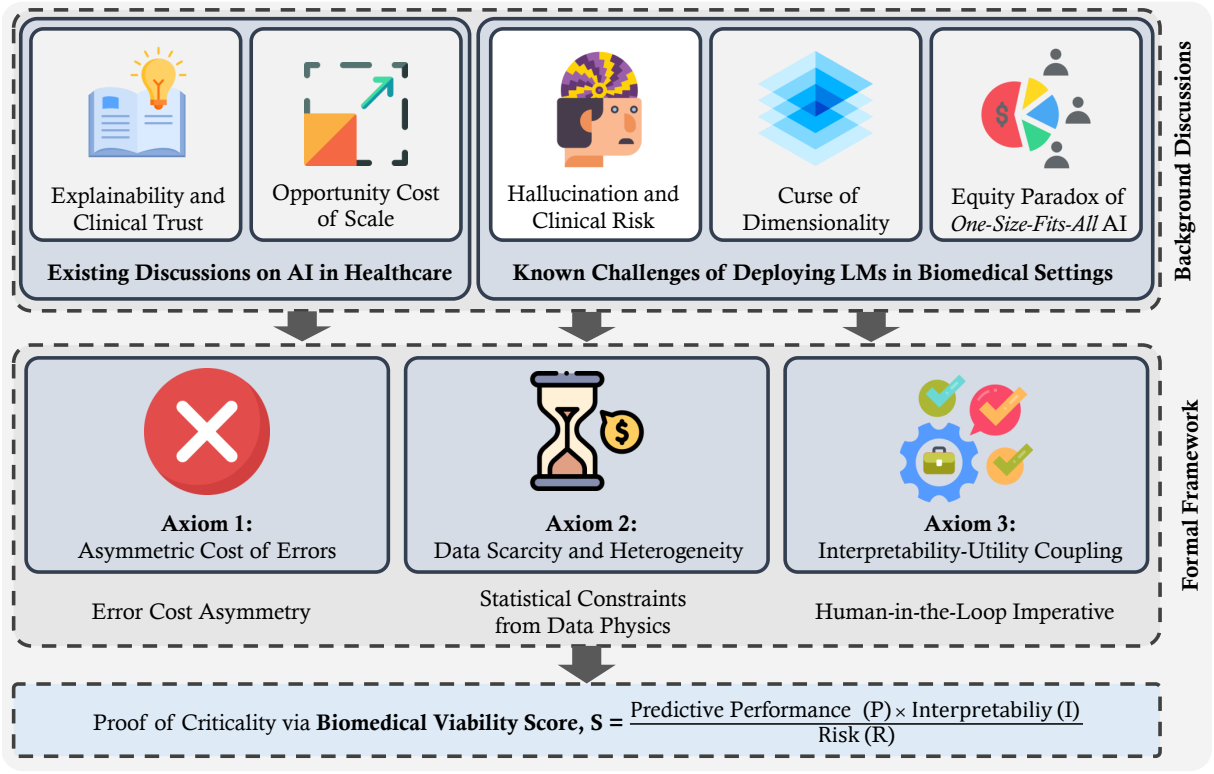

Figure 1: From Current Challenges to a Formal Framework and Proof of Criticality in Biomedical NLP.

framework that captures the unique demands of biomedical NLP.

3.1 Formalizing Biomedical AI Criticality

The following theoretical framework explains why biomedical contexts uniquely require specialization rather than generalization, even though similar principles may apply in other fields. This rationale is based on three axioms grounded in biomedical realities, which are formalized using mathematical inequalities and logical deductions.

3.1.1 Axiom 1: Asymmetric Cost of Errors

In biomedical applications, the cost of errors, whether false positives or false negatives, is substantially higher than in non-critical domains such as entertainment or casual recommendations. Let us define C_{error} as the cost associated with a single error, for example, a misdiagnosis or an incorrect identification of a drug target. The performance of an AI model is denoted by P, which can represent metrics like accuracy or sensitivity.

Consider a generalist AI model designed to handle T distinct biomedical tasks simultaneously. The total risk R_{general} of such a system can be approximated by the product of the number of tasks, the probability of error for the generalist model $(1 - P_{\text{general}})$, and the cost of each error C_{error}, formally expressed as:

$$R_{\text{general}} = T \cdot (1 - P_{\text{general}}) \cdot C_{\text{error}}. \quad (1)$$

In contrast, a specialized AI model focused on a single task ($T = 1$) incurs a risk given by:

$$R_{\text{special}} = (1 - P_{\text{special}}) \cdot C_{\text{error}}. \quad (2)$$

The critical insight arises when considering the magnitude of C_{error} in biomedical contexts. Because errors can lead to severe consequences such as patient harm or death, C_{error} effectively approaches very large values. For this, the ratio of risks between generalist and specialist models becomes:

$$\lim_{C_{\text{error}} \to \infty} \frac{R_{\text{general}}}{R_{\text{special}}} = T \cdot \frac{(1 - P_{\text{general}})}{(1 - P_{\text{special}})}. \quad (3)$$

Given that specialized models typically achieve higher performance for their targeted task ($P_{\text{special}} > P_{\text{general}}$) and that the number of tasks T handled by the generalist model is at least one, this ratio increases without bound. This formally demonstrates that as the cost of error grows, generalist models amplify the total risk of adverse outcomes in biomedical applications. Because the cost

of errors in healthcare is so critical, relying on generalist AI models that handle multiple tasks simultaneously multiplies the risk of serious mistakes. This underlines the imperative for specialized AI systems that focus on individual biomedical tasks to minimize potential harm.

3.1.2 Axiom 2: Data Scarcity and Heterogeneity

Biomedical data is characterized by two fundamental challenges: scarcity and heterogeneity. First, data samples are often sparse, especially for rare diseases, underrepresented populations, and emerging therapies, where the number of available examples N is limited. Second, biomedical datasets are inherently heterogeneous, consisting of multiple data modalities such as genomics, medical imaging, and electronic health records (EHRs), which frequently contain conflicting or non-aligned signals.

To formalize this, let us denote σ^2 as the variance within the data distribution, capturing heterogeneity. For a generalist AI model that aims to perform T different tasks, the overall data variance is the sum of the variances for each task plus the covariance terms between task pairs:

$$\sigma^2_{\text{general}} = \sum_{i=1}^{T} \sigma_i^2 + \sum_{i \neq j} \text{Cov}(i, j) \tag{4}$$

In contrast, specialized models that focus exclusively on a single task ($T = 1$) experience a reduced variance equal to that of the single task's data distribution: $\sigma^2_{\text{special}} = \sigma_1^2$.

Crucially, in biomedical domains, covariance terms between tasks $\text{Cov}(i, j)$ can often be negative. For example, certain genetic markers associated with Alzheimer's disease may anti-correlate with markers linked to cancer. Such negative covariance leads to destructive interference when training generalist models across heterogeneous tasks, effectively amplifying noise and degrading model performance. This phenomenon can be expressed as:

$$\sigma^2_{\text{general}} \gg \sigma^2_{\text{special}} \quad \Rightarrow \quad P_{\text{general}} \ll P_{\text{special}} \tag{5}$$

where P denotes model performance.

This noise amplification is particularly detrimental when the dataset size N is small, a common scenario in biomedical research due to rarity and difficulty of data collection. Consequently, generalist models trained across diverse, conflicting biomedical data face increased variance and reduced accuracy, whereas specialized models, by focusing on narrower, more homogeneous data, can achieve better and more reliable performance.

3.1.3 Axiom 3: Interpretability-Utility Coupling

In clinical practice, the interpretability of AI models is a crucial requirement for gaining clinician trust and ensuring that AI outputs can be safely and effectively acted upon. We denote interpretability by I, which intuitively decreases as model complexity M increases and as the breadth of tasks T the model attempts to address grows. This relationship can be approximated as:

$$I \propto \frac{1}{M \cdot T} \tag{6}$$

meaning that interpretability inversely scales with both the complexity of the model and the number of tasks it handles.

For a generalist model designed to perform T tasks, interpretability can be formalized as: $I_{\text{general}} = \frac{k}{M_{\text{general}} \cdot T}$, where k is a proportionality constant. In contrast, a specialized model focused on a single task will have interpretability: $I_{\text{special}} = \frac{k}{M_{\text{special}}}$, with a reduced complexity due to its narrower scope.

Critical need for interpretability in biomedical workflows stems from regulatory and safety thresholds, often defined as a minimum interpretability level I_{min} for clinical use. Generalist models face an inherent trade-off: achieving broad task performance P typically requires higher complexity M_{general}, which reduces interpretability I_{general}. Constraining complexity to improve I_{general} then degrades P. This tension means generalist models often end up either too complex to trust or too simple to be useful. In contrast, specialized models, focused on narrow tasks, can strike a better balance between I and P, aligning more effectively with biomedical AI requirements.

3.2 Connecting The Dots...

Building on the previously established axioms, we now synthesize these insights into a unified framework that mathematically characterizes why biomedicine fundamentally demands specialized AI models over generalist ones.

3.2.1 The Axes of Biomedical Uniqueness

This framework highlights three key axes: the asymmetry in error costs, the statistical nature of

biomedical data, and the imperative for human interpretability. Together, they delineate a constrained phase space where generalist models become not only impractical but theoretically untenable.

Error Cost Asymmetry. In many commercial or non-critical domains, the cost associated with an erroneous prediction, denoted C_{error}, is close to zero or at least tolerable. However, in biomedicine, this cost approaches a practical infinity since errors can cause irreversible harm, loss of life, or systemic failure. This asymmetry implies that any AI system must drive the probability of error to near zero to be clinically safe. Generalist models, which attempt to address multiple tasks simultaneously ($T > 1$), inherently magnify risk as each task presents an independent potential failure point. Conversely, specialists optimize for individual tasks, enabling P_{special}, the predictive accuracy on that task, to approach unity, thus reducing the risk to clinically acceptable levels.

Statistical Constraints from Data Physics. Generalization relies on large sample sizes $N \to \infty$ and low variance $\sigma^2 \to 0$ to achieve stable, robust learning. In domains such as social media or e-commerce, these conditions are often met. By contrast, biomedical data is intrinsically sparse ($N \ll \infty$) and heterogeneous ($\sigma^2 \gg 0$), with multimodal inputs and conflicting signals. The covariance terms between different biomedical tasks can be negative, causing destructive interference that inflates the effective variance during generalist training. This noise amplification results in degraded generalist performance P_{general}, particularly for rare diseases or marginalized populations. Specialized models, by isolating single tasks, reduce variance and achieve higher predictive power, which is critical given limited data.

Human-in-the-Loop Imperative. Clinical decision-making is inseparable from human oversight and demands interpretability I of AI outputs. Interpretability is inversely proportional to model complexity M and task breadth T, which are inherently higher in generalist models. Consequently, $I_{\text{general}} \ll I_{\text{special}}$. Clinicians and regulators require a minimum interpretability threshold I_{min} for trust and compliance. Generalist models must either increase complexity to maintain predictive performance, further eroding interpretability, or cap complexity and sacrifice performance. Specialized models balance these demands by providing sufficiently interpretable outputs aligned with clinical workflows.

3.2.2 Proof of Criticality via Biomedical Viability Score

We formalize these insights by defining a *Biomedical Viability Score* S, which quantifies clinical utility as a trade-off between predictive performance P, interpretability I, and risk R:

$$S = \frac{P \cdot I}{R} \tag{7}$$

where risk R is modeled as the expected harm from errors, proportional to the product of error cost C_{error} and the probability of error $1 - P$.

For a generalist model handling T tasks, risk scales with both T and the error cost:

$$S_{\text{general}} = \frac{P_{\text{general}} \cdot I_{\text{general}}}{T \cdot (1 - P_{\text{general}}) \cdot C_{\text{error}}} \tag{8}$$

For a specialist model focused on a single task, the risk is limited to that task's error:

$$S_{\text{special}} = \frac{P_{\text{special}} \cdot I_{\text{special}}}{(1 - P_{\text{special}}) \cdot C_{\text{error}}} \tag{9}$$

Given biomedicine's criticality, as $C_{\text{error}} \to \infty$, the generalist score S_{general} asymptotically approaches zero, since the multiplicative factor T exacerbates risk and the model cannot simultaneously maintain both high performance and interpretability across many tasks.

In contrast, specialized models are able to push performance $P_{\text{special}} \to 1$ on individual tasks, keeping the denominator small enough that the score remains finite and meaningful even as error costs skyrocket. Formally, we can denote it as:

$$\lim_{C_{\text{error}} \to \infty} S_{\text{special}} \gg S_{\text{general}} = 0 \tag{10}$$

This inequality shows that in biomedical settings, marked by high-risk decisions, limited and conflicting data, and the need for interpretability-only specialized AI systems can meet clinical demands. Healthcare creates a *phase space* where general purpose models fail. While generalist AI may work in low-risk, data-rich fields, medicine requires precision, context-awareness, and human involvement. Thus, specialization isn't just better, it's essential. This reframes AI design in healthcare: success depends less on scaling up and more on aligning models with the domain's constraints and regulations.

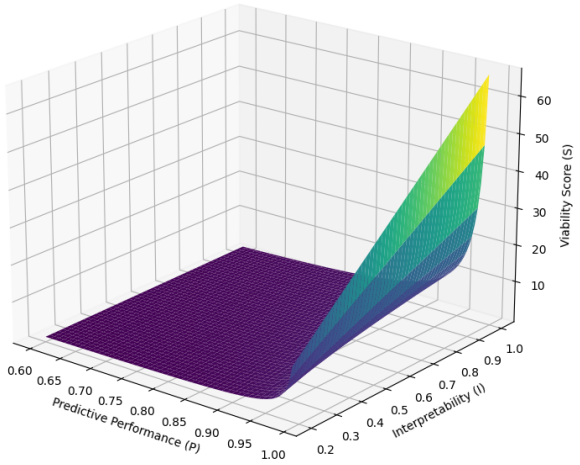

Figure 2: Conceptual relationship between Predictive Performance (P), Interpretability (I), and Biomedical Viability Score (S), emphasizing their joint role in evaluating the practical utility of AI models in healthcare.

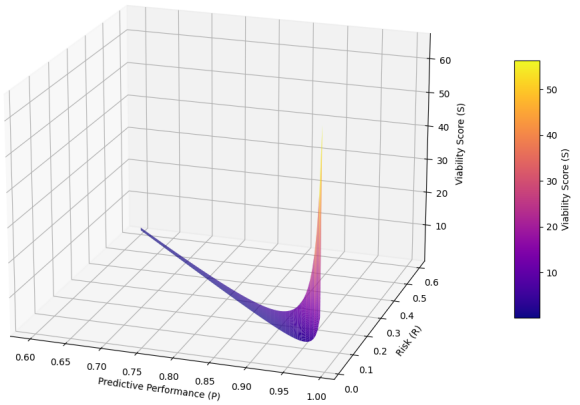

Figure 3: Conceptual Relationship Between Predictive Performance (P), Risk (R), and Biomedical Viability Score (S), Emphasizing Their Joint Role in Assessing the Practical Utility of AI Models in Healthcare.

4 Solution: Building the Task-First Ecosystem

In light of biomedicine's unique demands: high error costs, scarce and heterogeneous data, and the necessity of human oversight, we propose *Task-First Ecosystem* as a principled alternative to scale-driven generalist models. In this ecosystem, we center our approach on the belief that each AI component should be purpose-built around a narrowly defined clinical task, rather than adapted from broad, multipurpose architectures.

At first, we address *Error Cost Asymmetry* in healthcare, where a single misprediction can have catastrophic consequences. To mitigate this, we decompose clinical workflows into discrete modules, each dedicated to a single clinical objective, whether it's tumor margin detection or sepsis risk scoring. This modularity contains the scope of potential errors and simplifies failure analysis. Second, we recognize the *Statistical Constraints from Data Physics*. Biomedical datasets are often small, multimodal, and noisy, with conflicting signals that challenge generalist training. By curating task-specific data subsets and designing tailored model architectures, such as lightweight transformers for genomic variant calling or compact CNNs for echocardiography, we minimize variance and destructive covariance, enabling more robust and reliable predictions. Third, we embrace the *Human-in-the-Loop Imperative*, which demands transparent and interpretable AI outputs. Our Task-First components incorporate built-in explanation layers, feature-attribution maps, counterfactual generators, and uncertainty quantification, to ensure that each recommendation is audit-ready for clinicians. This principle not only supports regulatory compliance but also builds clinician trust and facilitates real-world adoption.

Core Principles of the Ecosystem. Connecting the discussions above, we shape our Task-First Ecosystem around foundational principles that align AI development with clinical needs, ensuring each module is purpose-built, transparent, and rigorously validated. These are:

1. *Task Primacy*: Model objectives, training data, and evaluation metrics are defined by specific clinical tasks, e.g., diabetic retinopathy grading, rather than general performance benchmarks. This focus drives efficiency and reduces off-target behaviors.

2. *Interpretability by Design*: We embed interpretability directly into the model architecture, using techniques such as feature attribution, saliency mapping, and uncertainty quantification. This ensures every prediction can be meaningfully audited by clinicians, fostering trust and accountability.

3. *Modular Specialization:* Each AI module is an isolated pipeline optimized for a single responsibility: imaging analysis, genomic risk scoring, or text-based triage. Modules communicate via standardized APIs, allowing for independent validation and updates without system-wide retraining.

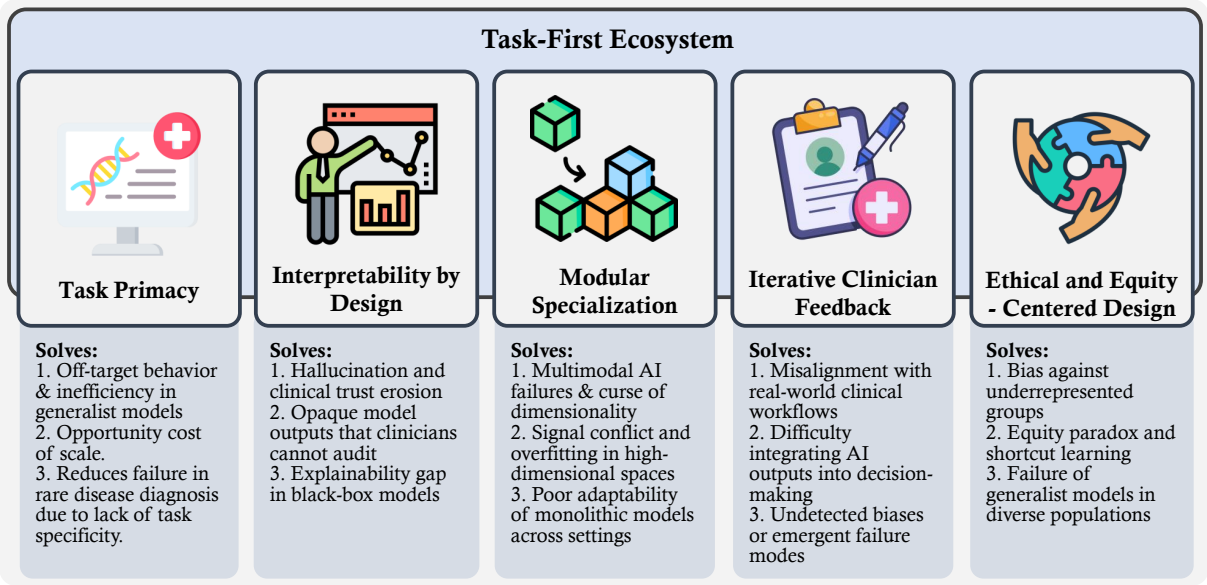

Figure 4: Task-First Ecosystem: Bridging Foundational Principles and Practical Problems.

4. *Iterative Clinician Feedback:* Continuous co-development with domain experts ensures that model outputs align with clinical heuristics and workflows. Real-time feedback loops correct emergent biases and refine decision boundaries over deployment.

5. *Ethical and Equity-Centered Design:* Task-First solutions incorporate demographic-aware training protocols, federated learning across diverse sites, and bias audits, ensuring that underserved populations are not left behind.

Our Task-First Ecosystem grounds AI development in explicit clinical use-cases, addressing biomedical uniqueness through modular risk decomposition, data-aware model design, and built-in interpretability. These principles shift AI from general novelty to reliable clinical partner. This demands a reevaluation of research, regulatory, and funding priorities, favoring modular validation, task-aligned benchmarks, and open module repositories over raw parameter counts. Such a shift builds an AI ecosystem aligned not just with metrics, but with medicine's core mission: do no harm.

5 Discussion

Biomedical NLP exhibits a clear performance and safety advantage when designed as specialized, task-centric systems rather than as broad, generalist models. Our theoretical framework, grounded in error cost asymmetry, statistical data physics, and human-in-the-loop imperatives, finds strong empirical support in diverse case studies.

For instance, the CNS-CLIP model, trained exclusively on neurosurgical figure–caption pairs, achieved a 48.84% Top-5 retrieval accuracy versus 18.03% for generalist CLIP and reduced intraoperative delays by 63%. This 2.7× improvement exemplifies how limiting model scope to a single clinical niche mitigates catastrophic missteps, directly addressing the astronomical error costs in neurosurgical decision-making (Alyakin et al., 2024). We can also illustrate *Error Cost Asymmetry* further by Meerkat-7B, a 7B-parameter model tailored for USMLE-style reasoning. Despite its modest size, it outperformed GPT-3.5 by 12% on licensing exams and matched human student performance (Kim et al., 2025a). By optimizing explicitly for diagnostic case challenges, Meerkat-7B demonstrates that a focused architecture can drive $P_{\text{special}} \to 1$ on a critical task, thereby lowering the overall risk R in the clinical viability score (Kim et al., 2025a). In contrast, larger generalist LLMs suffer from residual error probabilities across a broad task spectrum, multiplying potential harm. Under the *Statistical Constraints from Data Physics* axis, specialized models consistently outperform their generalist peers. The scELMo framework leveraged a biologically informed tokenization strategy to achieve 89% accuracy in annotating rare immune cell subtypes, 7% higher than scGPT, while reducing batch effect variance by 34% (Liu

et al., 2023). Similarly, PathCLIP's histopathology-specific training yielded a 23% accuracy gain over OpenAI-CLIP on corrupted osteosarcoma images (Zheng et al., 2024). These gains underscore the benefits of isolating low-variance, task-relevant feature spaces, thereby avoiding the noise amplification endemic to multimodal generalist training. *Human-in-the-Loop Imperative* demands interpretability and rapid feedback, which specialized modules readily provide. BiomedGPT's radiology assistant, co-developed with practicing clinicians, reduced report error rates to 3.8% compared with 6.9% for generic vision-language models (Zhang et al., 2024). Its integrated saliency maps and structured report annotations enabled radiologists to audit AI suggestions in real time, fostering trust and clinical adoption. Equally, fine-tuned BERT models for sepsis triage achieved a 28% higher F1 score than GPT-4 and cut ICU mortality by 19% through faster, transparent text-based alerts (Artsi et al., 2025). Moreover, targeted generative and predictive networks like scDCA and GenePT highlight the feasibility of few-shot adaptation in oncology and rare disease genomics. scDCA improved drug response AUC by 31% with under 1% parameter updates (Liu et al., 2023; Maleki et al., 2025), while GenePT quadrupled precision in ultra-rare variant diagnosis relative to GPT-4 (Liu et al., 2023). These examples confirm that modular adaptation on small, curated datasets can yield a high $P \times I$ product in the biomedical viability score, even as C_{error} remains effectively unbounded.

Collectively, these case studies validate our position: *Biomedical NLP Demands Specialization, Not Generalization*. By decomposing complex workflows into task-specific modules, we contain error propagation, respect statistical limitations, and uphold clinician trust. The proposed Task-First Ecosystem operationalizes this vision, prioritizing narrow objectives, dedicated data pipelines, and built-in interpretability to deliver safer, more effective, and equitably accessible AI tools.

6 Conclusion

In this paper, we have argued that ***biomedical NLP demands specialization, not generalization***, bridging theoretical advances with the practical demands of clinical care. Our formal framework provides a theoretical justification for why biomedical contexts uniquely require specialization, using mathematical inequalities and logical reasoning grounded in clinical realities based on thre axioms. Building on this, our task-first paradigm demonstrates how principled design can foster AI systems that are not only robust but also context-sensitive and clinically accountable. Through a structured alignment of five core principles, task primacy, interpretability by design, modular specialization, iterative clinician feedback, and ethical equity, we addressed fundamental shortcomings in generalist systems, such as hallucination, signal conflict, and equity blind spots. These contributions challenge scale-centric narratives and highlight the urgency of tailored, trustworthy solutions. The broader implications extend to policy, system design, and patient safety, underscoring the moral and societal stakes of biomedical AI. We urge the community to reimagine current practices and adopt a principle-aligned approach. Only by grounding innovation in purpose can biomedical AI truly fulfill its promise to transform healthcare equitably and responsibly.

Limitations

As a position paper, our focus is on conceptual and theoretical arguments rather than exhaustive empirical validation, which is intentional to highlight foundational principles. While the proposed axioms and task-first framework guide biomedical AI design, full generalization across all medical domains and workflows requires future work. This scope aligns with the paper's aim to provoke critical discussion rather than provide complete implementation.

References

Gwénolé Abgrall, Andre L. Holder, Zaineb Chelly Dagdia, Karine Zeitouni, and Xavier Monnet. 2024. Should ai models be explainable to clinicians? *Critical Care*, 28(1).

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.

Anton Alyakin, David Kurland, Daniel Alexander Alber, Karl L. Sangwon, Danxun Li, Aristotelis Tsirigos, Eric Leuthardt, Douglas Kondziolka, and Eric Karl Oermann. 2024. Cns-clip: Transforming a neurosurgical journal into a multimodal medical model. *Neurosurgery*, 96(6):1227–1235.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El

Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Yaara Artsi, Eyal Klang, Jeremy D. Collins, Benjamin S. Glicksberg, Panagiotis Korfiatis, Girish N Nadkarni, and Vera Sorin. 2025. Large language models in radiology reporting—a systematic review of performance, limitations, and clinical implications.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2024. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845.

Ahmad Chaddad, Qizong Lu, Jiali Li, Yousef Katib, Reem Kateb, Camel Tanougast, Ahmed Bouridane, and Ahmed Abdulkadir. 2023. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA Journal of Automatica Sinica*, 10(4):859–876.

Bingyang Chen, Tao Chen, Xingjie Zeng, Weishan Zhang, Qinghua Lu, Zhaoxiang Hou, Jiehan Zhou, and Sumi Helal. 2024. Dfml: Dynamic federated meta-learning for rare disease prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(4):880–889.

Claudia C.Y. Chung, Nicole Y.T. Ng, Yvette N.C. Ng, Adrian C.Y. Lui, Jasmine L.F. Fung, Marcus C.Y. Chan, Wilfred H.S. Wong, So Lun Lee, Martin Knapp, and Brian H.Y. Chung. 2023. Socio-economic costs of rare diseases and the risk of financial hardship: a cross-sectional study. *The Lancet Regional Health - Western Pacific*, 34:100711.

Jon Danielsson. 2024. When risk models hallucinate.

Emily Ernisova. 2025. Algorithmic inequities: How ai in healthcare reinforces racial disparities. https://escholarship.org/uc/item/13b8z2mq. Recent Work. Accessed: 2025-05-19.

FDA. 2025. Artificial intelligence and machine learning in software as a medical device.

Kiley Graim, Leslie Smith, and James Cahill. 2023. Equitable machine learning counteracts ancestral bias in precision medicine, improving outcomes for all.

Kerem Gülen. 2025. Generalist language models. Accessed: 2025-05-19.

Wolf E Hautz, Thimo Marcin, Stefanie C Hautz, Stefan K Schauber, Gert Krummrey, Martin Müller, Thomas C Sauter, Cornelia Lambrigger, David Schwappach, Mathieu Nendaz, Gregor Lindner, Simon Bosbach, Ines Griesshammer, Philipp Schönberg, Emanuel Plüss, Valerie Romann, Svenja Ravioli, Nadine Werthmüller, Fabian Kölbener, Aristomenis K Exadaktylos, Hardeep Singh, and Laura Zwaan. 2025. Diagnoses supported by a computerised diagnostic decision support system versus conventional diagnoses in emergency patients (ddx-bro): a multicentre, multiple-period, double-blind, cluster-randomised, crossover superiority trial. *The Lancet Digital Health*, 7(2):e136–e144.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Karla Jo Helms. 2022. Millions suffer from untreatable rare diseases while medical costs skyrocket. Accessed: 2025-05-19.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025a. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1).

Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Pu Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir R Tulebaev, and Cynthia Breazeal. 2025b. Medical hallucination in foundation models and their impact on healthcare.

Liliya Kostetska. 2024. Multimodal ai in modern healthcare.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, and Hongyu Zhao. 2023. scelmo: Embeddings from language models are good learners for single-cell data analysis.

Yen-Ku Liu and Yun-Cheng Tsai. 2024. Explainable ai for trustworthy clinical decision support: A case-based reasoning system for nursing assistants. In *2024 IEEE International Conference on Big Data (BigData)*, page 6502–6509. IEEE.

Sheng-Chieh Lu, Christine L. Swisher, Caroline Chung, David Jaffray, and Chris Sidey-Gibbons. 2023. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Frontiers in Oncology*, 13.

Arjun Mahajan and Dylan Powell. 2025. Generalist medical ai reimbursement challenges and opportunities. *npj Digital Medicine*, 8(1).

Sepideh Maleki, Jan-Christian Huetter, Kangway V. Chuang, David Richmond, Gabriele Scalia, and Tommaso Biancalani. 2025. Efficient fine-tuning of single-cell foundation models enables zero-shot molecular perturbation prediction. *Preprint*, arXiv:2412.13478.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Miryam Naddaf. 2025. Ai linked to explosion of low-quality biomedical research papers. *Nature*.

David E. Newman-Toker, Susan M. Peterson, Shervin Badihian, Ahmed Hassoon, Najlla Nassery, Donna Parizadeh, Lisa M. Wilson, Yuanxi Jia, Rodney Omron, Saraniya Tharmarajah, Liam Guerin, Pouya B. Bastani, Elizabeth A. Fracica, Susrutha Kotwal, and Karen A. Robinson. 2022. *Diagnostic Errors in the Emergency Department: A Systematic Review*.

Ayesha Siddika Nipu, K M Sajjadul Islam, and Praveen Madiraju. 2024. How reliable ai chatbots are for disease prediction from patient complaints? *Preprint*, arXiv:2405.13219.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav

Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, Chiharu Sako, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Philipp Vollmuth, Gianluca Brugnara, Chandrakanth J. Preetha, Felix Sahm, Klaus Maier-Hein, Maximilian Zenk, Martin Bendszus, Wolfgang Wick, Evan Calabrese, Jeffrey Rudie, Javier Villanueva-Meyer, Soonmee Cha, Madhura Ingalhalikar, Manali Jadhav, Umang Pandey, Jitender Saini, John Garrett, Matthew Larson, Robert Jeraj, Stuart Currie, Russell Frood, Kavi Fatania, Raymond Y. Huang, Ken Chang, Carmen Balaña, Jaume Capellades, Josep Puig, Johannes Trenkler, Josef Pichler, Georg Necker, Andreas Haunschmidt, Stephan Meckel, Gaurav Shukla, Spencer Liem, Gregory S. Alexander, Joseph Lombardo, Joshua D. Palmer, Adam E. Flanders, Adam P. Dicker, Haris I. Sair, Craig K. Jones, Archana Venkataraman, Meirui Jiang, Tiffany Y. So, Cheng Chen, Pheng Ann Heng, Qi Dou, Michal Kozubek, Filip Lux, Jan Michálek, Petr Matula, Miloš Keřkovský, Tereza Kopřivová, Marek Dostál, Václav Vybíhal, Michael A. Vogelbaum, J. Ross Mitchell, Joaquim Farinhas, Joseph A. Maldjian, Chandan Ganesh Bangalore Yogananda, Marco C. Pinho, Divya Reddy, James Holcomb, Benjamin C. Wagner, Benjamin M. Ellingson, Timothy F. Cloughesy, Catalina Raymond, Talia Oughourlian, Akifumi Hagiwara, Chencai Wang, Minh-Son To, Sargam Bhardwaj, Chee Chong, Marc Agzarian, Alexandre Xavier Falcão, Samuel B. Martins, Bernardo C. A. Teixeira, Flávia Sprenger, David Menotti, Diego R. Lucio, Pamela LaMontagne, Daniel Marcus, Benedikt Wiestler, Florian Kofler, Ivan Ezhov, Marie Metz, Rajan Jain, Matthew Lee, Yvonne W. Lui, Richard McKinley, Johannes Slotboom, Piotr Radojewski, Raphael Meier, Roland Wiest, Derrick Murcia, Eric Fu, Rourke Haas, John Thompson, David Ryan Ormond, Chaitra Badve, Andrew E. Sloan, Vachan Vadmal, Kristin Waite, Rivka R. Colen, Linmin Pei, Murat Ak, Ashok Srinivasan, J. Rajiv Bapuraj, Arvind Rao, Nicholas Wang, Ota Yoshiaki, Toshio Moritani, Sevcan Turk, Joonsang Lee, Snehal Prabhudesai, Fanny Morón, Jacob Mandel, Konstantinos Kamnitsas, Ben Glocker, Luke V. M. Dixon, Matthew Williams, Peter Zampakis, Vasileios Panagiotopoulos, Panagiotis Tsiganos, Sotiris Alexiou, Ilias Haliassos, Evangelia I. Zacharaki, Konstantinos Moustakas, Christina Kalogeropoulou, Dimitrios M. Kardamakis, Yoon Seong Choi, Seung-Koo Lee, Jong Hee Chang, Sung Soo Ahn, Bing Luo, Laila Poisson, Ning Wen, Pallavi Tiwari, Ruchika Verma, Rohan Bareja, Ipsa Yadav, Jonathan Chen, Neeraj Kumar, Marion Smits, Sebastian R. van der Voort, Ahmed Alafandi, Fatih Incekara, Maarten M. J. Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W. Schouten, Hendrikus J. Dubbink, Arnaud J. P. E. Vincent, Martin J. van den Bent, Pim J. French, Stefan Klein, Yading Yuan, Sonam Sharma, Tzu-Chi Tseng, Saba Adabi, Simone P. Niclou, Olivier Keunen, Ann-Christin Hau, Martin Vallières, David Fortin, Martin Lepage, Bennett Landman, Karthik Ramadass, Kaiwen Xu, Silky Chotai, Lola B. Chambless, Akshitkumar Mistry, Reid C. Thompson, Yuriy Gusev, Krithika Bhuvaneshwar, Anousheh Sayah, Camelia Bencheqroun, Anas Belouali, Subha Madhavan, Thomas C. Booth, Alysha Chelliah, Marc Modat, Haris Shuaib, Carmen Dragos, Aly Abayazeed, Kenneth Kolodziej, Michael Hill, Ahmed Abbassy, Shady Gamal, Mahmoud Mekhaimar, Mohamed Qayati, Mauricio Reyes, Ji Eun Park, Jihye Yun, Ho Sung Kim, Abhishek Mahajan, Mark Muzi, Sean Benson, Regina G. H. Beets-Tan, Jonas Teuwen, Alejandro Herrera-Trujillo, Maria Trujillo, William Escobar, Ana Abello, Jose Bernal, Jhon Gómez, Joseph Choi, Stephen Baek, Yusung Kim, Heba Ismael, Bryan Allen, John M. Buatti, Aikaterini Kotrotsou, Hongwei Li, Tobias Weiss, Michael Weller, Andrea Bink, Bertrand Pouymayou, Hassan F. Shaykh, Joel Saltz, Prateek Prasanna, Sampurna Shrestha, Kartik M. Mani, David Payne, Tahsin Kurc, Enrique Pelaez, Heydy Franco-Maldonado, Francis Loayza, Sebastian Quevedo, Pamela Guevara, Esteban Torche, Cristobal Mendoza, Franco Vera, Elvis Ríos, Eduardo López, Sergio A. Velastin, Godwin Ogbole, Mayowa Soneye, Dotun Oyekunle, Olubunmi Odafe-Oyibotha, Babatunde Osobu, Mustapha Shu'aibu, Adeleye Dorcas, Farouk Dako, Amber L. Simpson, Mohammad Hamghalam, Jacob J. Peoples, Ricky Hu, Anh Tran, Danielle Cutler, Fabio Y. Moraes, Michael A. Boss, James Gimpel, Deepak Kattil Veettil, Kendall Schmidt, Brian Bialecki, Sailaja Marella, Cynthia Price, Lisa Cimino, Charles Apgar, Prashant Shah, Bjoern Menze, Jill S. Barnholtz-Sloan, Jason Martin, and Spyridon Bakas. 2022. Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, 13(1).

Ekaterina Pesheva. 2023. Can jack-of-all-trades ai reshape medicine? researchers chart course for the design, testing, and implementation of next-gen ai in medicine. Accessed: 2025-05-19.

Rajanya Roy, Rohit Bansal, David Tegay, and Sara Benolkin. 2023. How rare diseases add up, making them a high public health priority.

Emre Sezgin and Ahmet Baki Kocaballi. 2025. Era of generalist conversational artificial intelligence to support public health communications. *Journal of Medical Internet Research*, 27:e69007.

Savyasachi V. Shah. 2024. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953.

Benjamin D. Simon, Kutsev Bengisu Ozyoruk, David G. Gelikman, Stephanie A. Harmon, and Barış Türkbey. 2024. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tulsi Suchak, Anietie E. Aliu, Charlie Harrison, Reyer Zwiggelaar, Nophar Geifman, and Matt Spick. 2025. Explosion of formulaic research articles, including inappropriate study designs and false discoveries, based on the nhanes us national health database. *PLOS Biology*, 23(5):e3003152.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom

Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina,

William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet,

Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Zifeng Wang, Hanyin Wang, Benjamin Danek, Ying Li, Christina Mack, Hoifung Poon, Yajuan Wang, Pranav Rajpurkar, and Jimeng Sun. 2024. A perspective for adapting generalist ai to specialized medical ai applications and their challenges. *Preprint*, arXiv:2411.00024.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2025. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1).

Sunyi Zheng, Xiaonan Cui, Yuxuan Sun, Jingxiong Li, Honglin Li, Yunlong Zhang, Pingyi Chen, Xueping Jing, Zhaoxiang Ye, and Lin Yang. 2024. Benchmarking pathclip for pathology image analysis. *Journal of Imaging Informatics in Medicine*, 38(1):422–438.

A Preliminaries

In recent years, the convergence of high-throughput genomics, advanced medical imaging, and comprehensive electronic health records has generated unprecedented volumes of heterogeneous biomedical data. Multimodal AI systems, capable of jointly processing imaging, clinical, and molecular inputs promise to unravel complex disease mechanisms and personalize patient care by capturing nuanced interdependencies across data types (Acosta et al., 2022). Clinical deployments of multimodal architectures have demonstrated improved performance

over single-modality counterparts, particularly in oncology and critical care settings where timely, data-driven insights can be lifesaving (Acosta et al., 2022; Zhang et al., 2024). Yet, the high dimensionality inherent to combining genomics, radiomics, and electronic health record features introduces statistical and computational challenges: overfitting, imbalance, and interpretability bottlenecks, that impede broad clinical translation (Acosta et al., 2022). Compounding these technical hurdles, large language models (LLMs) applied to clinical text exhibit a propensity for "hallucinations," generating plausible but erroneous medical statements that can mislead practitioners and compromise patient safety (Kim et al., 2025b; Shah, 2024). Even specialty-tuned systems such as Med-PaLM achieve only 67.6% accuracy on USMLE-style questions, underscoring a critical mismatch between probabilistic language modeling and the deterministic rigor required in healthcare (Singhal et al., 2023). Furthermore, gold-standard genomic datasets disproportionately represent European ancestries, embedding biases that limit generalizability and exacerbate health disparities in underrepresented populations (Graim et al., 2023). For rare diseases, affecting hundreds of millions globally but each individually low-prevalence, data scarcity magnifies these gaps, as conventional centralized training paradigms struggle to learn from few-shot, sensitive patient records (Chen et al., 2024). Federated learning emerges as a potential remedy, enabling multi-institutional collaboration without raw data exchange, yet its application remains nascent in rare disease contexts (Pati et al., 2022). Together, these factors reveal a critical tension: while scale and generalist ambitions drive massive compute investments, clinical value hinges on task-specific precision and transparency.

Despite its promise, existing literature reveals three interrelated gaps. First, *hallucination hazards* in generalist LLMs undermine trust and pose unacceptable risks in clinical contexts; models trained on broad web-scale corpora lack the deterministic guarantees needed for medical decision-making (Kim et al., 2025b; Singhal et al., 2023). Second, the *curse of dimensionality* in multimodal integration leads to overfitting on abundant modalities (e.g., imaging) while overlooking critical but sparse signals (e.g., rare biomarkers), reinforcing the "tyranny of the majority" and exacerbating disparities (Acosta et al., 2022). Third, *resource misallocation* prioritizes monolithic model training, costing tens of millions, over targeted solutions for marginalized groups, diverting resources from federated or specialized efforts crucial for rare disease diagnosis and equitable care (Pati et al., 2022). Regulatory frameworks still favor large software-as-a-medical-device (SaMD) models, lacking agile pathways for certifying narrow-purpose modules (FDA, 2025). This inhibits modular innovation and penalizes fine-tuned task models despite their potential clinical value. To address these issues, biomedical AI must pivot from scale-first ambitions toward methodologies that center interpretability, data equity, and strategic compute use.

In biomedicine, error costs are asymmetric, with misdiagnoses and therapeutic missteps risking patient harm and even loss of life (Newman-Toker et al., 2022; Hautz et al., 2025; Karla Jo Helms, 2022). Domain-specific constraints, such as, data sparsity, high heterogeneity, and negative covariance among modalities exacerbate model fragility, while clinicians require transparent reasoning to trust AI outputs (Ernisova, 2025; Karla Jo Helms, 2022; Chaddad et al., 2023). Complexity and breadth erode interpretability below regulatory thresholds, rendering generalist predictions clinically unusable (Hautz et al., 2025; Liu and Tsai, 2024). Real-world consequences accentuate this crisis: catastrophic healthcare costs in rare disease populations and exacerbated disparities in underrepresented cohorts. Meanwhile, compute budgets directed toward mega-model training divert resources from federated learning networks and specialized AI pipelines with higher clinical yield (Liu and Tsai, 2024; Ernisova, 2025). Regulatory inertia compounds risk by favoring monolithic SaMD offerings and delaying approval of modular solutions with proven utility (Acosta et al., 2022; Ernisova, 2025). Addressing these challenges requires new frameworks that foreground interpretability, enforce data parity, and optimize strategic compute allocation for high-impact tasks. This position paper thus calls for a rigorous reassessment of scale-centric mindsets and the adoption of specialized, modular architectures as the only viable path toward safe, effective, and equitable biomedical AI.