Impressions on a Survey of Life Sciences and Medically Focused LLMs from 2022 to 2026

Austin James[1], Xavier-Lewis Palmer[2], Lucas Potter[2] and Timilehin Valentina Oyelakin[3]
[1]University of East London, UK
[2]BIOSView Labs, Dayton, USA
[3]Miva Open University, Nigeria

austinjamesigwe@gmail.com (Corresponding author)
BiosView1@proton.me (Corresponding author)

Abstract: Biomedical foundation models now address protein and genomic sequences, single-cell transcriptomes, clinical narratives, medical images and longitudinal disease histories. Although some employ related language-modelling approaches, their outputs carry different biological meanings and support different scientific or clinical claims. This narrative synthesis examines 35 purposively selected model-focused publications published from 2022 to October 2026, alongside methodological, clinical-evaluation and governance literature. The publications were selected to illustrate contrasts across applications, not to estimate the prevalence of particular models or validation practices. The synthesis advances three analytical propositions: similarities in computational methods do not establish equivalence of biological meaning; model performance must be interpreted in relation to intended task, population and evaluation setting; and linking predictions across biological scales introduces evidentiary requirements beyond those established for individual components. To structure these comparisons, we propose a seven-facet descriptive framework encompassing provenance, function, modality, domain, evaluation, trustworthiness and deployment. Four model–task examples, protein-structure prediction, single-cell annotation, hospital readmission prediction, and medical image question answering, illustrate how apparently similar claims of model performance rest on different data, assessments, and operational contexts. Two clinical studies make the distinction concrete. In a prospective, non-interventional deployment across 29,286 encounters, NYUTron yielded an evaluation AUROC of 0.787 for 30-day all-cause readmission, though operational false-positive rates at optimal decision thresholds still bound clinical utility. In a separate randomised clinical-vignette study involving 50 physicians, access to GPT-4 did not produce a statistically significant improvement in diagnostic-reasoning scores compared with conventional resources. These findings assess different questions, and neither demonstrates improved patient outcomes. Population-level disease prediction raises further concerns about cohort composition and generalisability, while model availability and deployment arrangements do not, by themselves, establish privacy or safety. This framework serves as a descriptive taxonomy rather than a universal performance benchmark or prescriptive clinical certification. It directs future work toward task-specific evaluation, explicit evidence status, and empirical assessment of uncertainty and error at interfaces between specialist systems or within generalist alternatives. The central contribution is to make the limits of evidence visible before asserting cross-scale integration or clinical readiness.

Keywords: Biomedical foundation models, Narrative synthesis, Clinical evaluation, Genomic AI, Single-cell models, Multimodal AI

1. Introduction

The application of language-model architectures and foundation-model approaches to biomedical data has produced systems with overlapping computational methods but distinct scientific purposes and evidentiary foundations. Protein and genomic models predict or generate biological sequence representations (Lin et al., 2023; Nguyen et al., 2024), single-cell models analyse gene-expression patterns (Cui et al., 2024), biomedical text models process scientific literature (Luo et al., 2022), and medical vision–language systems integrate images and text (Sellergren et al., 2026). Models trained on electronic health records (EHRs) additionally predict clinical events and disease trajectories (Jiang et al., 2023; Shmatko et al., 2025). Treating these heterogeneous architectures under the blanket term 'medical large language models' obscures fundamental divergences in input spaces, task constraints, and validation standards.

These distinctions have practical consequences. While protein-structure prediction advances molecular research, it does not independently verify biological function, therapeutic safety, or clinical efficacy (Lin et al., 2023). Similarly, medical question-answering benchmarks and simulated diagnostic consultations assess different aspects of model performance (Singhal et al., 2023; Tu et al., 2025), neither of which independently demonstrates improved patient outcomes. In a randomised clinical-vignette study, giving physicians access to GPT-4 did not significantly improve diagnostic-reasoning scores compared with conventional resources (Goh et al., 2024). EHR-based predictions also depend on data collection practices, outcome definitions and population characteristics; even prospective predictive performance does not establish the clinical benefit of acting on model outputs (Jiang et al., 2023; Shmatko et al., 2025). Model appraisal must therefore consider the task, data environment, evaluation design and intended setting alongside the model itself.

This narrative synthesis advances three interconnected analytical propositions. First, computational similarities across biological and clinical representations do not establish equivalence of biological meaning. Second, interpret model performance in relation to its intended task, population, and context, rather than through superficially comparable benchmark scores. Third, connecting outputs across molecular, cellular, and clinical scales introduces additional evidentiary requirements at model interfaces. These propositions structure the subsequent comparisons and identify questions for future empirical investigation rather than conclusions independently validated by this synthesis.

2. Basis and Boundaries of the Narrative Synthesis

The discussion uses a purposively selected, source-linked set of 35 original model studies published between 2022 and October 2026. Studies were chosen to make contrasts visible across biological scale, input modality, model objective, evaluation setting, and deployment arrangements; examples include generative and non-generative systems, open and institutionally integrated programmes, and models with varying degrees of biological or clinical testing. Primary model reports were privileged for capability claims, while evaluation trials and governance guidance were considered separately. The evidence inventory identifies the selected publications. Selection did not use repository popularity, download activity or performance thresholds. This narrative approach offers breadth and explanatory contrast, not a systematic search, complete inventory or estimate of field-wide prevalence.

'Biomedical language-orientated foundation model' is used here as a practical umbrella category. It includes generative language models, pretrained non-generative encoders, models of biological symbol sequences and biomedical vision–language systems. The term does not imply that a protein sequence encoder is functionally equivalent to a conversational LLM, or that a model trained on clinical text is automatically suitable for patient care. Scientific applications are described at molecular, genomic, cellular, patient and population scales, while recognising that some programmes span boundaries.

The synthesis asks what each model represents, what the available evidence actually demonstrates, and what new evidence would be required to transfer an inference to another biological scale or into healthcare. The seven facets in Section 6 help answer those questions. Independent inter-rater or external construct-validation studies illustrate, but do not establish, their classification utility.

3. Scientific Modelling from Molecules to Cells

3.1 Protein and Genomic Sequences

Protein language models illustrate how related sequence representations support distinct predictive and generative tasks. ProtGPT2 and ProGen2 investigate protein sequence generation and the modelling of natural protein properties, while the ESM-2/ESMFold programme focuses on predicting protein structure from learned sequence representations (Ferruz, Schmidt and Höcker, 2022; Nijkamp et al., 2023; Lin et al., 2023). ESM3 extends protein design by integrating sequence, structure, and function information, and experimental characterisation of a generated fluorescent protein demonstrates functionality in a specific setting (Hayes et al., 2025). These studies provide different forms of computational and experimental evidence but do not establish general safety, manufacturability or clinical utility for generated proteins.

Genomic models extend sequence analysis across longer contexts and diverse biological systems. HyenaDNA addresses long-range genomic modelling at single-nucleotide resolution, while DNABERT-2 and the Nucleotide Transformer investigate transferable representations for genomic classification and prediction tasks (Nguyen et al., 2023; Zhou et al., 2024; Dalla-Torre et al., 2025). Evo extends sequence modelling and design across molecular and genomic scales, including experimentally tested CRISPR-associated systems, while Evo 2 broadens training across genomic sequences representing all domains of life (Nguyen et al., 2024; Brixi et al., 2026). Nevertheless, performance depends on the organisms, sequence distributions, evaluation tasks and experimental conditions considered. Benchmark accuracy and predicted variant effects should therefore be distinguished from experimentally established phenotypes and broader biological generalisability.

3.2 Cellular Representation and Biological Context

Single-cell foundation models provide a distinct analytical level for investigating cellular states, gene expression and regulatory relationships. scBERT focuses on cell-type annotation, while Geneformer supports context-dependent predictions in gene-network biology, including selected experimentally evaluated interventions (Yang, F. et al., 2022; Theodoris et al., 2023). scGPT and scFoundation extend pretrained representations across

cellular annotation, data integration and perturbation-related tasks (Cui et al., 2024; Hao et al., 2024). Subsequent models, including scPRINT and CellFM, expand training scale and examine gene-network inference, cellular representations and perturbation prediction (Kalfon et al., 2025; Zeng et al., 2025). scGPT's pretraining on more than 33 million cells illustrates the scale of representation learning, although dataset size alone does not establish causal validity or generalisation beyond evaluated conditions (Cui et al., 2024).

Cellular modelling therefore warrants interpretation distinct from sequence-level prediction. Gene-expression profiles depend on cell type, tissue, developmental and disease context, perturbation state and measurement technology. Consequently, cross-dataset generalisation, batch effects, and experimental evidence must be considered alongside predictive accuracy (Yang, F. et al., 2022; Hao et al., 2024; Zeng et al., 2025). Although computational models can identify candidate regulatory relationships, mechanistic interpretations require evidence appropriate to the claimed causal relationship. Geneformer illustrates the value of targeted experimental validation, whereas benchmark-based network inference alone cannot establish that every predicted relationship is causal (Theodoris et al., 2023; Kalfon et al., 2025).

4. Patient-level and Multimodal Systems

4.1 From Language Competence to Clinical Tasks

Biomedical language models such as BioGPT and GatorTron demonstrated different applications of language pretraining to biomedical text generation and EHR-based information processing (Luo et al., 2022; Yang, X. et al., 2022). Subsequent models, including GatorTronGPT, PMC-LLaMA, BioMistral and HuatuoGPT, explored clinical text generation, biomedical pretraining, instruction tuning and medical consultation tasks (Peng et al., 2023; Wu et al., 2024; Labrak et al., 2024; Zhang et al., 2023). These applications are not interchangeable: clinical information extraction, conversational assistance and diagnostic support require distinct performance criteria and risk assessments (Gallifant et al., 2025).

Med-PaLM and Med-PaLM 2 advanced evaluation of medical question answering and clinician-rated responses, while Med-PaLM M extended this work to multimodal biomedical tasks (Singhal et al., 2023; Singhal et al., 2025; Tu et al., 2024). Such studies provide evidence of performance on controlled knowledge- and reasoning-oriented tasks, but do not independently establish improved clinical decisions, fewer adverse events or patient benefit (Vasey et al., 2022). NYUTron offers a different evidentiary example through EHR-based prediction, including a prospective, single-arm, non-interventional evaluation of hospital readmission risk (Jiang et al., 2023). This establishes prospective predictive performance within one health system, rather than the effectiveness of AI-assisted clinical interventions. The distinction concerns intended use, study design and demonstrated outcomes, not simply model capability.

4.2 Images, Consultations and Situated Assessment

Multimodal biomedical models further complicate direct comparisons. LLaVA-Med and Med-Flamingo investigate biomedical image–language tasks; CONCH focuses on computational pathology; and RadFM and VILA-M3 address radiological and specialist vision–language applications (Li et al., 2023; Moor et al., 2023; Lu et al., 2024; Wu et al., 2025; Nath et al., 2025). MedGemma likewise demonstrates performance comparable to or exceeding selected comparison models across reported medical text and imaging benchmarks (Sellergren et al., 2026). However, benchmark performance alone does not establish clinical readiness. Image provenance, acquisition variation, modality-specific errors and integration into clinical workflows require evaluation appropriate to the intended application (World Health Organization, 2024; Vasey et al., 2022).

Conversational systems illustrate the distinction between simulation and clinical effectiveness. AMIE was evaluated through simulated diagnostic consultations, while subsequent research assessed disease-management reasoning in structured, multi-visit virtual examinations involving comparisons with physicians (Tu et al., 2025; Liévin et al., 2026). These studies provide evidence of conversational and management performance under controlled conditions, not demonstrated effectiveness in routine patient care. Similarly, in a randomised clinical-vignette study involving 50 physicians, access to GPT-4 produced median diagnostic-reasoning scores of 76%, compared with 74% using conventional resources; the adjusted difference was two percentage points (95% confidence interval −4 to 8; P = 0.60) (Goh et al., 2024). The difference was not statistically significant, although this does not establish equivalence. An exploratory model-alone comparison also yielded higher scores than the conventional-resource physician group. Consequently, standalone model performance and the effects of integrating AI into clinician workflows require separate evaluation (Vasey et al., 2022).

5. From Patient Histories to Population-level Patterns

Health-event language models offer approaches to modelling disease trajectories across individuals and cohorts. Foresight generates patient timelines from biomedical concepts extracted from EHR free text, whereas Delphi-2M models sequences of recorded diagnoses and associated health information (Kraljevic et al., 2024; Shmatko et al., 2025). Delphi-2M was trained on approximately 0.4 million UK Biobank participants and externally evaluated using data from 1.9 million Danish individuals to predict risks across more than 1,000 diseases. Mean AUC declined from 0.69 in longitudinal UK Biobank testing to 0.67 in Danish external testing, providing evidence of cross-cohort predictive transfer with some reduction in performance (Shmatko et al., 2025). These findings do not establish universal generalisability, causal inference or improved population health. Healthcare access, diagnostic coding, cohort selection and historical patterns of missingness remain consequential. Disease-specific prediction methods may also outperform generalist models in particular settings, precluding claims of universal superiority (Shmatko et al., 2025).

A proposed multiscale connection can be expressed as an evidentiary sequence: genomic variation may affect molecular function; molecular perturbations may alter cellular states; cellular phenotypes may relate to measured patient conditions; and longitudinal health records may inform population-level risk estimates. Research on cellular foundation models demonstrates capabilities in gene-network inference and perturbation-related tasks (Theodoris et al., 2023; Cui et al., 2024) but does not establish the causal relationships linking these scales to clinical outcomes. Each connection requires independent validation appropriate to its biological context, measurement methods and intended use. Figure 1 therefore depicts a conceptual sequence of evidentiary questions rather than a validated, end-to-end biomedical AI system.

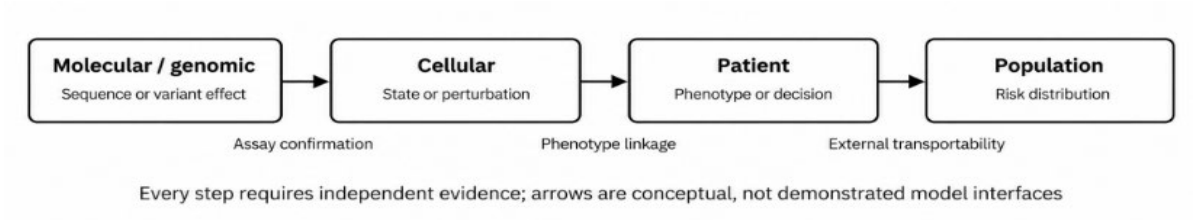

Figure 1: A proposed multiscale evidence pathway. Each transition adds an independent validation requirement; the diagram does not depict an implemented system

6. Seven Facets for Meaningful Comparison

The proposed framework organises biomedical AI models and their intended applications across seven complementary facets: (i) provenance, describing model development and adaptation; (ii) function, identifying intended tasks; (iii) modality, specifying input and output representations; (iv) domain, identifying biological or clinical context; (v) evaluation, documenting assessment methods and settings; (vi) trustworthiness, recording evidence concerning reliability, risks, limitations and safeguards; and (vii) deployment, distinguishing reported operating arrangements from potential delivery options. The facets form a non-hierarchical, multi-label descriptive framework rather than a validated measure of clinical readiness. Multiple attributes may apply, and undocumented information should not be interpreted as evidence of absence. Table 1 defines the facets, while Table 2 illustrates their application to four selected model–task examples.

Performance must be interpreted in relation to intended function, data representation, application domain and evaluation setting (Gallifant et al., 2025; Vasey et al., 2022). Provenance provides information relevant to reproducibility and adaptation, while deployment distinguishes research availability from institutional operation. The trustworthiness facet records available evidence and unresolved limitations without establishing overall safety or reliability (World Health Organization, 2024). Table 2 compares ESMFold, scGPT, NYUTron and MedGemma 4B across molecular, cellular and clinical applications (Lin et al., 2023; Cui et al., 2024; Jiang et al., 2023; Sellergren et al., 2026). Rather than ranking these models, the comparison shows how differences in tasks, evaluation evidence, and operating arrangements shape interpretations of demonstrated capabilities and associated risks.

Table 1: Seven complementary facets for interpreting biomedical AI models. Note: Facets may contain multiple, non-exclusive attributes; examples are illustrative

Facet	Guiding question	Illustrative attributes
Provenance	How was the model developed and adapted?	Training sources, pretraining, fine-tuning, architecture and version
Function	What task is the model intended to perform?	Generation, classification, annotation, prediction, and consultation
Modality	What information enters and leaves the model	Protein and DNA sequences, transcriptomes, EHR text, images, predicted structures, and risk scores
Domain	In which biological context or clinical context is it applied?	Molecular, genomic, cellular, hospital/patient, population
Evaluation	How and where has performance been demonstrated/assessed?	Benchmarks, laboratory tests, retrospective/prospective studies, simulation, intervention
Trustworthiness	What evidence addresses risk and reliability?	Uncertainty, Bias, calibration, privacy, oversight, monitoring
Deployment	Under what operating arrangements is it available?	Research release, local hosting, hosted service, institutional, controlled access, integration

Table 2: Illustrative application of the seven-facet framework to four biomedical model–task examples. Note: This comparison is descriptive, not a ranking of model performance or clinical readiness. 'Not established' indicates that the cited evidence does not support the specified claim, rather than proving it is absent. Sources: Lin et al. (2023); Cui et al. (2024); Jiang et al. (2023); Sellergren et al.l (2026)

Facet	ESMFold	scGPT	NYUTron	MedGemma 4B
Provenance	ESM-2 protein representations and structure-model training	Pretrained on over 33 million cells	Clinical-note pretraining and task-specific fine-tuning	Medical adaptation of Gemma 3
Function	Protein-structure prediction	Cell-type annotation	30-day readmission prediction	Medical image question answering
Modality	Protein sequence → 3D structure	Single-cell expression → cell label	EHR discharge text → risk score	Medical image and text → generated text
Domain	Molecular biology	Cellular biology	Hospital care	Medical imaging
Evaluation	Protein-structure benchmarks	Annotation benchmarks; other cellular tasks evaluated separately	Retrospective and prospective prediction; prospective AUC 0.787	Medical imaging and text benchmarks
Trustworthiness	Structural accuracy assessed; biological function not established	Task performance assessed; causal mechanisms not established	Calibration examined; reduction in readmissions untested	Benchmark limitations documented; clinical safety not established
Deployment	Research model	Research code and model	Institutional EHR-linked inference engine	Released model weights for research and adaptation

7. Interpreting the Evidence: Conditional Validity and Responsible Integration

7.1 Evaluation is Task- and Setting-dependent

The term 'validated' can encompass distinct evidentiary states, including computational benchmarking, laboratory confirmation, expert appraisal, simulated interaction, retrospective testing, prospective observation and interventional evaluation. NYUTron achieved an area under the receiver operating characteristic curve (AUC) of 0.787 in a prospective, single-arm, non-interventional readmission-prediction study involving 29,286 discharged encounters (Jiang et al., 2023). This establishes prospective predictive performance within one health system, not a causal reduction in readmissions. By contrast, Goh et al. (2024) randomised physician access to

GPT-4 during clinical-vignette assessments and found no statistically significant improvement in diagnostic-reasoning scores. These studies address different questions: one evaluates a prospective model predictions, while the other examines human–AI reasoning under controlled conditions. Neither demonstrates improved patient outcomes, nor does the absence of statistical significance establish equivalence. Their findings therefore cannot be reduced to a single measure of clinical readiness.

Evaluation must instead reflect the intended task and evidentiary claim. Protein language models may generate sequences with computationally predicted properties (Ferruz et al., 2022), whereas claims of biological function require experimental confirmation, as illustrated by the characterisation of ESM3-generated fluorescent proteins (Hayes et al., 2025). Cell-state and perturbation predictions require appropriate independent datasets and, where relevant, experimental validation (Theodoris et al., 2023; Cui et al., 2024). EHR prediction additionally requires assessment of calibration, transportability and prospective workflow performance, while clinical decision support requires evidence proportionate to its intended use and associated risks. DECIDE-AI provides guidance for reporting early-stage clinical evaluations, while TRIPOD-LLM addresses transparent reporting of LLM studies; neither independently establishes model validity (Vasey et al., 2022; Gallifant et al., 2025). Common benchmark leaderboards cannot substitute for these distinct evidentiary requirements.

7.2 Privacy and Deployment are Design Questions

Open-weight and hosted models offer different deployment arrangements rather than a universal privacy–capability trade-off. Open-weight systems may enable institutional hosting and greater control over data processing, subject to licensing, infrastructure, cybersecurity and maintenance requirements. Hosted services may offer managed updates and auditing capabilities but require scrutiny of contractual obligations, data handling and operational safeguards (World Health Organization, 2024). MedGemma shows that medical vision–language models with accessible weights can perform strongly on selected benchmarks, but neither model availability nor benchmark results support safe clinical deployment (Sellergren et al., 2026). Model access, clinical validity and data governance are therefore separate considerations.

Governance must address the complete operating arrangement, including data residency, consent, access controls, subgroup performance, calibration where applicable, human oversight, model updates, incident response and accountability. WHO guidance identifies broad ethical and governance considerations for large multimodal models, whereas FDA guidance on predetermined change control plans addresses modifications to regulated AI-enabled device software functions rather than biomedical research models generally (World Health Organization, 2024; US Food and Drug Administration, 2025). Distinguishing these scopes avoids implying regulatory authorisation or clinical readiness from model accessibility, technical capability or benchmark performance alone.

7.3 Integration Should Remain a Testable Proposition

Specialised biomedical models demonstrate potentially complementary capabilities across biological scales. Evo 2 predicts effects associated with selected genomic variants (Brixi et al., 2026); single-cell models investigate gene networks, cellular states and perturbation responses (Theodoris et al., 2023; Cui et al., 2024); and Delphi-2M estimates disease risks using longitudinal patient histories (Shmatko et al., 2025). However, these individual capabilities do not establish a validated multiscale system. Combining them may propagate uncertainty, introduce incompatible assumptions, or create errors at component interfaces. For example, a predicted variant effect could enter downstream analysis without adequate experimental confirmation, while differences in biological contexts and patient populations could undermine subsequent inference. These are plausible integration risks requiring direct investigation, not demonstrated failures of a particular pipeline.

A generalist model may also prove preferable to a specialised pipeline for a bounded task, but neither approach should be privileged without matched evaluation. Liévin et al. (2026) demonstrate agent-assisted disease-management reasoning in structured, multi-visit virtual encounters, providing evidence of performance under simulated conditions rather than real-world clinical effectiveness or molecular-to-clinical integration. Comparative evaluation of specialist pipelines and generalist alternatives should therefore use comparable cases, evidence access and oversight conditions while separately measuring predictive performance, safety, latency, cost and operational complexity. Interface-level assessment should additionally examine uncertainty propagation, conflicting recommendations, traceability and responsibility for correcting errors. This leads to the third proposition: cross-scale integration requires validating not only individual components but also the relationships and information exchanges that connect them.

8. Limitations and Considerations Towards Future Work

This narrative synthesis draws on 35 model-focused publications selected for explanatory contrast rather than exhaustive retrieval. Consequently, the examples cannot establish the prevalence of particular architectures, validation practices or deployment arrangements across biomedical AI research. The emphasis on English-language, indexed, and well-documented systems may introduce selection bias, while incomplete reporting of proprietary models, individual checkpoints, and institutional implementations further limits comparison. Findings across molecular, cellular and clinical domains therefore remain interpretive rather than statistically representative.

The proposed seven-facet framework has not been evaluated for independent coding reliability, completeness, or utility in decision-making. The four illustrative model–task comparisons demonstrate its descriptive application, not formal validation or comparative clinical readiness. Moreover, absent or incomplete reporting should not be interpreted as evidence that a capability or safeguard does not exist. Accordingly, the three propositions advanced in this synthesis remain analytical arguments requiring subsequent empirical investigation.

Future research should operationalise the seven facets through explicit coding definitions, evidence-status categories and reproducible assessment procedures. Applying these procedures to broader, systematically selected model collections could establish coding consistency, examine framework coverage and identify validation gaps. Further task-specific studies should evaluate cross-scale integration, uncertainty propagation and the comparative performance and safety of specialist and generalist architectures under matched conditions. Such work could develop the framework into a more standardised evidence-mapping instrument, while any use as a benchmarking or regulatory assessment tool would require additional validation.

9. Conclusion

Biomedical language-oriented models are best understood as scientifically distinct instruments with overlapping computational methods. The synthesis shows why architectural resemblance alone cannot bridge biological levels, why performance means vary with use and setting, and why integrations create additional obligations to test each evidentiary link. A seven-facet classification makes those distinctions visible without pretending that different tasks share a single validation standard. The most promising research agenda therefore is not to pursue a universal model, but to explicitly compare specialist and generalist approaches under relevant experimental, clinical, and governance conditions.

Ethics Declaration: Ethics approval was not required for the work described.

AI Declaration: Gemini, Grammarly and GPT-6 were used at different stages to assist with drafting, reorganisation, source organisation and language editing. Source claims were checked against available publications during revision.

References

Brixi, G. et al. (2026) “Genome modelling and design across all domains of life with Evo 2”, Nature, Vol 652, pp 1349–1361. https://doi.org/10.1038/s41586-026-10176-5.

Cui, H. et al. (2024) “scGPT: toward building a foundation model for single-cell multi-omics using generative AI”, Nature Methods, Vol 21, pp 1470–1480. https://doi.org/10.1038/s41592-024-02201-0.

Dalla-Torre, H. et al. (2025) “Nucleotide Transformer: building and evaluating robust foundation models for human genomics”, Nature Methods, Vol 22, No. 2, pp 287–297. https://doi.org/10.1038/s41592-024-02523-z.

Ferruz, N., Schmidt, S. and Höcker, B. (2022) “ProtGPT2 is a deep unsupervised language model for protein design”, Nature Communications, Vol 13, article 4348. https://doi.org/10.1038/s41467-022-32007-7.

Gallifant, J. et al. (2025) “The TRIPOD-LLM reporting guideline for studies using large language models”, Nature Medicine, Vol 31, No. 1, pp 60–69. https://doi.org/10.1038/s41591-024-03425-5.

Goh, E. et al. (2024) “Large language model influence on diagnostic reasoning: a randomized clinical trial”, JAMA Network Open, Vol 7, No. 10, article e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.

Hao, M. et al. (2024) “Large-scale foundation model on single-cell transcriptomics”, Nature Methods, Vol 21, pp 1481–1491. https://doi.org/10.1038/s41592-024-02305-7.

Hayes, T. et al. (2025) “Simulating 500 million years of evolution with a language model”, Science, Vol 387, pp 850–858. https://doi.org/10.1126/science.ads0018.

Jiang, L.Y. et al. (2023) “Health system-scale language models are all-purpose prediction engines”, Nature, Vol 619, pp 357–362. https://doi.org/10.1038/s41586-023-06160-y.

Kalfon, J. et al. (2025) “scPRINT: pre-training on 50 million cells allows robust gene network predictions”, Nature Communications, Vol 16, article 3607. https://doi.org/10.1038/s41467-025-58699-1.

Kraljevic, Z. et al. (2024) “Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study”, The Lancet Digital Health, Vol 6, No. 4, pp e281–e290. https://doi.org/10.1016/S2589-7500(24)00025-6.

Labrak, Y. et al. (2024) “BioMistral: a collection of open-source pretrained large language models for medical domains”, Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, pp 5848–5864. https://doi.org/10.18653/v1/2024.findings-acl.348.

Li, C. et al. (2023) “LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day”, Advances in Neural Information Processing Systems, Vol 36, pp 28541–28564. https://doi.org/10.52202/075280-1240.

Liévin, V. et al. (2026) “Towards conversational artificial intelligence for disease management”, Nature, Vol 655, pp 1292–1299. https://doi.org/10.1038/s41586-026-10764-5.

Lin, Z. et al. (2023) “Evolutionary-scale prediction of atomic-level protein structure with a language model”, Science, Vol 379, pp 1123–1130. https://doi.org/10.1126/science.ade2574.

Lu, M.Y. et al. (2024) “A visual-language foundation model for computational pathology”, Nature Medicine, Vol 30, pp 863–874. https://doi.org/10.1038/s41591-024-02856-4.

Luo, R. et al. (2022) “BioGPT: generative pre-trained transformer for biomedical text generation and mining”, Briefings in Bioinformatics, Vol 23, No. 6, article bbac409. https://doi.org/10.1093/bib/bbac409.

Moor, M. et al. (2023) “Med-Flamingo: a multimodal medical few-shot learner”, Proceedings of the 3rd Machine Learning for Health Symposium, Proceedings of Machine Learning Research, Vol 225, pp 353–367. https://proceedings.mlr.press/v225/moor23a.html.

Nath, V. et al. (2025) “VILA-M3: enhancing vision-language models with medical expert knowledge”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025), pp 14788–14798. https://openaccess.thecvf.com/content/CVPR2025/html/Nath_VILA-M3_Enhancing_Vision-Language_Models_with_Medical_Expert_Knowledge_CVPR_2025_paper.html.

Nguyen, E. et al. (2023) “HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution”, Advances in Neural Information Processing Systems, Vol 36. https://papers.neurips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html.

Nguyen, E. et al. (2024) “Sequence modeling and design from molecular to genome scale with Evo”, Science, Vol 386, article eado9336. https://doi.org/10.1126/science.ado9336.

Nijkamp, E. et al. (2023) “ProGen2: exploring the boundaries of protein language models”, Cell Systems, Vol 14, No. 11, pp 968–978.e3. https://doi.org/10.1016/j.cels.2023.10.002.

Peng, C. et al. (2023) “A study of generative large language model for medical research and healthcare”, npj Digital Medicine, Vol 6, article 210. https://doi.org/10.1038/s41746-023-00958-w.

Sellergren, A. et al. (2026) “An open vision-language model for diverse medical applications”, Nature Medicine, Vol 32, No. 10, pp 3695–3704. https://doi.org/10.1038/s41591-026-04626-w.

Shmatko, A. et al. (2025) “Learning the natural history of human disease with generative transformers”, Nature, Vol 647, pp 248–256. https://doi.org/10.1038/s41586-025-09529-3.

Singhal, K. et al. (2023) “Large language models encode clinical knowledge”, Nature, Vol 620, pp 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Singhal, K. et al. (2025) “Toward expert-level medical question answering with large language models”, Nature Medicine, Vol 31, pp 943–950. https://doi.org/10.1038/s41591-024-03423-7.

Theodoris, C.V. et al. (2023) “Transfer learning enables predictions in network biology”, Nature, Vol 618, pp 616–624. https://doi.org/10.1038/s41586-023-06139-9.

Tu, T. et al. (2024) “Towards generalist biomedical AI”, NEJM AI, Vol 1, No. 3. https://doi.org/10.1056/AIoa2300138.

Tu, T. et al. (2025) “Towards conversational diagnostic artificial intelligence”, Nature, Vol 642, pp 442–450. https://doi.org/10.1038/s41586-025-08866-7.

US Food and Drug Administration (2025). Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions, US Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence.

Vasey, B. et al. (2022) “Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI”, Nature Medicine, Vol 28, pp 924–933. https://doi.org/10.1038/s41591-022-01772-9.

World Health Organization (2024). Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models, World Health Organization, Geneva. https://www.who.int/publications/i/item/9789240084759.

Wu, C. et al. (2024) “PMC-LLaMA: toward building open-source language models for medicine”, Journal of the American Medical Informatics Association, Vol 31, No. 9, pp 1833–1843. https://doi.org/10.1093/jamia/ocae045.

Wu, C. et al. (2025) “Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data”, Nature Communications, Vol 16, article 7866. https://doi.org/10.1038/s41467-025-62385-7.

Yang, F. et al. (2022) “scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data”, Nature Machine Intelligence, Vol 4, pp 852–866. https://doi.org/10.1038/s42256-022-00534-z.

Yang, X. et al. (2022) “A large language model for electronic health records”, npj Digital Medicine, Vol 5, article 194. https://doi.org/10.1038/s41746-022-00742-2.

Zeng, Y. et al (2025) “CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells”, Nature Communications, Vol 16, article 4679. https://doi.org/10.1038/s41467-025-59926-5.

Zhang, H. et al. (2023) “HuatuoGPT, towards taming language model to be a doctor”, Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, pp 10859–10885. https://doi.org/10.18653/v1/2023.findings-emnlp.725.

Zhou, Z. et al. (2024) “DNABERT-2: efficient foundation model and benchmark for multi-species genomes”, Proceedings of the International Conference on Learning Representations (ICLR 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/b633e7052970b8f5aa1a69164d99e9e8-Abstract-Conference.html.