 © Authors of the article / Authors of the article

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ/ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI: https://doi.org/10.60797/COMP.2026.12.3 EDN: ZFXUCR

ARTIFICIAL INTELLIGENCE TOOLS FOR BUILDING SCIENTIFIC JOURNALS IN GENOMICS AND BIOTECHNOLOGY RESEARCH: A COMPREHENSIVE REVIEW AND ETHICAL FRAMEWORK

Review article

Budhyanto V.[1], Jurgens S.[2], Silalahi D.[3], Cicilia C.O.[4,] *, Rohadi H.[5], Retnomawarti R.[6]
[1] ORCID : 0009-0000-2176-6417;
[2] ORCID : 0000-0002-6175-4725;
[3] ORCID : 0000-0002-4786-4696;
[4] ORCID : 0009-0003-3231-4629;
[5] ORCID : 0009-0000-9738-5004;
[6] ORCID : 0009-0002-4624-2059;
[1, 2, 3, 4, 5, 6] Satriabudi Dharma Setia Foundation, Jakarta, Indonesia

* Corresponding author (cindy.o.cicilia[at]dharma.or.id)

Suggested: 16.07.2026; Accepted: 12.10.2026; Published: 29.10.2026

Abstract

Recent advances in large language models (LLMs), retrieval-augmented research agents, and biology-specific foundation models have transformed the scientific publishing workflow in genomics and biotechnology, yet differences in performance, reliability, accessibility, and ethical considerations continue to complicate tool selection, particularly in resource-constrained research environments. This review evaluates contemporary artificial intelligence (AI) tools that support the end-to-end scientific publishing process and proposes an ethical framework for their responsible use in biomedical research. A structured narrative review was conducted using peer-reviewed literature, preprints, benchmark reports, model cards, and international guidance documents published between January 2024 and May 2026. The reviewed tools were categorized into literature and research agents, biology-specific foundation models, and general-purpose frontier LLMs. The analysis found that frontier LLMs provide strong scientific reasoning capabilities, retrieval-augmented systems improve literature discovery and citation grounding, and biology-specific foundation models enhance genomic, structural biology, and multi-omics analyses. No single model was optimal across all tasks, rather, the most effective approach involved integrating complementary reasoning, retrieval, and domain-specific models within a unified workflow. Key ethical considerations included transparency of AI use, citation verification, data privacy, authorship accountability, and the distinction between legitimate biomedical research and dual-use research of concern. Overall, an integrated AI toolchain combining frontier LLMs, retrieval-grounded research agents, and biology-specific foundation models offers a practical and ethically responsible strategy for genomics and biotechnology publishing, provided that its implementation is guided by established principles from the International Committee of Medical Journal Editors (ICMJE), Committee on Publication Ethics (COPE), and World Health Organization (WHO) frameworks.

Keywords: artificial intelligence, large language models, genomics, bioinformatics, dual-use research.

ИНСТРУМЕНТЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ РАЗРАБОТКИ НАУЧНЫХ ЖУРНАЛОВ В ОБЛАСТИ ГЕНОМИКИ И БИОТЕХНОЛОГИИ: ВСЕСТОРОННИЙ ОБЗОР И ЭТИЧЕСКИЕ РАМКИ

Обзор

Будьянто В.[1], Юргенс С.[2], Силалахи Д.[3], Сисилия С.О.[4,] *, Рохади Н.[5], Ретномаварти Р.[6]
[1] ORCID : 0009-0000-2176-6417;
[2] ORCID : 0000-0002-6175-4725;
[3] ORCID : 0000-0002-4786-4696;
[4] ORCID : 0009-0003-3231-4629;
[5] ORCID : 0009-0000-9738-5004;
[6] ORCID : 0009-0002-4624-2059;
[1, 2, 3, 4, 5, 6] Фонд Сатриабуди Дхарма Сетиа, Джакарта, Индонезия

* Корреспондирующий автор (cindy.o.cicilia[at]dharma.or.id)

Предложена: 16.07.2026; Принята: 12.10.2026; Опубликована: 29.10.2026

Аннотация

Последние достижения в области больших языковых моделей (LLM), исследовательских агентов с функцией поиска и базовых моделей, специализированных для биологии, преобразовали рабочий процесс научных публикаций в области геномики и биотехнологии; однако различия в производительности, надёжности, доступности и этических соображениях по-прежнему затрудняют выбор инструментов, особенно в условиях ограниченных ресурсов. В данном обзоре дается оценка современных инструментов искусственного интеллекта (ИИ), поддерживающих процесс научных публикаций от начала до конца, и предлагается этическая концепция их ответственного использования в биомедицинских исследованиях. Был проведен структурированный нарративный обзор с использованием рецензируемой литературы, предварительных публикаций, отчетов по тестированию, описаний моделей и

 © Authors of the article / Authors of the article

международных руководящих документов, опубликованных в период с января 2024 года по май 2026 года. Рассматриваемые инструменты были разделены на следующие категории: литературные и исследовательские агенты, базовые модели для биологии и передовые LLM общего назначения. Анализ показал, что передовые LLM обладают мощными возможностями научного рассуждения, системы с расширенными возможностями поиска улучшают обнаружение литературы и обоснование цитирования, а базовые модели для биологии повышают эффективность геномного анализа, структурной биологии и мультиомического анализа. Ни одна из моделей не оказалась оптимальной для всех задач; напротив, наиболее эффективный подход заключался в интеграции взаимодополняющих моделей рассуждений, поиска и моделей, специализированных для конкретной области, в рамках единого рабочего процесса. К ключевым этическим аспектам относятся прозрачность использования ИИ, проверка цитирований, конфиденциальность данных, ответственность авторов, а также разграничение между легитимными биомедицинскими исследованиями и вызывающими озабоченность исследованиями двойного назначения. В целом, интегрированный набор инструментов ИИ, сочетающий передовые модели большого языка (LLM), исследовательские агенты, основанные на поиске информации, и базовые модели, специализированные на биологии, предлагает практическую и этически ответственную стратегию для публикаций в области геномики и биотехнологии при условии, что его внедрение будет осуществляться в соответствии с установленными принципами, закрепленными в документах Международного комитета редакторов медицинских журналов (ICMJE), Комитета по публикационной этике (COPE) и Всемирной организации здравоохранения (ВОЗ).

Ключевые слова: искусственный интеллект, большие языковые модели, геномика, биоинформатика, исследования двойного назначения.

Introduction

Scientific publishing in genomics and biotechnology faces three major challenges: an exponentially growing primary literature, with PubMed indexing more than 150 publications containing the keyword "foundation model" in 2024 alone, up from fewer than ten per year before 2023 [1]. With rising methodological complexity that requires the integration of bulk sequencing, single-cell omics, structural biology, and clinical metadata, and persistent under-resourcing of laboratories in low- and middle-income countries (LMICs), where many neglected and endemic diseases, including tuberculosis (TB), for which Indonesia ranks second globally for incidence remain prevalent [2], [3].

This study is a structured narrative review with an explicit comparative-evaluation design, rather than a PRISMA-compliant systematic review or meta-analysis. The narrative format was chosen because the evidence base spanning peer-reviewed articles, preprints, vendor technical reports, and frequently updated benchmark leaderboards is heterogeneous and not amenable to statistical pooling, while the objective is to compare tools and synthesize governance recommendations rather than to estimate a single effect size. Searches and data extraction were conducted between February and May 2026 and were restricted to sources published or released between January 2024 and May 2026.

These developments are not, however, uniformly accessible or aligned with the realities of LMIC biomedical research. An emerging concern is the potential for excessive safety restrictions: certain frontier models, some commercially deployed frontier models, have been documented to decline requests related to computational structural biology, cybersecurity teaching, and routine bioinformatics requests at rates that potentially affect research efficiency [14], [15]. For researchers designing amplicons for tuberculosis nucleic acid amplification assays, a routine diagnostic activity overly restrictive model responses may unnecessarily limit access to legitimate scientific assistance, with potentially greater impacts on investigators working in resource-constrained, high-disease-burden settings.

Editorial and ethical guidance has begun to respond: the International Committee of Medical Journal Editors (ICMJE), the Committee on Publication Ethics (COPE), and the World Health Organization (WHO) have each issued recommendations on AI authorship, disclosure, and dual-use research. What is still missing is an integrated, operational account that simultaneously maps the fast-changing AI tool landscape across the entire genomics publishing workflow, anchors comparative claims to explicitly cited benchmarks rather than vendor marketing, and reconciles legitimate diagnostic and therapeutic biotechnology with dual-use research of concern (DURC) in a form usable by a resource-constrained institution. Existing reviews tend to treat tool capability, citation reliability, ethical compliance, and equity of access in isolation, and none provides a consolidated, governance-ready framework calibrated to a high-TB-burden LMIC such as Indonesia. Addressing that gap is the purpose of this review.

This review (i) provides a comprehensive overview of the current AI tool landscape across the full publication workflow, (ii) provides comparative performance data anchored to benchmark numbers and explicitly cited primary sources, (iii) articulates an ethical framework that explicitly distinguishes legitimate diagnostic and therapeutic biotechnology from DURC, and (iv) recommends a *tiered toolchain* appropriate for an Indonesian biotechnology foundation operating under ICMJE, COPE, WHO, and Indonesian Ministry of Health (Kemenkes) governance.

Methods

This study is a structured narrative review with an explicit comparative-evaluation design, rather than a PRISMA-compliant systematic review or meta-analysis. The narrative format was chosen because the evidence base, whereas spanning peer-reviewed articles, preprints, vendor technical reports, and frequently updated benchmark leaderboards is heterogeneous and not amenable to statistical pooling, while the objective is to compare tools and synthesize governance recommendations rather than to estimate a single effect size. Searches and data extraction were conducted between February and May 2026 and were restricted to sources published or released between January 2024 and May 2026.

For each tool, we extracted and evaluated: developer, release date and current version, model parameters and context-window specifications, principal benchmark scores, licensing and pricing, agentic capability, and reported failure modes. When vendor-reported benchmark results differed from independent evaluations, such discrepancies were explicitly noted and

 © Authors of the article / Authors of the article

discussed. No external funding was received for this review. YSDS Research Team All literature screening, manuscript preparation, and verification were performed internally by the authors.

2.1. Information Sources, Search Tools, and Materials

Because this is a literature-based review, no laboratory instruments, reagents, or biological materials were used, the analogous "materials" are the bibliographic corpora and primary documents analyzed, and the "tools" are the search and benchmarking platforms used to retrieve and verify them. Five categories of source were consulted:

(i) peer-reviewed primary literature indexed in PubMed and Web of Science, using the search strings "large language model" AND ("genomics" OR "biotechnology" OR "literature review"], "foundation model" AND "biology", and "AI" AND ["ICMJE" OR "COPE" OR "publication ethics");

(ii) preprints on bioRxiv and arXiv;

(iii) vendor technical reports and model cards from OpenAI, Google DeepMind, Anthropic, Meta, xAI, Alibaba, DeepSeek, Mistral, EvolutionaryScale, Arc Institute, InstaDeep, FutureHouse, and the Allen Institute for AI;

(iv) public benchmark leaderboards, including Artificial Analysis, Vellum, the Epoch AI Benchmarking Hub, and the Hugging Face InstaDeep Nucleotide Transformer leaderboard;

(v) international guidance documents, namely the WHO 2022 Global Guidance Framework, the ICMJE Recommendations (January 2024 update), and the COPE 2023 and 2024 position statements on AI and authorship.

2.2. Exclusion and Verification Criteria

A recurring concern in this literature is that many widely cited performance figures originate from the organizations selling the tools. To avoid treating such claims as equivalent to independently validated evidence, extracted claims were classified into four evidence tiers, with the tier determining their eligibility for comparative and ranking statements. Claims without a stated evaluation protocol were classified as Tier D, claims based on benchmarks that were not publicly specified or obtainable, or that used unnamed or non-version-pinned comparators, were excluded from comparative analyses. Where both vendor and independent results were available, independent estimates were prioritized, while vendor figures were reported only when materially different. Vendor evaluations using vendor-authored benchmarks were retained as Tier C, whereas community leaderboards lacking documented submission and verification procedures were excluded. Accordingly, Undermind retrieval-enrichment and relevance-classification results were classified as Tier C and excluded from rankings, FutureHouse/PaperQA2 and OpenScholar results were classified as Tier B because their evaluation sets were developed by the respective groups, and an independent eight-chatbot citation study was classified as Tier A and used to support conclusions on citation reliability. Community-based deep-research error estimates were excluded because their evaluation and verification procedures were insufficiently documented.

2.3. Research Procedure

Step I. Source identification. Candidate sources were identified by running the search strings above across the bibliographic databases and preprint servers, supplemented by hand-searching the reference lists of key articles and the official release pages of each model developer.

Step II. Screening and eligibility. Records were screened for relevance to the genomics and biotechnology publishing workflow. A source was eligible if it reported tool capabilities, benchmark performance, licensing terms, or governance guidance relevant to one or more workflow stages (ideation, literature review, sequence or structure analysis, drafting, or peer review). Tools without verifiable documentation, and benchmark claims that could not be traced to a primary source, were excluded from the comparative tables.

Step III. Data extraction and charting. For each tool, the following items were extracted into a common charting template: developer, release date and current version, model parameters and context-window specifications, principal benchmark scores, licensing and pricing, agentic capability, and documented failure modes. Extracted items were grouped into three tool families, including literature and research agents, biology-specific foundation models, and general-purpose frontier LLMs that were summarized in Tables 1–3.

Step IV. Evidence synthesis. The charted data were synthesized narratively by workflow stage to derive the recommended tiered toolchain and the institutional ethics recommendations. Where vendor-reported benchmark results diverged from independent evaluations, both figures were reported, and the discrepancy was discussed rather than resolved by selecting a single value.

2.4. Ethical Clearance

This review analyzed only published literature, preprints, publicly released technical documentation, and aggregate benchmark data, it involved no human participants, no animal subjects, and no identifiable personal or genomic data. Institutional ethics-committee approval was therefore not required. Any subsequent empirical work that operationalizes the recommendations herein particularly studies involving human genomic data under the IGNITE network will require separate ethical approval from the relevant institutional review board and must comply with Indonesia's Personal Data Protection Law (UU PDP No. 27/2022).

2.5. Data Analysis

No statistical meta-analysis was performed, consistent with the heterogeneous and non-randomized nature of the evidence. Analysis was qualitative and comparative: benchmark figures were tabulated on a common scale to permit cross-tool comparison, tools were ranked within each workflow stage by a transparent combination of capability, citation reliability, licensing openness, and suitability for data-sovereign deployment, and recurrent failure modes notably citation fabrication and over-restriction were coded thematically across sources. All literature screening, data extraction, verification of citations against Crossref and PubMed, and manuscript preparation were performed internally by the authors. No external funding was received for this review.

 © Authors of the article / Authors of the article

Result and discussion

This section reports the findings of the comparative review and interprets them against the existing literature and the operational realities of LMIC genomics research. Following the structure of the evidence, the results are presented in five parts: AI tools for literature, ideation, and citation workflows (Section 3.1), biology-specific foundation models (Section 3.2), general-purpose frontier LLMs and their willingness to engage with diagnostic biotechnology (Section 3.3), the integration of these tools across the publication cycle (Section 3.4), and the ethical framework that follows from them (Section 3.5). For each tool family, we first present the key capability and benchmark data (summarised in Tables 1–3), then interpret what those data imply for genomics publishing, compare them with previously published evaluations, offer mechanistic explanations for the observed performance and failure modes, and close with the broader implications for resource-constrained, high-disease-burden settings. The over-restriction of legitimate diagnostic requests and the design of Mycobacterium tuberculosis amplicons recur throughout as concrete tests of whether each tool and the policy governing it serves equitable global-health research.

3.1. AI tools for literature, ideation, and citation workflows

3.1.1. Retrieval-augmented scientific agents

OpenScholar (Allen Institute for AI / University of Washington, released November 2024) combines a specialized 8-billion-parameter retrieval-augmented LM with the OpenScholar-DataStore (OSDS), which is reported by its developers to contain approximately 45 million open-access papers from Semantic Scholar, along with 237 million corresponding passage embeddings. On the developers' ScholarQABench, OpenScholar achieved a Citation F1 score of 81% on the ScholarQABench benchmark, and in blinded PhD-level evaluators preferred OpenScholar-8B and OpenScholar-GPT4o responses over expert-authored responses 51% and 70% of the time respectively, versus 32% for unaugmented GPT-4o, which produces unsupported or fabricated citations in 78–90% of evaluated cases [12].

FutureHouse Platform — Crow, Falcon, Owl, Phoenix, Finch, Robin (launched 1 May 2025) is built on the open-source PaperQA2 system. According to published benchmark evaluations, PaperQA2 was reported to achieve higher retrieval accuracy than participating PhD- and postdoctoral-level researchers under the evaluated conditions at retrieving information from the scientific literature, as measured using LitQA2, a piece of the LAB-Bench evaluation set [13], on RAG-QA-Arena (science portion), PaperQA2 was reported to be 12.4 percentage points higher than the next-best system [13]. Crow functions as a literature question-answering agent, Falcon performs deep multi-document synthesis with access to OpenTargets and other databases, Owl evaluates whether similar ideas have previously been reported in the literature, Phoenix assists in planning chemistry experiments. The platform is freely accessible via web user interface and application programming interface (API) with rate limits and has been used by independent groups for systematic reviews of Parkinson's disease genes, reportedly outperforming general-purpose agents in the evaluated tasks [16].

Undermind AI, founded by two Massachusetts Institute of Technology quantum physics doctoral graduates, uses an iterative LLM-based search strategy designed to mimic aspects of human literature exploration. The company's whitepaper reports a ten-fold enrichment of highly relevant search results within the top hits relative to Google Scholar and reported approximately 98% classification accuracy under the company's evaluation protocol in classification of paper relevance, However, these results originate from self-published evaluations and have not yet undergone peer review [17].

Elicit, SciSpace, Consensus, Scite.ai, ResearchRabbit, Semantic Scholar, and Perplexity Pro serve partially overlapping functions (Table 1). Independent published evaluations show that, for structured evidence extraction across many papers (for example, PICO tables), Elicit demonstrated the strongest performance in evaluated evidence-extraction tasks, SciSpace supports a broader range of activities spanning literature discovery and manuscript drafting, and Consensus is particularly effective for rapid evidence synthesis [18]. Scite provides citation classification features that categorize citations as supporting, contrasting, or merely mentioning—a feature potentially useful for verifying claims before clinical or guideline use. ResearchRabbit currently offers a free-access model and integrates two-way with Zotero, generating visual citation-network maps from seed papers [19]. Perplexity Pro and Perplexity Deep Research are efficient and user-friendly but have been reported to generate citation inaccuracies on academic tasks: a 2025 study of eight AI chatbots in bibliographic reference retrieval reported that only 26.5% of the references were fully correct, 33.8 % partially correct, and 39.8% were either erroneous or entirely fabricated, with Grok and DeepSeek the were the only systems in that evaluation that did not produce false references, while Microsoft Copilot, Perplexity, and Claude showed the highest rates of reference-generation errors [20]. Community benchmarks have similarly estimated Perplexity Deep Research (R1 mode) substantially higher error rates than competing systems on equivalent research tasks [21]. Perplexity-generated citations should therefore be independently verified prior to scholarly use.

STORM (Stanford OVAL, NAACL 2024) synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking produces source-supported, long-form articles by simulating multi-perspective expert conversations grounded on curated internet sources. Versus an outline-driven retrieval-augmented generation (RAG) baseline on the FreshWiki dataset, STORM articles were evaluated as 25% more organized and 10% broader in coverage. STORM is open-source and may be particularly useful for drafting introduction and background sections of review articles, but the authors explicitly identify propagation of source-related biases and spurious associations between unrelated facts as limitations [22].

3.1.2. Comparative summary

Table 1 - AI tools for literature review and citation workflows (as of May 2026)

DOI: https://doi.org/10.60797/COMP.2026.12.3.1

Tool	Developer	Principal strength	Citation accuracy / failure mode	Access
OpenScholar-8B / -	Ai2 / UW	Citation F1 81% on	Best citation	Open weights, demo

 © Authors of the article / Authors of the article

Tool	Developer	Principal strength	Citation accuracy / failure mode	Access
GPT4o		ScholarQABench, experts preferred over expert-written 51–70%	grounding among open systems	free
PaperQA2 / Crow / Falcon	FutureHouse	Outperforms PhD biologists on LitQA2, SOTA on RAG-QA-Arena science	Strong, full-text access to papers	Free web + API
Elicit	Elicit Labs	Structured PICO data extraction	Good, manual verification still required	Free + paid (~USD 10/mo)
SciSpace	SciSpace (Typeset)	Chat-with-PDF, multi-source AI search	Moderate	Free + paid (from USD 12/mo)
Consensus	Consensus	<30-second evidence synthesis with consensus meters	Moderate	Free + Pro (from USD 15/mo)
Scite.ai	Scite	Citation context (supporting/contrasting)	Excellent for context	Free + paid (from USD 12/mo)
ResearchRabbit	ResearchRabbit Inc.	Visual citation graphs, Zotero two-way sync	n/a (discovery)	Free
Semantic Scholar	Ai2	214M+ paper index, free API	Strong baseline	Free
Perplexity Pro / Deep Research	Perplexity AI	Fast general search	Among the highest hallucination rates in a peer-reviewed test of eight chatbots [20]	USD 20/mo Pro
Undermind	Undermind Inc.	Iterative agentic search	Self-reported only	Paid subscription
STORM	Stanford OVAL	Long-form grounded outlines	Source-bias transfer	Open-source

3.2. Biology-specific foundation models

3.2.1. Genomic and DNA foundation models

Evo 2 Developed via a collaboration between the Arc Institute and NVIDIA, Evo 2 represents a 40-billion-parameter genomic foundation model built upon the StripedHyena 2 architecture. By leveraging a 1-megabase context length and trained on a large-scale genomic corpus comprising approximately 9.3 trillion DNA base pairs from over 128,000 species, Evo 2 achieves 90% accuracy in classifying BRCA1 variant pathogenicity. Additionally, the model supports multi-modal predictions (DNA, RNA, and proteins) and generative genomic design. To ensure biosecurity, eukaryotic viruses were excluded from training, with red-team evaluations verifying the safety of the open-source weights now hosted on NVIDIA BioNeMo and Amazon Web Services.

AlphaFold 3 Advancing beyond the protein-centric predictions of its predecessor, AlphaFold 3 models complex biomolecular interactions, including protein couplings with ligands, nucleic acids, antibodies, and ions. The architecture yields at least a 50% accuracy improvement over conventional prediction and physics-based docking methods, notably doubling performance in specific interaction categories within the PoseBusters ligand benchmark. Academic and non-commercial utilization of AlphaFold 3 is supported through restricted code releases and the AlphaFold Server, while structurally equivalent models such as Boltz-1 and Chai-1 offer broader utility under permissive open licenses.

ESM-3 and ESM C Recent advancements in biological foundation models are highlighted by ESM-3/ESM-C and Nucleotide Transformer V2. ESM-3 and ESM-C are multimodal generative protein language models trained on 2.78 billion proteins using one trillion teraflops of compute. They enable simultaneous reasoning over protein sequence, structure, and function, as evidenced by the *de novo* synthesis of esmGFP—a novel green-fluorescent protein exhibiting 58% identity to natural relatives, representing roughly 500 million years of evolutionary divergence. While its base 1.4-billion-parameter model is open-source, larger variants up to 98 billion parameters are deployed via the EvolutionaryScale Forge API and AWS/NVIDIA.

 © Authors of the article / Authors of the article

Nucleotide Transformer V2 For nucleotide-level analysis, Nucleotide Transformer V2 offers a family of genomic models (50M to 2.5B parameters) trained on human and multispecies reference genomes. Its largest 2.5B-parameter variant outperforms DNABERT-2, HyenaDNA, and Enformer across 18 benchmarks.

DNABERT-2, HyenaDNA, Caduceus-Ph, and GROVER. Comparative benchmarking of prominent DNA foundation models (DNABERT-2, HyenaDNA, Caduceus-Ph, and GROVER) has highlighted significant trade-offs based on architectural design. On human-genome benchmarks, DNABERT-2 shows robust consistency, whereas Nucleotide Transformer V2 offers superior sensitivity in epigenetic-modification detection. For long-range genomic dependencies, HyenaDNA leverages a long-convolution framework to process inputs up to 1 million tokens, conversely, the bidirectional Mamba-based Caduceus-Ph is highly effective for localized, single-nucleotide variant tasks. Despite these advancements, evidence indicates that general-purpose models underperform specialized tools in complex downstream applications, such as gene-expression prediction and causal quantitative-trait loci identification, emphasizing the continued necessity of task-specific architectures in genomic pipelines.

3.2.2. Single-cell and gene-expression models

Geneformer, scGPT, scFoundation, scBERT, and GenePT. Prominent foundation models in the single-cell domain include Geneformer, scGPT, scFoundation, scBERT, and GenePT. Recent iterations such as Geneformer V2-316M and scGPT utilize expansive training datasets, scaling up to 104 million and 33 million single-cell transcriptomes, respectively. Despite supporting diverse tasks like multi-omic integration and cell-type annotation, their out-of-the-box utility remains constrained, independent benchmarks indicate that zero-shot implementations can be outperformed in clustering by traditional pipelines like scVI or Harmony. To circumvent expensive pre-training architectures, alternative approaches like GenePT utilize NCBI gene description embeddings to achieve comparable downstream efficacy. These findings underscore a critical methodological implication for genomic pipelines: foundation models must undergo targeted fine-tuning for specific downstream applications rather than relying on baseline zero-shot embeddings for novel biological discovery.

3.2.3. Biomedical text and chemistry models

BioGPT In the biomedical and chemical domains, BioGPT represents a specialized GPT-2-based architecture pre-trained on PubMed literature. The architecture scales from a 355-million-parameter base version to a 1.5-billion-parameter large variant, achieving strong performance metrics across relation-extraction tasks (BC5CDR, KD-DTI, and DDI) and reaching up to 81.0% accuracy on the PubMedQA benchmark. This paradigm is further complemented by ProtGPT2 and ChemBERTa, which process non-textual modalities including generative protein sequences and SMILES chemical strings, respectively. While structurally smaller than current general-purpose foundation models, these specialized architectures provide distinct advantages for targeted workflows, offering robust performance and enabling secure, on-premises deployment where strict data privacy is required.

3.2.4. Primer and amplicon design augmented by AI

While AI-native primer design platforms have yet to surpass the empirical reliability of traditional thermodynamic modeling pipelines (e.g., NCBI Primer-BLAST, Primer3Plus, and OligoAnalyzer), frontier reasoning LLMs like GPT-5 and Gemini 2.5 Pro offer valuable scaffolding capabilities. These advanced models streamline the design workflow by recommending conserved genomic targets, defining primer-pair parameters (GC content, melting temperature, and secondary structure avoidance), assessing off-target specificity, and generating *in silico* validation frameworks. This hybrid approach is highly applicable to diagnostic amplicon design for pathogens such as *Mycobacterium tuberculosis*, where targeting established loci including *IS6110*, *rpoB*, *katG*, *inhA*, and *hsp65* is standard practice. This integration of LLM assistance directly supports the development pipeline, with specific technical constraints regarding over-restriction addressed in Section 3.5.

3.3. General-purpose frontier LLMs and willingness to engage with diagnostic biotechnology

3.3.1. Benchmark performance on graduate-level scientific reasoning

Table 2 - Frontier LLM performance on graduate-level scientific reasoning (GPQA Diamond) and key context and release data, as of May 2026

DOI: https://doi.org/10.60797/COMP.2026.12.3.2

Model	Developer	Release	GPQA Diamond	Context window	Notes on willingness for diagnostic biotech
GPT-5 / GPT-5 Pro	OpenAI	Aug 2025	88.4–89.4% (Pro, no tools)	256K–400K	Generally engaged, medical-domain stress-tested
GPT-5.2 / GPT-5.2 Pro	OpenAI	Late 2025	92.4–93.2%	400K	State-of-the-art reasoning
Gemini 2.5 Pro	Google DeepMind	Mar 2025	~86.4%	1,048,576 (1M) tokens	Strong AI Co-Scientist multi-agent (Nature 2025)
Grok 4 / Grok 4 Heavy	xAI	Jul 2025	87–88.9% (Heavy)	256K (API)	Less restrictive on scientific

 © Authors of the article / Authors of the article

Model	Developer	Release	GPQA Diamond	Context window	Notes on willingness for diagnostic biotech
					content
Claude Opus 4.5	Anthropic	Late 2025	~87%	200K	Documented over-restriction, 0.23 % benign refusal vs. 0.05% Sonnet 4.5
DeepSeek V3 / V3.1 / R1	DeepSeek-AI	Dec 2024 – 2026	68.4 → 81 (R1-0528)	128K	Open-weights MIT license
Llama 4 Scout / Maverick	Meta	2026	~75–80 %	10M (Scout)	Open weights, best for self-hosted long context
Mistral Large 3	Mistral AI	2026	~75%	128K	EU-based, Apache 2.0 (Large 3 / Small 4)
Qwen 3 / 3.5	Alibaba	2026	81.7 (Qwen 3.5-9B), 90+ flagship	Up to 1M	Apache 2.0, multilingual (29–201 languages)

Frontier reasoning improvements have compressed what was a 6- to 12-month lead for closed models into a 2- to 3-month lead. Per the U.S. National Institute of Standards and Technology Center for AI Standards and Innovation evaluation published in May 2026, CAISI evaluations indicate that DeepSeek V4's capabilities lag behind the frontier by about eight months [29]. Qwen 3.5 and Llama 4 had not been formally CAISI-evaluated at the time of that report.

3.3.2. The over-restriction problem-empirical evidence

Implementing LLMs in diagnostic biotechnology requires careful management of the 'over-restriction' problem, wherein safety classifiers mistake legitimate scientific prompts for policy violations. Benchmarking data reveals that alignment guardrails vary significantly across architectures, for instance, Claude-3.5-Sonnet demonstrates a 73% refusal rate on dual-use chemistry benchmarks, whereas models like GPT-3.5-turbo and Grok-2 exhibit substantially lower refusal rates at 10% and 20%, respectively, and Mistral processes all queries without intervention. This behavioral gradient indicates that highly aligned models, particularly within the Claude 4.x family, frequently trigger false positives on structural biology and biochemical tasks, an issue further compounded by extended thinking features. Consequently, for uninterrupted genomic workflows and PCR amplicon design, this pipeline prioritizes the utilization of low-refusal architectures, including GPT-5/5.2, Gemini 2.5 Pro, and open-weights models (DeepSeek V3.1, Qwen 3.5, Llama 4), while restricting the use of Claude models strictly to non-technical tasks such as manuscript editing and ethical-section drafting.

3.3.3. Agentic deep research

Gemini AI Co-Scientist. Recent advancements in autonomous discovery are highlighted by frontier agentic systems like Gemini AI Co-Scientist and Grok 4 Heavy. Gemini AI Co-Scientist employs a multi-agent framework utilizing Gemini 2.5 to autonomously synthesize, debate, and refine scientific hypotheses through orchestrated generation, reflection, proximity, and ranking steps [31]. The scalability of this approach has led to its deployment across major research institutions, including all 17 U.S. National Laboratories [32]. Similarly, Grok 4 Heavy utilizes a tool-enabled multi-agent architecture that achieves a 44.4% score on the challenging 'Humanity's Last Exam' benchmark, substantially higher than Gemini 2.5 Pro at 26.9%, while registering high proficiencies on AIME (100%) and GPQA (87%) [33]. While these commercial models demonstrate state-of-the-art reasoning, the underlying agentic workflows can be computationally emulated within academic pipelines by integrating open-access resources like FutureHouse Crow and the Falcon API.

3.4. Workflow Integration Across the Publication Cycle

 © Authors of the article / Authors of the article

Table 3 - Recommended AI tool by workflow stage for genomics and biotechnology manuscripts

DOI: https://doi.org/10.60797/COMP.2026.12.3.3

Stage	Recommended primary tool	Backup or specialist	Cost / API Constraints	Notes
1. Ideation / hypothesis generation	Gemini AI Co-Scientist, GPT-5 Pro	FutureHouse Owl (prior-art check)	Frontier consumer subscription ≈ USD 20–25 per user per month, premium reasoning tiers roughly an order of magnitude higher. AI Co-Scientist is not generally available, so its effective cost to an unaffiliated institution is access-limited rather than priced.	Co-Scientist debate agents particularly strong for novel hypotheses
2. Literature review	OpenScholar, FutureHouse Crow/Falcon	Elicit (structured tables), Semantic Scholar	OpenScholar open weights, free, self-hosting the 8B model needs ≈1×24 GB accelerator. FutureHouse free at web and API tier. Elicit freemium, paid ≈ USD 10–12 per user per month.	Citation accuracy matters more than speed
3. Citation management	Zotero + ResearchRabbit two-way sync	Scite.ai for citation context	Zotero and ResearchRabbit free, Zotero cloud storage metered above the free quota. Scite ≈ USD 12–20 per user per month or institutional licence. Crossref and PubMed APIs free.	Always verify against Crossref/DOI
4. Primer / amplicon design	Primer-BLAST + GPT-5/Gemini for criteria reasoning	Evo 2 for sequence-level checks	Primer-BLAST and Primer3Plus free (NCBI / public). LLM scaffolding cost as row 1. Evo 2 metered by GPU-hour in cloud, or capital cost on-premises.	LLM should augment, not replace, thermodynamic tools
5. CRISPR guide design	CRISPOR / CHOPCHOP + LLM scaffolding	Evo 2 / Nucleotide Transformer V2 (off-target)	CRISPOR and CHOPCHOP free web services. Nucleotide Transformer V2 2.5B runs on ≈1×40 GB accelerator, Evo 2 as row 4.	Off-target prediction remains the limiting step, treat model output as a shortlist for wet-lab validation
6. Bioinformatics analysis	GPT-5 / Gemini 2.5 Pro for code	DeepSeek V3.1 / Llama 4 on-premise for sensitive data	Frontier API metered per million tokens. **Budget caveat:** self-hosting flagship DeepSeek V3.1 is not a low-	Verify outputs against GATK, nf-core

 © Authors of the article / Authors of the article

Stage	Recommended primary tool	Backup or specialist	Cost / API Constraints	Notes
			cost option it requires a multi-accelerator node even quantised. Llama 4 Scout and 8–32B-class models are the realistic sovereign choices at foundation scale.	
7. Single-cell / multi-omics	scGPT / Geneformer (fine-tuned)	scVI/Harmony baselines	All open-source and free. The real cost is fine-tuning compute (≈1×24-48 GB accelerator) plus expert curation time for labelled data.	Avoid zero-shot foundation-model embeddings for novel discovery [26]
8. Protein structure / interaction	AlphaFold 3 (Server, free non-commercial)	Boltz-1, Chai-1 (open-source clones), ESM-3	AlphaFold Server free for non-commercial use. Boltz-1 and Chai-1 free under permissive licences, ≈1×24–40 GB accelerator. ESM-3 largest variants metered via Forge API.	Choose the permissively licensed equivalents wherever commercialisation is foreseeable
9. Data interpretation/visualization	GPT-5 / Gemini 2.5 Pro	Claude (best plotting code style)	Covered by the row 1 subscription, no incremental cost.	Verify all numerical claims
10. Manuscript drafting	Claude Opus 4.5 (writing quality)	GPT-5, Gemini 2.5 Pro	Frontier consumer subscription ≈ USD 20–25 per user per month.	Disclose per ICMJE, never list AI as author
11. Statistical analysis	GPT-5 with Python tool	DeepSeek R1 for cost-sensitive	Covered by the row 1 subscription where code execution is included. DeepSeek R1 open weights free, but subject to the same multi-accelerator hosting constraint as row 6.	Rerun in R/Python with auditable scripts
12. Figure generation	BioRender, matplotlib via GPT-5	DALL-E 3 / Imagen 3 for schematic	matplotlib free. **BioRender is the significant recurring cost:** individual academic plans run to the low hundreds of USD per year and institutional licences higher and publication rights require a paid plan, as the free tier does not license figures for journal publication.	Disclose any AI-generated images
13. Peer-review	Claude Opus 4.5 for	GPT-5 for technical	Covered by the row	Do not upload

 © Authors of the article / Authors of the article

Stage	Recommended primary tool	Backup or specialist	Cost / API Constraints	Notes
preparation	rebuttal tone	depth	10 subscription.	confidential comments to public models (ICMJE 2024)
14. Journal selection	Elsevier Journal Finder, Springer Nature Journal Suggester, SciSpace match	Manual scope review	Publisher finders and JANE free, SciSpace freemium.	Never rely on a single publisher's suggestion as a scope assessment
15. Plagiarism / AI-detection	iThenticate (canonical)	Originality.ai (AI-text screening)	iThenticate is normally reached through an institutional Similarity Check membership, standalone per-manuscript pricing is material for an institution submitting at volume. Originality.ai is credit-based.	Disclose AI use upfront — best defence

3.5. Ethical Framework

3.5.1. The Indonesian biotechnology research context

The optimization of diagnostic infrastructure is critical in Indonesia, which retains the second-highest global burden of TB amidst significant geographic heterogeneity and underreporting. Addressing calls for domestic diagnostic innovation, localized studies in West Java have validated the efficacy of regional tools, for instance, local interferon-gamma release assays show equivalence to international standards, while targeted next-generation sequencing workflows exhibit a 91–100% positive percent agreement with CBNAAT and line-probe assays for multi-drug resistance screening. To scale these genomic breakthroughs under constrained laboratory settings, the integration of AI-assisted tools for amplicon design, primer specificity verification, and variant interpretation is imperative. Rather than serving as peripheral enhancements, these computational workflows act as vital mechanisms that elevate the analytical throughput and global competitiveness of domestic research initiatives.

This context also underscores the importance of developing and benchmarking domain-specific Small Language Models (SLMs) tailored to Indonesian medical and biomedical texts. General-purpose and frontier LLMs are predominantly optimized for English-dominant corpora and may therefore exhibit language-specific inaccuracies when processing Indonesian medical terminology, clinical expressions, abbreviations, or mixed Indonesian-English biomedical content. SLMs trained or fine-tuned on high-quality Indonesian biomedical corpora could improve terminology fidelity and contextual accuracy while substantially reducing computational, infrastructure, and inference costs compared with frontier-scale models. Their development should therefore be accompanied by standardized Indonesian-language benchmarks covering biomedical terminology, clinical reasoning, scientific information retrieval, diagnostic interpretation, and genomic workflows. Such benchmarking would determine whether smaller, locally adapted models can provide sufficiently reliable performance for Indonesian biomedical research while offering a more computationally accessible and deployable alternative for resource-constrained institutions.

3.5.2. Distinguishing legitimate research from dual-use research of concern

According to the WHO 2022 global guidance framework, Dual-Use Research of Concern (DURC) involves life-science research with a high potential for misappropriation that could threaten public health, agriculture, or national security [37], [38]. Within this ethical framework, routine diagnostic biotechnology is distinct from DURC. Specifically, the development of PCR amplicons for *Mycobacterium tuberculosis* (a Risk Group 3 pathogen) relies on well-established, publicly accessible genomic data and WHO-approved diagnostic baselines, such as the GeneXpert MTB/RIF platform. Automated AI alignment systems that refuse to process these requests fail to distinguish benign diagnostic development from actual biological risks. Because the data generated in routine primer design offers no novel risk beyond existing tools like NCBI Primer-BLAST, these diagnostic workflows should be classified as low-risk, compliant activities supported by standard laboratory biosafety protocols.

3.5.3. ICMJE and COPE compliance

The ICMJE January 2024 update [39], [40] and the COPE position statement on authorship and AI tools41,42 converge on five operational requirements that the YSDS Research Team endorses and recommends YSDS adopt as institutional policy:

AI cannot be listed as an author because it cannot take responsibility for the work, sign conflict-of-interest declarations, hold copyright, or be held legally accountable [41], [42].

AI use must be disclosed in both the cover letter and a dedicated subsection of the Methods (or an equivalent disclosure section), specifying which tool, which version, and for what purpose [39], [40].

 © Authors of the article / Authors of the article

Authors remain fully responsible for accuracy, integrity, and originality of all content including AI-generated text and must verify factual claims and citations [40], [41].

Confidentiality must be preserved: peer reviewers and editors must not upload submitted manuscripts to public AI tools without explicit author permission, because doing so may breach privileged communication [40].

AI cannot be cited as a source in the reference list, because LLM outputs are non-reproducible and not authoritative scientific records [40].

3.5.4. Data privacy for human genomic data

Compliance with Indonesia's Personal Data Protection Law (UU PDP No. 27/2022, fully enforced from October 2024) requires researchers to mirror European Union GDPR standards, including strict adherence to data-subject consent, purpose limitation, and breach notification protocols. Because human genomic datasets carry inherent risks of re-identification, this project adopts the following multi-tiered data privacy protocol:

Endpoint Restrictions: identifying genomic sequences and linked phenotypic data are strictly prohibited from being transmitted to public LLM endpoints.

Sovereign Infrastructure: sensitive analytical workflows will prioritize on-premises or sovereign-cloud infrastructures, such as self-hosted implementations of DeepSeek V3.1, Qwen 3.5, and Llama 4, or platforms ensuring regional data residency like Mistral Large 3.

Enterprise-Grade Compliance: any integration of commercial cloud LLMs will be restricted to enterprise tiers that explicitly guarantee zero data retention within their data-processing addenda, including OpenAI Enterprise, Google Vertex AI, and Anthropic hosted via Amazon Bedrock.

3.5.5. Bias against non-Western populations

Biomedical foundation models and LLMs exhibit significant geographic and demographic biases due to the overrepresentation of European-ancestry data, which reduces their generalizability in non-Western context. Empirical evidence shows that polygenic risk scores (PRS) generated by European-trained architectures suffer up to a 4.5-fold reduction in predictive accuracy when applied to non-European cohorts. To mitigate these systemic biases within the Indonesian genomics context, this project adopts a three-tiered validation framework:

Population-Specific Cross-Checking: all variant calls will be routinely cross-referenced against population-specific frequency databases, specifically leveraging the South Asian and East Asian subsets of gnomAD, supplemented by domestic Indonesian genomic registries where accessible.

Model Selection Caution: critical evaluation will be applied to single-cell foundation models trained predominantly on Caucasian-derived cell lines (e.g., HeLa, K562) to prevent skewed analytical baselines.

Ancestry Transparency: any reporting or manuscript preparation utilizing AI-generated phenotypic or genotypic predictions will explicitly disclose the ancestral composition of the training datasets to ensure proper contextual interpretation.

Limitation

(i) This is a structured narrative review, not a PRISMA systematic review, selection bias may favour widely publicised tools.

(ii) Benchmark figures change rapidly, the comparisons in Sections 3.1 to 3.3 reflect a snapshot as of May 2026.

(iii) Several tool claims (Undermind's ten-fold retrieval superiority, FutureHouse's "superhuman" PhD comparison) derive from vendor-published whitepapers or preprints not yet peer-reviewed at journal venue, and we have flagged these accordingly.

(iv) We have not independently re-run benchmarks.

(v) The ethical framework is grounded in WHO, ICMJE, and COPE guidance, institutional ethics review at YSDS, Indonesian Kemenkes, and any partnering institution remains required.

(vi) Indonesian-language LLM performance and Indonesian biomedical literature coverage in foundation models are under-studied and a priority for future work.

Conclusion

A researcher leading a Panin Group-funded biotechnology foundation in Jakarta does not need a single "best" AI model. They need a toolchain: GPT-5 or Gemini 2.5 Pro for general reasoning and scaffolding, FutureHouse Crow/Falcon or OpenScholar for citation-grounded literature work, Evo 2, AlphaFold 3, and ESM-3 for sequence and structure tasks, an open-weights model (DeepSeek V3.1 or Qwen 3.5) for data-sovereign deployment of identifying genomic data, and disciplined institutional policy aligned with ICMJE, COPE, WHO 2022, and Indonesia's Personal Data Protection Law. Claude Opus 4.x retains a role for writing polish and rebuttal-letter tone, with the caveat that its current over-restriction makes it unsuitable as a primary tool for diagnostic-biotechnology experimental design. The ethical imperative cuts both ways: scientific publishers must demand transparency, but model providers must calibrate restrictions to the realities of the diseases that disproportionately afflict LMIC populations were chief among them, in Indonesia, tuberculosis. The TB amplicon example is not a marginal case, it is a litmus test for whether AI safety policies actually serve global health.

 © Authors of the article / Authors of the article

Благодарности

Автор выражает благодарность исследовательской группе YSDS за проведение внутренней экспертизы, а также группе университетов IGNITE 21 за непрерывный диалог по вопросам ответственного использования искусственного интеллекта в биомедицинских исследованиях в Индонезии.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Acknowledgement

The author thanks the YSDS Research Team for internal review, and the IGNITE 21-university network for ongoing dialogue on responsible AI use in Indonesian biomedical research.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Foundation Models for Biomedical Research // Watershed Bio. — 2024. — URL: https://watershed.bio/resources/foundation-models-for-biomedical-research (accessed: 16.07.2026).

2. Hasanah N. Variation of tuberculosis prevalence across diagnostic approaches and geographical areas of Indonesia / N. Hasanah, A. Surya, B.D. Nugraha [et al.] // PLOS ONE. — 2021. — Vol. 16, № 10. — P. e0258809. — DOI: 10.1371/journal.pone.0258809.

3. Burhan E. Tuberculosis in Indonesia: challenges and future directions / E. Burhan, A.Y. Soeroto, F. Isbaniah [et al.] // The Lancet Respiratory Medicine. — 2025. — DOI: 10.1016/S2213-2600(25)00168-7.

4. Introducing GPT-5 // OpenAI. — 2025. — URL: https://openai.com/index/introducing-gpt-5/ (accessed: 16.07.2026).

5. Introducing GPT-5.2 // OpenAI. — 2025. — URL: https://openai.com/index/introducing-gpt-5-2/ (accessed: 16.07.2026).

6. GPT-5 Benchmarks // Vellum AI. — 2025. — URL: https://www.vellum.ai/blog/gpt-5-benchmarks (accessed: 16.07.2026).

7. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next-Generation Agentic Capabilities // Gemini Team, Google. — arXiv, 2025. — arXiv:2507.06261.

8. Brixi G. Genome modeling and design across all domains of life with Evo 2 / G. Brixi, M.G. Durrant, J. Ku [et al.] // Nature. — 2026. — URL: https://www.researchgate.net/publication/389218623_Genome_modeling_and_design_across_all_domains_of_life_with_Evo_2 (accessed: 16.07.2026).

9. Abramson J. Accurate structure prediction of biomolecular interactions with AlphaFold 3 / J. Abramson, J. Adler, J. Dunger [et al.] // Nature. — 2024. — Vol. 630, № 8016. — P. 493–500. — DOI: 10.1038/s41586-024-07487-w.

10. Hayes T. Simulating 500 million years of evolution with a language model (ESM3) / T. Hayes, R. Rao, H. Akin [et al.] // Science. — 2025. — Vol. 387. — Art. eads0018. — DOI: 10.1126/science.ads0018.

11. Dalla-Torre H. Nucleotide Transformer: building and evaluating robust foundation models for human genomics / H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla [et al.] // Nature Methods. — 2024. — DOI: 10.1038/s41592-024-02523-z.

12. Asai A. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs / A. Asai, J. He, R. Shao [et al.]. — arXiv, 2024. — arXiv:2411.14199.

13. Skarlinski M.D. Language agents achieve superhuman synthesis of scientific knowledge (PaperQA2) / M.D. Skarlinski, S. Cox, J.M. Laurent [et al.]. — arXiv, 2024. — arXiv:2409.13740.

14. The Register. Claude Opus 4.7 has turned into an overzealous query cop. — 2026. — URL: https://www.theregister.com/software/2026/04/23/claude-opus-47-has-turned-into-an-overzealous-query-cop/ (accessed: 16.07.2026).

15. Anthropic. Claude Opus 4.5 System Card. — 2025. — URL: https://thezvi.substack.com/p/claude-opus-45-model-card-alignment (accessed: 16.07.2026).

16. Accelerating scientific discovery with AI (FutureHouse) // MIT News. — 2025. — URL: https://news.mit.edu/2025/futurehouse-accelerates-scientific-discovery-with-ai-0630 (accessed: 16.07.2026).

17. Hartke T. Benchmarking the Undermind Search Assistant: whitepaper / T. Hartke, J. Ramette. — Undermind Inc., 2024. — URL: https://www.undermind.ai/whitepaper.pdf (accessed: 16.07.2026).

18. Diaz H.J.V. Assessing the Effectiveness of AI Tools (Elicit, SciSpace, and Consensus) in Literature Review and Research / H.J.V. Diaz [et al.] // Canadian Journal of Information and Library Science. — 2025. — URL: https://ojs.lib.uwo.ca/index.php/cjils/article/view/23075 (accessed: 16.07.2026).

19. ResearchRabbit: Zotero integration documentation. — 2024. — URL: https://cshl.libguides.com/research_rabbit/zotero-integration (accessed: 16.07.2026).

20. Roldán-García M.D.M. Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are fully accurate / M.D.M. Roldán-García, J. García-Nieto [et al.]. — arXiv, 2025. — arXiv:2505.18059.

21. Perplexity Made Up Information Fake: 2026 Fix Guide // AI Q&A Hub. — 2026. — URL: https://www.aiqnahub.com/perplexity-made-up-information-fake/ (accessed: 16.07.2026).

 © Authors of the article / Authors of the article

22. Shao Y. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models (STORM) / Y. Shao, Y. Jiang, T.A. Kanell [et al.] // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). — 2024. — P. 6252–6278. — DOI: 10.18653/v1/2024.naacl-long.347.

23. Tang Z. Benchmarking DNA foundation models for genomic and genetic tasks / Z. Tang [et al.] // Nature Communications. — 2025. — URL: https://www.nature.com/articles/s41467-025-65823-8 (accessed: 16.07.2026).

24. Theodoris C.V. Transfer learning enables predictions in network biology (Geneformer) / C.V. Theodoris, L. Xiao, A. Chopra [et al.] // Nature. — 2023. — Vol. 618. — P. 616–624.

25. Cui H. scGPT: Toward Building a Foundation Model for Single-Cell Multi-omics using Generative AI / H. Cui, C. Wang, H. Maan [et al.] // Nature Methods. — 2024. — URL: https://www.biorxiv.org/content/10.1101/2023.04.30.538439 (accessed: 16.07.2026).

26. Kedzierska K.Z. Zero-shot evaluation reveals limitations of single-cell foundation models / K.Z. Kedzierska, L. Crawford, A.P. Amini [et al.] // Genome Biology. — 2025. — PMCID: PMC12007350.

27. Chen Y. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT / Y. Chen, J. Zou // bioRxiv. — 2024. — DOI: 10.1101/2023.10.16.562533.

28. Luo R. BioGPT: generative pre-trained transformer for biomedical text generation and mining / R. Luo, L. Sun, Y. Xia [et al.] // Briefings in Bioinformatics. — 2022. — Vol. 23, № 6. — Art. bbac409. — DOI: 10.1093/bib/bbac409.

29. U.S. National Institute of Standards and Technology, Center for AI Standards and Innovation (NIST CAISI). CAISI Evaluation of DeepSeek V4 Pro. — 2026. — URL: https://www.nist.gov/news-events/news/2026/05/caisi-evaluation-deepseek-v4-pro (accessed: 16.07.2026).

30. Noever D.A. Forbidden Science: Dual-Use AI Challenge Benchmark and Scientific Refusal Tests / D.A. Noever, F. McKee. — arXiv, 2025. — arXiv:2502.06867.

31. Gottweis J. Towards an AI co-scientist / J. Gottweis [et al.]. — Google DeepMind / Nature, 2025. — URL: https://deepmind.google/blog/co-scientist-a-multi-agent-ai-partner-to-accelerate-research/ (accessed: 16.07.2026).

32. Accelerating the Genesis Mission with Gemini for Government // Google Cloud Blog. — 2025. — URL: https://cloud.google.com/blog/topics/public-sector/how-google-public-sector-and-google-deepmind-can-power-the-genesis-mission-and-a-new-era-of-scientific-discovery (accessed: 16.07.2026).

33. Grok 4 Technical Brief // Artificial Analysis benchmark. — 2026. — URL: https://artificialanalysis.ai/models/grok-4; https://x.ai/news/grok-4 (accessed: 16.07.2026).

34. Burhan E. Tuberculosis in Indonesia: challenges and future directions / E. Burhan, A.Y. Soeroto, F. Isbaniah [et al.] // The Lancet Respiratory Medicine. — 2025. — DOI: 10.1016/S2213-2600(25)00168-7.

35. Anggraini P. Standard E TB-Feron ELISA and Standard F TB-Feron FIA positivity rates and agreement with QuantiFERON-TB Gold Plus among TB high-risk population in Bandung, Indonesia / P. Anggraini [et al.]. — 2025. — PMCID: PMC11980364.

36. Pradipta A. Drug resistance of Mycobacterium tuberculosis in West Java, Indonesia / A. Pradipta, D. Virya Chen, Ih. Fauzan [и др.] // Current Research in Microbial Sciences. — 2025. — Vol. 8. — Art. 100219.

37. Global Guidance Framework for the Responsible Use of the Life Sciences: Mitigating Biorisks and Governing Dual-Use Research // World Health Organization. — Geneva : WHO, 2022. — 114 p.

38. Technical Advisory Group on the Responsible Use of the Life Sciences and Dual-Use Research (TAG-RULS DUR). — URL: https://www.who.int/groups/technical-advisory-group-on-the-responsible-use-of-the-life-sciences-and-dual-use-research-(tag-ruls-dur) (accessed: 16.07.2026).

39. Up-Dated ICMJE Recommendations // International Committee of Medical Journal Editors (ICMJE). — 2024. — URL: http://www.icmje.org/news-and-editorials/updated_recommendations_jan2024.html (accessed: 16.07.2026).

40. Recommendations: Use of Artificial Intelligence in Publishing // International Committee of Medical Journal Editors (ICMJE). — 2024. — URL: https://www.icmje.org/recommendations/browse/artificial-intelligence/ (accessed: 16.07.2026).

41. Authorship and AI tools : Position Statement // COPE Council (Committee on Publication Ethics). — 2023. — URL: https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools (accessed: 16.07.2026).

42. Levy N. Responsibility is not required for authorship / N. Levy // Journal of Medical Ethics. — 2024. — DOI: 10.1136/jme-2024-109912.

43. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research / National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. — Bethesda, MD, 1979. — 20 p.

44. Ethical Principles for Medical Research Involving Human Participants / World Medical Association (WMA). — Helsinki, 2024. — URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ (accessed: 16.07.2026).

45. Martin A.R. Clinical use of current polygenic risk scores may exacerbate health disparities / A.R. Martin, M. Kanai, Y. Kamatani [et al.] // Nature Genetics. — 2019. — Vol. 51. — P. 584–591. — DOI: 10.1038/s41588-019-0379-x.

46. Grand challenge in biosafety and biosecurity // Frontiers in Bioengineering and Biotechnology. — 2024. — Vol. 12. — Art. 1538723. — DOI: 10.3389/fbioe.2024.1538723.

47. National Strategy for Tuberculosis Control 2020–2024. — Jakarta : Kemenkes RI, 2020. — xvi, 218 p.

48. Personal Data Protection Law. — Jakarta : Kementerian Sekretariat Negara RI, 2022. — 70 p.

49. Susilawati T.N. A recent update of the diagnostic methods for tuberculosis and their applicability in Indonesia: a narrative review / T.N. Susilawati, R. Larasati // Medical Journal of Indonesia. — 2019.

50. Yadav S. Evolution of ICMJE recommendations: a beacon of ethical guidance / S. Yadav // Cureus. — 2024. — Vol. 16, № 5. — Art. e59792.

51. Hu W. Comprehensive benchmarking of AlphaFold3 for predicting biomacromolecules and their interactions / W. Hu, Y. Wang [et al.] // Briefings in Bioinformatics. — 2025. — Vol. 26, № 6. — Art. bbaf616. — DOI: 10.1093/bib/bbaf616.

 © Authors of the article / Authors of the article

52. Wright C.Y. The African Journal Partnership Program's guidance on the use of AI in scholarly publishing / C.Y. Wright, M. Lartey, K. Khomsi [et al.] // Ghana Medical Journal. — 2024. — Vol. 58, № 1. — DOI: 10.4314/gmj.v58i1.1.

53. A Standardized Framework for Integrating and Benchmarking Single-Cell Foundation Models. — bioRxiv, 2024. — DOI: 10.1101/2024.11.22.624786.

54. Hitawala A. AlphaFold3, a secret sauce for predicting mutational effects on protein–protein interactions / A. Hitawala, J.J. Gray. — bioRxiv, 2024. — DOI: 10.1101/2024.05.25.595871.

55. Grok 4 and Grok 4.3 — Intelligence, Performance & Price Analysis / Artificial Analysis. — 2025–2026. — URL: https://artificialanalysis.ai/models/grok-4 ; https://artificialanalysis.ai/models/comparisons/grok-4-3-vs-grok-4 (accessed: 16.07.2026).

56. Launching the FutureHouse Platform. Superintelligent AI Agents for Scientific Discovery // FutureHouse. — 2025. — URL: https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents (accessed: 16.07.2026).

57. Interpreting Evo 2: Arc Institute's Next-Generation Genomic Foundation Model // Goodfire AI, Arc Institute. — 2026. — URL: goodfire.ai (accessed: 16.07.2026).

58. The ICMJE recommendations on AI: Advice for authors and peer reviewers // Editage Insights. — 2024. — URL: https://www.editage.com/insights/the-icmje-recommendations-on-ai-advice-for-authors-and-peer-reviewers (accessed: 16.07.2026).

59. ICMJE recommendations update 2024: what's new and what's next? // The Publication Plan. — 2024. — URL: thepublicationplan.com (accessed: 16.07.2026).

60. AlphaFold 3 vs AlphaFold 2: what changed for drug discovery / Mindwalk AI. — 2024–2025. — URL: mindwalk.ai (accessed: 16.07.2026).

61. Grok 4 Launches With Benchmark Records and Idiosyncratic Behavior // DeepLearning.AI. — 2025. — URL: deeplearning.ai (accessed: 16.07.2026).