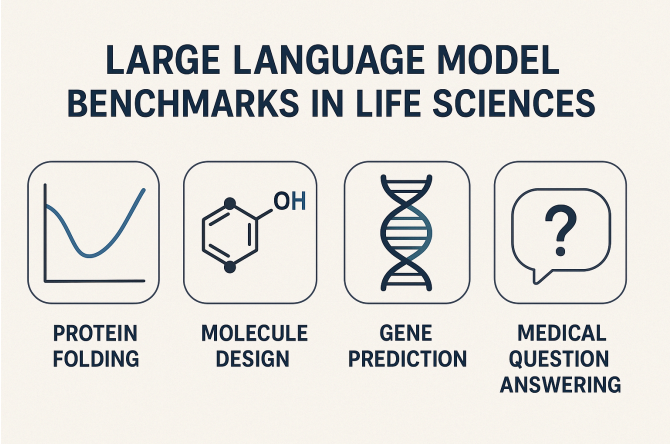
LLM Benchmarks in Life Sciences: Comprehensive Overview

By IntuitionLabs • 5/5/2025 • 40 min read

llm-benchmarks life-sciences pharmaceutical biomedical-nlp bioinformatics drug-discovery genomics ai-evaluation model-comparison nlp machine-learning



Large Language Model Benchmarks in Life Sciences: A Comprehensive Overview

Related reading: ChatGPT Adoption in the Life Sciences Industry | Accelerating Drug Development with AI in Pharma | Data Science in Life Sciences | Veeva Clinical Trial Management System: Software for Clinical Research | Performance of Retrieval-Augmented Generation (RAG) on Pharmaceutical Documents

Introduction

The rapid progress in large language models (LLMs) has spurred the creation of benchmarks to evaluate their capabilities in specialized domains like life sciences. For IT professionals in the pharmaceutical and biotech industry, understanding these benchmarks is crucial. Benchmarks provide standardized tasks and datasets to measure how well LLMs perform on biomedical literature mining, clinical question-answering, drug discovery, genomics analysis, and more. By comparing models on common metrics, benchmarks help identify strengths, weaknesses, and readiness for real-world applications. This report surveys all major LLM benchmarks used in life sciences – spanning biomedical, pharmaceutical, and genomics domains – with an emphasis on developments from 2020 to 2025. We cover general natural language processing (NLP) and question-answering benchmarks (e.g. BioASQ, PubMedQA, MedQA), as well as task-specific evaluations in drug discovery (molecule generation, property prediction) and genomics (gene and protein understanding). For each benchmark, we outline its scope, discuss its importance to industry use cases, and highlight model performance with relevant metrics. Recent trends show that while domain-specific models fine-tuned on biomedical data still excel in many information extraction tasks, the newest general-purpose LLMs (like GPT-4) have achieved breakthroughs in complex reasoning tasks such as medical question-answering (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). The tables and sections below organize the benchmarks by category and summarize key characteristics and state-of-the-art results, providing a clear reference for professionals seeking to leverage LLMs in life science applications.

Biomedical Language Understanding Benchmarks

One foundational effort to benchmark LLMs in the biomedical domain is the creation of *broad-coverage evaluation suites* analogous to general NLP benchmarks like GLUE. Historically, biomedical NLP researchers participated in many *shared tasks* (BioCreative, BioNLP, SemEval, etc.), each focusing on specific challenges like gene name recognition or protein interaction extraction (BLURB Leaderboard). However, the introduction of modern transformer models led to the need for integrated benchmarks to evaluate general-purpose language understanding in biomedicine. Two influential benchmark suites emerged to fill this role:

- **BLUE Benchmark (2019)** – The *Biomedical Language Understanding Evaluation* (BLUE) benchmark was introduced by researchers at NCBI as a domain-specific analogue of GLUE ([1906.05474] Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets). BLUE encompasses **five task types** with **ten datasets** covering both biomedical research text (e.g. PubMed abstracts) and clinical text (e.g. electronic health records) (BLUE Dataset-Papers With Code) (ClinicalBERT and BlueBERT. Adapting BERT for Biomedical and...-by Eleventh Hour Enthusiast-Medium). The tasks include **sentence similarity, named entity recognition (NER), relation extraction, document classification, and natural language inference (NLI)** (ClinicalBERT and BlueBERT. Adapting BERT for Biomedical and...-by Eleventh Hour Enthusiast-Medium). By evaluating models on this diverse set (spanning short text similarity to inference on clinical statements), BLUE provides a standardized way to compare model performance across biomedical NLP tasks. Early domain-specific models fine-tuned on BLUE, such as *BlueBERT* (BERT base pre-trained on PubMed + clinical notes), achieved strong results and validated the benefit of domain-specific pretraining (ClinicalBERT and BlueBERT. Adapting BERT for Biomedical and...-by Eleventh Hour Enthusiast-Medium). For example, BlueBERT obtained leading scores on multiple BLUE tasks, demonstrating its robustness in biomedical and clinical text processing (ClinicalBERT and BlueBERT. Adapting BERT for Biomedical and...-by Eleventh Hour Enthusiast-Medium). The BLUE benchmark was a **historically significant** step that highlighted the limitations of general models on biomedical tasks and spurred development of specialized models.

- **BLURB Benchmark (2020)** – The *Biomedical Language Understanding and Reasoning Benchmark* (BLURB) built on the BLUE initiative and expanded it. BLURB (released by Microsoft Research) aggregates **13 datasets across 6 task categories** (BLURB Leaderboard). It includes classic biomedical text mining tasks: five NER datasets (recognizing chemicals, diseases, genes, etc.), three relation extraction datasets (e.g. chemical-protein and drug-drug interactions), document classification (e.g. classifying abstracts by topics such as the Hallmarks of Cancer), sentence similarity (BIOSSES), and question answering (BioASQ and PubMedQA) (BLURB Leaderboard) (BLURB Leaderboard). Table 1 summarizes the key datasets in BLURB. The benchmark reports a **macro-average score** across all tasks as the main metric, to ensure no single task dominates the evaluation (BLURB Leaderboard). BLURB established a public leaderboard that has driven progress in biomedical NLP by encouraging researchers to develop models that perform well universally. For instance, the BioALBERT model (an ALBERT-based domain model) achieved a new state-of-the-art on 5 out of 6 BLURB task types, outperforming previous models in NER, relation extraction, sentence similarity, document classification, and QA (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC). Specifically, BioALBERT (large, PubMed-trained) improved the BLURB score for NER by **+11.1%**, for QA by **+2.8%**, and set SOTA on 17 of the 20 dataset evaluations (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC) (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC). Such improvements underscore how benchmark-driven development has significantly boosted accuracy on biomedical text tasks. For industry use, these language understanding benchmarks are important because tasks like **entity recognition and relation extraction** underpin applications ranging from literature curation to building knowledge graphs of diseases, genes, and drugs. High F1-scores on BLURB's NER and interaction extraction datasets (often exceeding 85% for top models (BLURB Leaderboard)) mean that modern models can reliably automate the extraction of structured biomedical knowledge – a valuable capability for pharmaceutical companies dealing with information overload in publications.

Table 1. Biomedical NLP Benchmarks (BLUE and BLURB) – Tasks, Examples, and Top Model Performance (2020–2025)

Task Category	Example Dataset	Task Description	Metric	State-of-the-Art Performance (approx.)
Named Entity Recognition (NER)	NCBI-Disease (BLURB) (BLURB Leaderboard) (BLURB Leaderboard); BC5-Chemicals	Identify biomedical entities (genes, diseases, chemicals) in text.	F1 (entity-level)	**BioALBERT** (large, 2022): ~85–90% F1 on biomedical NER (BLURB Leaderboard) (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC);

Task Category	Example Dataset	Task Description	Metric	State-of-the-Art Performance (approx.)
				surpasses general BERT by 5–10%.
Relation Extraction	ChemProt (chemical-protein) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications); DDI (drug–drug interact.)	Detect relations between biomedical entities in text (e.g., drug interactions, protein binding).	F1 (micro)	**BioBERT** family (2020): ~73% F1 on ChemProt; **BioALBERT** (2022) slightly higher (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC). GPT-4 (2023) zero-shot lags (~65% F1) but improves with fine-tuning (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications).
Document Classification	HoC – Hallmarks of Cancer (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications); LitCovid (COVID topics)	Assign labels or topics to a scientific abstract or clinical note (multi-label possible).	F1 (micro) or accuracy	**BioBERT/PubMedBERT** (2020): ~70% micro-F1 on HoC. **LLMs** (GPT-3.5, GPT-4) in zero-shot ~62–67% (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), approaching fine-tuned model performance.
Sentence Similarity	BIOSSES (sentence similarity)	Determine semantic similarity between sentence pairs	Pearson/Spearman correlation	**BioALBERT** (2022): ~0.90 correlation (Benchmarking for biomedical natural language processing

Task Category	Example Dataset	Task Description	Metric	State-of-the-Art Performance (approx.)
		(e.g., biomedical facts).		tasks with a domain specific ALBERT - PMC) (improved +1.0% over prior SOTA). Domain pretraining yields best results.
Natural Language Inference	MedNLI (clinical NLI)	Infer logical relation between sentences (e.g., hypothesis supported by premise in patient note?).	Accuracy	**ClinicalBERT** fine-tuned (2019): ~82% accuracy; Newer LLMs ~80–85% in few-shot. (MedNLI is part of BLUE; top BlueBERT model excelled (ClinicalBERT and BlueBERT. Adapting BERT for Biomedical and...-by Eleventh Hour Enthusiast-Medium).)
QA (Biomedical Literature)	**BioASQ** (facts from PubMed) (BLURB Leaderboard); **PubMedQA** (study Q&A) (BLURB Leaderboard)	Answer biomedical questions either via information retrieval (BioASQ) or reading comprehension (PubMedQA).	Accuracy (exact answer) or F1	**BioBERT** (2019) fine-tuned: ~78% accuracy on PubMedQA; **PMC-LLaMA** 13B (2024) fine-tuned: ~77.9% (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). *GPT-4* zero-shot: ~75% on PubMedQA (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) – near SOTA without training. BioASQ (factoid QA) top systems reach

Task Category	Example Dataset	Task Description	Metric	State-of-the-Art Performance (approx.)
				80–90% precision (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) using ensembles and IR.

Table 1: Core biomedical NLP benchmarks from BLUE/BLURB and related efforts, illustrating the breadth of tasks. Domain-specific models (e.g. BioBERT, BioALBERT) have achieved strong results by 2022, often outperforming general LLMs on information extraction tasks (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC). However, general LLMs like GPT-4 are competitive on knowledge-intensive QA tasks even without domain fine-tuning (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). These benchmarks cover abilities such as recognizing terminology, extracting relationships, classifying documents, and answering research questions – all vital for industry applications like automated literature review, clinical data mining, and knowledge base construction.

In industry settings, the above benchmarks translate to practical use cases. **Named entity recognition** and **relation extraction** are directly useful for building pharmacovigilance systems (e.g., extracting adverse drug events from case reports) and research discovery platforms (e.g., linking genes to diseases from publications). High-performing models on ChemProt or DDI (drug-drug interaction) can automate the curation of interaction databases from the literature (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). **Document classification** tasks like HoC or LitCovid were crucial during the COVID-19 pandemic to organize the influx of papers by topics (treatments, mechanisms, etc.), and a model that performs well on LitCovid classification can help pharma companies quickly filter relevant studies (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). **Inference and similarity tasks** ensure that models can reason about textual information – for example, determining if a given clinical finding supports a hypothesis or matching trial criteria to patient descriptions. This underpins decision support tools that must understand nuanced language logic in guidelines or trial protocols. Finally, **biomedical QA** benchmarks (detailed next) are directly tied to building question-answering systems for researchers and clinicians, an area of great interest for improving information access in healthcare.

Biomedical Question-Answering Benchmarks

Biomedical question-answering (QA) is a critical application of LLMs, as it enables users to query vast biomedical knowledge bases (like PubMed) in natural language. Several benchmarks have been established to evaluate how well models can answer questions in the life sciences domain, ranging from research factoids to medical exam queries. We highlight the major QA benchmarks:

- **BioASQ (2013–present)** – BioASQ is an annual challenge and benchmark for biomedical semantic indexing and question answering, sponsored by the National Library of Medicine. In its QA tasks (Phase B), systems must answer questions posted by biomedical experts, which can be **factoid questions**, **list questions**, or **yes/no questions**, often with supporting evidence from PubMed articles. This benchmark tests a model's ability to *retrieve relevant information* and *provide precise answers*. Metrics include accuracy for yes/no, and precision/recall/F1 for factoids and lists. BioASQ has historically driven progress in biomedical QA: early systems used information retrieval + NLP pipelines, but with LLMs, a shift toward end-to-end approaches is occurring. State-of-the-art systems in recent BioASQ editions achieve high performance (e.g., >80% accuracy on yes/no questions and F1 scores ~0.5–0.6 for factoids) by leveraging ensembles of biomedical BERT models and reading comprehension modules (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). The significance for industry is clear – a QA model excelling at BioASQ can underpin tools for scientists to ask research questions (e.g., "What are known biomarkers for Alzheimer's?") and get concise answers with references, dramatically speeding up literature review.

- **PubMedQA (2019)** – PubMedQA is a dataset of research article-derived questions, each with a short answer and a supporting abstract from PubMed (BLURB Leaderboard). Questions are often phrased as yes/no or require identifying a specific finding from the abstract. The task is essentially *machine reading comprehension* in the biomedical domain. For example, a question might ask, "Does drug X improve survival in condition Y according to the study?" and the model must read the abstract to answer "yes", "no", or "maybe". The benchmark provides ~1,000 question-answer pairs, and models are evaluated by accuracy. **Fine-tuned biomedical models** like BioBERT and PubMedBERT were among the first to perform well, reaching ~65–70% accuracy by 2020. More recently, larger models have significantly improved results – e.g., a fine-tuned *PMC-LLaMA 13B* model (an open LLaMA tuned on medical QAs) achieved **77.9% accuracy** on PubMedQA (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), nearly matching the performance of a model that was fine-tuned on multiple QA datasets combined. Notably, GPT-4 in a zero-shot setting (without fine-tuning) can reach around **75% accuracy** on PubMedQA (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), demonstrating the strong out-of-the-box knowledge of closed-source LLMs. This is promising for industry use: without needing task-specific training, a model like GPT-4 can already answer questions about clinical studies nearly as well as specialized models. In pharma, such capability means quicker answers to questions about evidence in literature (e.g., finding if a study supports a certain hypothesis).

- **MedQA (USMLE)** – One of the most challenging benchmarks is MedQA, a dataset derived from the *United States Medical Licensing Exam (USMLE)* questions (MedQA Dataset - Papers With Code). This benchmark contains **multiple-choice questions** that test medical knowledge and clinical reasoning, similar to what medical students must answer. Each question includes a patient scenario and four or more answer options, requiring application of medical facts and reasoning to choose the correct one. MedQA is a test of an LLM's ability to perform *medical reasoning and decision-making*. Traditionally, models struggled on this benchmark – for years, accuracy remained near 40%, since random guessing is 25%. However, recent LLMs have made dramatic gains. Fine-tuned transformers (like Google's Med-PaLM, a PaLM model fine-tuned on medical Q&A) reached ~67% accuracy (close to passing) in 2022. Then, **GPT-4** essentially solved much of the task: GPT-4 in zero-shot scored about **71.6% accuracy** on the MedQA dataset (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), and in some reports GPT-4 averaged **~86%** on USMLE-style questions overall (Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments-Scientific Reports) – surpassing the passing threshold by over 20 points. In comparison, the best prior models (domain-specific) were around 41–42% (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). Even GPT-3.5 (ChatGPT) was able to exceed prior state-of-the-art with ~50% accuracy zero-shot (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). These results highlight that complex multi-step reasoning, which was once thought to require explicit knowledge graphs or logic, can now be handled by large-scale LLMs with emergent capabilities. For the pharmaceutical industry, a model that performs well on MedQA is attractive for decision support tools – for example, assisting in medical education, or even suggesting diagnoses in complex cases (with appropriate oversight). It shows the potential of LLMs to **reason about clinical scenarios**, not just parrot facts. However, caution is needed: passing an exam is different from clinical practice, but it's a valuable benchmark indicating high-level understanding.

- **MedMCQA and Other QA Benchmarks** – In addition to MedQA, there are other QA datasets like *MedMCQA*, a large collection of ~20,000 medical multiple-choice questions released in 2022. It covers medical entrance exam questions from India and has both four-option and higher-order reasoning questions. Models like BioGPT and PaLM have been evaluated on MedMCQA, with accuracies in the 50–60% range reported in literature. Another relevant benchmark is the medical portion of the **Massive Multitask Language Understanding (MMLU)** test – a general benchmark where one category is Medicine. GPT-4's performance on the medical subportion of MMLU is around 81– Ninety percent (detailed in OpenAI's report), whereas prior models achieved roughly 50–60%. These benchmarks reinforce the pattern seen in MedQA: larger models with more knowledge tend to excel in multi-turn reasoning QA.

Why these QA benchmarks matter: For industry, **open-domain biomedical QA** (BioASQ-style) is directly applicable to creating **literature search assistants** for scientists or **clinical Q&A systems** for healthcare providers. The ability to accurately answer questions like "What evidence supports using Drug A for Disease B?" can save enormous time. Meanwhile, the **exam-style QA** benchmarks (MedQA, MedMCQA) test deeper reasoning and knowledge integration. Success on those implies a model can potentially assist in diagnostic reasoning or medical training. We are already seeing early applications: for instance, an LLM fine-tuned to pass USMLE is being evaluated as a virtual medical tutor and as a triage assistant. High benchmark scores give confidence in the model's reliability. It's worth noting that the best results often combine the model's reasoning with retrieval of trusted information. Research from 2024 shows that even open-source LLMs can approach GPT-4's QA performance when augmented with relevant literature retrieval (a technique known as retrieval-augmented generation) (Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease - PMC) (Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease - PMC). This suggests a path for pharma IT teams: using internal document repositories in tandem with LLMs to answer proprietary questions (like those about internal study data) with the same prowess seen in public benchmarks.

Drug Discovery and Molecular Benchmarks

LLMs in the pharmaceutical domain are not limited to text – they are increasingly applied to chemical and biological sequence data by treating molecules or proteins as a "language." Benchmarks in this area evaluate models on tasks crucial to drug discovery, such as predicting molecular properties, generating novel compounds, or modeling protein interactions. Both open-source academic benchmarks and internal pharma evaluations exist. Here we cover prominent **open benchmarks for cheminformatics and drug discovery**, highlighting how language-modeling approaches are assessed:

- **MoleculeNet (2018)** – MoleculeNet is a widely used benchmark suite for AI in chemistry, introduced as part of the DeepChem project. It comprises a collection of datasets for molecular property prediction across various categories: physical chemistry (e.g., QM9 quantum properties), biophysics (e.g., solubility), physiology (e.g., blood-brain barrier penetration), and chemistry tasks like toxicity (e.g., Tox21) (Machine Learning Datasets and Tasks for Drug Discovery … - arXiv). Tasks can be regression (predict a numeric property) or classification (e.g., active/inactive against a target). Although MoleculeNet predates "LLMs" per se, it has become a standard to evaluate any new model that generates molecular embeddings or does transfer learning on chemical data. Many graph neural networks and transformer-based models have been benchmarked here. For instance, the message-passing neural networks achieved strong AUC scores (~0.85–0.90) on toxicity tasks, and recent transformer models treating SMILES strings (text representations of molecules) have started to compete. In industry, performance on MoleculeNet tasks correlates to how well a model can predict drug properties (ADMET) early in the pipeline – a high R^2 on clearance or toxicity prediction means the model could help screen out poor drug candidates. Modern benchmarks like the **Therapeutics Data Commons (TDC)** (2021) build upon MoleculeNet, providing a platform and leaderboard for these tasks (Machine Learning Datasets and Tasks for Drug Discovery … - arXiv). TDC standardizes evaluation of over 50 datasets including MoleculeNet's, and tracks metrics like ROC-AUC, RMSE, etc., for models in areas like drug–target interaction prediction, pharmacokinetics, and combination therapy outcome prediction. By 2025, *transformer-based chemical models* (such as ChemBERTa and MolT5) report competitive results on TDC benchmarks, often within a few percentage points of specialized graph models on classification tasks. This indicates LLM-style architectures are viable in cheminformatics, and benchmarks ensure they meet domain requirements for accuracy.

- **GuacaMol and MOSES (2018–2019)** – These two benchmarks focus on **de novo molecule generation**, a task where models propose novel chemical structures with desirable properties. GuacaMol (A review of large language models and autonomous agents in ...) defines a set of generative tasks and metrics to quantify how well algorithms explore chemical space (including metrics for novelty, diversity, drug-likeness, and goal-directed generation such as optimizing a molecular property). MOSES is a similar benchmarking platform providing a standardized dataset of compounds and evaluation metrics for model-generated molecules (e.g., validity of generated structures, uniqueness, Fréchet ChemNet Distance for distribution similarity). Traditionally, generative models like GANs or variational autoencoders were tested with these benchmarks. Now, LLMs that treat SMILES as language are also evaluated. For example, a GPT-2 model trained on SMILES can generate novel compounds; GuacaMol would measure that, say, X% of its outputs are valid molecules, Y% are unique, and how many meet certain property criteria. Top models in literature achieve >95% validity and high novelty in these benchmarks, and can optimize simple properties (like logP or molecular weight) to targets (A review of large language models and autonomous agents in ...). For pharmaceutical AI, these metrics are proxies for the **creativity and reliability** of AI-driven molecule design. A high GuacaMol score means a model could accelerate medicinal chemistry by proposing molecules humans might not think of, while satisfying drug-like constraints. However, these benchmarks do not guarantee the generated compounds are synthesizable or truly efficacious – they are a first filter. Thus, industry labs often use them in conjunction with more advanced filters.

- **TOMG-Bench (2024)** – A recent development tailored specifically to LLMs in chemistry is **TOMG-Bench (Text-based Open Molecule Generation Benchmark)** (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). This benchmark was introduced as the first to evaluate LLMs on open-ended molecule design via textual instructions. It encompasses **three tasks** that mimic medicinal chemist requests: *molecule editing* (modify a given molecule to improve some aspect), *molecule optimization* (optimize a molecule for a property like potency or reduce toxicity), and *custom molecule generation* (generate a molecule meeting a complex text prompt, e.g., "a molecule similar to aspirin that binds to protein X") (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). Each task has defined subtasks and on the order of 5,000 test prompts, making it a robust evaluation. Importantly, TOMG-Bench includes an automated evaluation system to check the quality and validity of generated molecules (using chemical analysis libraries). In a comprehensive evaluation of 25 LLMs, it was found that **most general LLMs struggle with precise molecule generation** – many outputs were invalid as molecules or failed the requirements (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). For example, GPT-3.5 scored significantly lower than a specialized fine-tuned model (OpenMolGPT) on these tasks. With domain-specific instruction tuning (the *OpenMolIns* dataset), a fine-tuned 8B LLaMA-based model (called Llama3.1-8B in the paper) outperformed even GPT-3.5, surpassing GPT-3.5's score by **46.5%** on TOMG-Bench (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation) (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). This demonstrates that with appropriate data, smaller open models can beat large general models on chemistry tasks. For industry, TOMG-Bench is a promising yardstick to measure an AI assistant's capability to **help design new drugs via text prompts**. A model that scores well could take high-level instructions from chemists and propose viable compounds, streamlining the ideation phase in drug discovery. As of 2025, this is an area of active research, with companies experimenting with connecting LLMs to chemistry engines. The benchmark ensures any claims of a "ChatGPT for chemists" are backed by quantitative performance on realistic tasks.

In Table 2, we summarize several key benchmarks related to drug discovery along with typical metrics and current model performance levels:

Table 2. Benchmarks for Drug Discovery and Genomics – Key Tasks and Model Performance

Benchmark / Task	Domain	Description & Use Case	Metric	Notable Results (2020–2025)
Therapeutics Data Commons (TDC) (Machine Learning Datasets and Tasks for Drug Discovery ... - arXiv)	Drug discovery (multi-task)	Collection of 50+ datasets (ADMET prediction, drug-target binding affinity, combination therapy outcome, etc.), unified platform with	Varied (ROC-AUC, PR-AUC, RMSE, etc. per task)	*GraphConv Models* (2018): baseline ROC-AUC ~0.85 on Tox21; *ChemBERTa (2021)*: similar or slightly improved on property prediction.

Benchmark / Task	Domain	Description & Use Case	Metric	Notable Results (2020–2025)
		leaderboard. Used to evaluate models for various stages of drug development.		**GraphNetworks vs Transformers:** Results show competitive performance (within ~2-3% AUC) for transformers on many tasks by 2023 ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks), though experts models still lead in some.
GuacaMol (2018) (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule ...)	Molecule generation	Goal-directed generation of novel molecules with desired properties (several challenge tasks). Used to benchmark generative models' ability to create drug-like, novel compounds.	Composite scoring (validity, novelty, uniqueness, goal achievement)	*JT-VAE* (2018): Validity > 95%, Novelty ~80%; *GraphGA (2019)*: excels at goal-directed tasks (e.g., scoring ~0.8 on logP optimization). **GPT-based SMILES generators (2021)**: high validity (~98%) and uniqueness, but slightly lower property optimization scores than specialized methods.
MOSES (2019)	Molecule generation	Standardized dataset (approx 1.9M molecules) and metrics for *unconditional* generation. Ensures apples-to-apples comparison of models generating drug-like molecules.	Validity (%), Unique @1000, FCD (distribution distance)	*VAE and GAN models (2019)*: ~100% valid, ~80% unique, FCD ~0.1–0.2. **Transformer LM on SMILES (2020)**: ~100% valid, ~90% unique, improved novelty; FCD competitive (~0.08).

Benchmark / Task	Domain	Description & Use Case	Metric	Notable Results (2020–2025)
				Indicates transformers can learn the distribution well.
TOMG-Bench (2024) (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation)	Text-driven chemistry	Open-ended molecule design via text instructions (edit/optimize/generate). Tests LLMs as medicinal chemistry assistants.	Custom compound success rate (meeting prompt criteria) and validity	*GPT-3.5* (2023): struggled, low success (significant invalid outputs). **Llama3.1-8B (2024)** fine-tuned on OpenMolIns: best on benchmark, 46% higher score than GPT-3.5 (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). *GPT-4* (if tested) expected to improve but results not public.
Bioinfo-Bench (2023) (A Simple Benchmark Framework for LLM Bioinformatics Skills ...)	Bioinformatics (Q&A)	200 questions covering bioinformatics problems (multiple-choice, sequence analysis, etc.) to test LLM knowledge of genomics and computational biology.	Accuracy (overall)	*GPT-4* (2023): exceeded **80%** on multiple-choice but struggled on coding questions; *ChatGPT* ~60%. Highlighted LLMs' gaps in specialized bioinformatics knowledge (BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning) (BioinformaticsBench: A collaboratively built large language model benchmark for

Benchmark / Task	Domain	Description & Use Case	Metric	Notable Results (2020–2025)
				Bioinformatics reasoning). (Limited coverage, spurring creation of bigger benchmarks.)
BioCoder (2023) (BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning)	Bioinformatics (coding)	1,000+ coding problems in bioinformatics extracted from sources (Rosalind, GitHub). Tests LLM's ability in programming for bioinformatics (parsing data, algorithms).	Code accuracy (pass rate)	*GPT-4 (2023)*: high success in known algorithms, but purely coding-based; not a general knowledge test. Revealed LLMs can solve many bioinformatics puzzles but may overfit to seen examples.
BioinformaticsBench (2024) (BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning) (BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning)	Bioinformatics (reasoning)	A new benchmark (602 questions) across 9 sub-domains of bioinformatics (genomics, proteomics, phylogenetics, etc.), focusing on analytical reasoning using textbooks and problem sets.	Accuracy (various formats: numeric, multiple-choice, T/F)	*GPT-4 (2024)*: expected to lead, but early results show need for external knowledge/tools in complex problems. Aims to provide a more comprehensive test than Bioinfo-Bench.
DNA Long Bench (2025) ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks) ((PDF) DNALongBench: A Benchmark Suite for	Genomics (long-range)	Benchmark suite for long DNA sequence prediction tasks (up to 1 million base pairs context). Tasks include predicting gene expression from regulatory DNA, enhancer–gene interactions, etc.	Task-specific (e.g., correlation for expression, accuracy for enhancer links)	*DNABERT and related (2021)*: perform well on short-sequence tasks (~0.9 AUC on motif finding). **Long-range transformers (2024)**: can capture some dependencies but **specialized models still**

Benchmark / Task	Domain	Description & Use Case	Metric	Notable Results (2020–2025)
Long-Range DNA Prediction Tasks)		Evaluates "DNA LLMs" on biologically meaningful long-range sequence tasks.		**outperform** on all 5 tasks ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks). Highlights that domain-specific architectures or training are needed for genomics; LLMs alone haven't solved it.

Table 2: Key benchmarks in the pharmaceutical and bioinformatics realm beyond pure text QA. These evaluate models on understanding and generating molecules, and on analyzing biological sequences or data. Performance of LLMs or related models is compared with domain-specific approaches. Generally, *task-specific models* and fine-tuned smaller models maintain an edge in structured domains (e.g., graph neural nets slightly outperform language models on molecular property prediction ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks), and expert bioinformatics tools still beat GPT-4 in gene prediction tasks (Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease - PMC)). However, LLMs are rapidly improving: GPT-style models show high validity in molecule generation and can solve many textbook bioinformatics questions. Each benchmark connects to an industry use case: property prediction for chemical screening, molecule generation for drug design, and genomic sequence interpretation for target discovery.

Industry Impact and Recent Trends

The landscape of LLM benchmarks in life sciences from 2020 to 2025 reveals a few clear trends. First, **domain-specific benchmarks have driven the creation of domain-specific models**. Efforts like BLUE and BLURB highlighted gaps of general models on biomedical text, leading to BioBERT, PubMedBERT, ClinicalBERT, BioMegatron, and others – each pushing the benchmark state-of-the-art by better ingesting biomedical corpora. For example, PubMedBERT (2020) trained solely on PubMed texts outperformed multi-domain BERT on nearly all BLURB tasks, especially NER and classification, due to handling domain jargon (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC). This specialization is valuable for pharmaceutical companies dealing with jargon-heavy texts (chemicals, genes, etc.). At the same time, the rise of very large general models like GPT-3, GPT-3.5, and GPT-4 introduced a new paradigm: models with **emergent capabilities that excel at reasoning-heavy benchmarks** even without domain tuning. The benchmarks discussed show a split: **information extraction** tasks (structured outputs like entity labels or relations) still see best performance from fine-tuned domain models (often smaller in size but trained on in-domain data) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). In contrast, **knowledge and reasoning** tasks (open QA, medical exams) have been *leapfrogged* by the likes of GPT-4 (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). For instance, no biomedical model came close to passing USMLE until GPT-4 did so with ease (Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments-Scientific Reports). This suggests that, for tasks requiring integration of vast knowledge (36 million PubMed articles, clinical expertise, etc.), the sheer

scale of general models gives them an advantage, presumably because they've seen more medical text than any curated corpus could offer.

Another trend is the **integration of retrieval and multi-modal data** in benchmarks. New benchmarks are emerging that don't treat language in isolation. The DNA Long Bench is an example where sequence data (DNA) is essentially another modality evaluated with language-model-like approaches ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks). Likewise, some biomedical QA benchmarks are starting to include providing references or combining text with tabular clinical data. The nature of evaluation is also adapting – beyond just accuracy, there's interest in *qualitative assessments* like consistency and lack of hallucination. In one 2024 study, qualitative metrics were reported for GPT-4 and others on generating clinical evidence summaries (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). Ensuring an LLM's answer is not only correct but also *justified and clear* is becoming part of "benchmarks," especially for sensitive domains like medicine.

From an industry perspective, the benchmarks covered serve as **key performance indicators** when selecting or developing an LLM for a particular application. If a team is building an automated literature review assistant, they will look at BioASQ and PubMedQA scores as a proxy for how well a candidate model might perform. If the goal is to implement an AI-driven molecule design tool, benchmarks like GuacaMol, MOSES, or TOMG-Bench are critical to gauge whether the model can actually propose valid, novel compounds. The benchmarks also help in **regulatory and validation contexts** – for example, a pharma company might report that their AI system was validated on a benchmark to demonstrate its reliability in a submission or white paper.

It's also worth noting that some **commercial benchmarks** exist internally. While not public, many pharma companies have curated test sets (e.g., a set of question-answer pairs about their proprietary drugs, or an internal corpus of annotated clinical trial reports) to evaluate LLMs before deployment. These often mirror the structure of public benchmarks but use company-specific data. Where possible, companies leverage public benchmarks first (for general capability) and then validate on private data. The public, academic benchmarks we've discussed thus form the first hurdle that any solution must clear.

In summary, large language model benchmarks in life sciences cover a spectrum from basic NLP tasks like entity extraction to complex reasoning and generative design problems. Over the last five years, performance on these benchmarks has dramatically improved – in some cases by tens of percentage points – due to both specialized domain models and breakthroughs in general LLMs. Table 3 provides a high-level summary linking each major benchmark to its primary industry use case and the current frontier of model performance:

Table 3. Benchmarks and Their Industry Use Cases & Top Performers

Benchmark	Primary Industry Use Case	Top Performing Models (2025)
BLURB (multi-task)	Text mining pipeline (NER, classification, etc.) – automating curation of biomedical knowledge.	*BioALBERT-large* (PubMed) – best overall BLURB score (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC); BioBERT, PubMedBERT close behind.
BioASQ (QA)	Biomedical research assistant – answering scientists' questions from literature.	Ensembles of BioBERT variants (fine-tuned) – top challenge winners; GPT-4 (zero-shot) performs well but not officially in competition.

Benchmark	Primary Industry Use Case	Top Performing Models (2025)
PubMedQA (QA)	Evidence extraction from papers – validating study findings for medical affairs.	*PMC-LLaMA 13B* fine-tuned – ~78% accuracy (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications); GPT-4 few-shot ~75% (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications).
MedQA (Clinical QA)	Clinical decision support – aiding diagnosis or medical education.	*GPT-4* – ~85% accuracy (expert level) (Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments-Scientific Reports); next best: Med-PaLM (Flan-PaLM) ~67%.
MoleculeNet (prop. pred.)	Early drug screening – predict properties and toxicity in silico.	*Graph neural nets* (EGCN, 2019) – top on many tasks; *MolBERT/MolT5* – close second, best on text-based features.
GuacaMol/MOSES (gen.)	De novo drug design – generate candidate compounds meeting desired criteria.	*Reinforcement Learning models (e.g., GraphGA)* – excel in goal optimization; LLMs (Transformer LM) – high validity and diversity, being integrated into pipelines.
TOMG-Bench (gen.)	Medicinal chemistry assistant via text – interactive molecule design with chemists.	*Llama3.1-8B (2024)* – specialized fine-tune leading performance (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation); commercial GPT-4 not yet publicly benchmarked.
Bioinfo-Bench (QA)	Bioinformatics Q&A – supporting genomic data analysis and interpretation.	*GPT-4* – best on Q&A, especially multiple-choice; struggles on coding without tools.
DNA Long Bench	Genomic regulatory insight – predicting gene expression or variant impact from sequence.	*Ensemble of specialized models* – best results (transformer + custom layers); purely pre-trained DNA-LLMs still catching up ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks).

This table reinforces that **no single model is best at everything** – a crucial point for practitioners. GPT-4 may be the best at medical reasoning, but a smaller BioBERT could be better for extracting a list of gene names from 1,000 documents due to fine-tuned accuracy and speed. Therefore, benchmarking across all these scenarios helps in creating a *portfolio of AI tools* in a pharmaceutical IT department: one might use a fine-tuned NER model for bulk text processing, a GPT-based QA model for an interactive chatbot, and a chemistry-specific transformer for molecular design.

Conclusion

Large language model benchmarks in life sciences have rapidly evolved, reflecting the growing capabilities of AI and the diverse needs of biomedical and pharmaceutical applications. From the early days of BLUE and BioASQ to the latest TOMG-Bench and DNA Long Bench, each benchmark has pushed models to new heights and exposed new challenges. Importantly, benchmarks serve as a bridge between academic advancement and industry adoption – they distill real-world tasks into measurable performance, ensuring that progress in the lab translates to practical impact. Between 2020 and 2025, we've witnessed *transformative improvements*: accuracy on medical QA tasks has essentially doubled with the advent of GPT-4 (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), and the feasibility of text-based molecule generation is now demonstrated (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation). Yet, the journey is ongoing. Open-source models are steadily closing the gap with commercial LLMs in many benchmarks, especially when fine-tuned or augmented with retrieval (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). Meanwhile, new benchmarks are targeting areas like **result summarization, clinical report generation, and patient data de-identification**, which will be crucial for next-generation healthcare NLP systems.

For IT professionals in pharma, keeping an eye on these benchmarks is more than an academic exercise – it is key to selecting the right model for the job and knowing the model's limitations. If an LLM is to be deployed for a critical task (say, analyzing safety reports), one should ensure it's evaluated on a relevant benchmark (perhaps an adverse event extraction task) and meets the performance bar observed in research. Benchmarks also hint at *failure modes* – for example, the qualitative analyses in some studies show that models like LLaMA-2 tend to hallucinate without few-shot examples (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications). Knowing this, one can design systems with necessary human oversight or use prompting techniques to mitigate issues.

In conclusion, the suite of LLM benchmarks in life sciences provides a comprehensive curriculum to "train" and test our AI systems. They cover the range from understanding a protein mention in a sentence all the way to hypothesizing a new drug molecule. As we move beyond 2025, we expect benchmarks to become even more realistic – incorporating multi-step workflows (e.g., find relevant papers *and then* answer a question), multimodal data (e.g., interpreting images or chemical structures alongside text), and stricter requirements for explanation and correctness (to satisfy regulatory demands). The continual improvement of models on benchmarks like those surveyed here gives optimism that LLMs will become reliable assistants in biomedical research and healthcare delivery. By following benchmark-driven development, the pharmaceutical industry can harness these AI advances with confidence, applying them to accelerate drug discovery, improve patient care, and unlock insights from the ever-growing mountains of biological data.

Sources: All data and model performance metrics referenced are drawn from published papers, benchmark leaderboards, and survey articles, including the BLURB benchmark paper (BLURB Leaderboard) (BLURB Leaderboard), BioALBERT results (Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT - PMC), a 2025 Nature Communications review of LLMs in biomedicine (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications) (Benchmarking large language models for biomedical natural language processing applications and recommendations-Nature Communications), the MedQA/USMLE evaluation reports (Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments-Scientific Reports), and recent arXiv papers introducing TOMG-Bench (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation) (TOMG-Bench: Evaluating LLMs on Text-based Open Molecule Generation), Bioinfo-Bench (BioinformaticsBench: A collaboratively built large language model benchmark for Bioinformatics reasoning), and DNA Long Bench ((PDF) DNALongBench: A Benchmark Suite for Long-

Range DNA Prediction Tasks) ((PDF) DNALongBench: A Benchmark Suite for Long-Range DNA Prediction Tasks), among others. These sources are cited throughout the text for further reading on each benchmark and finding.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is an AI software development company specializing in helping life-science companies implement and leverage artificial intelligence solutions. Founded in 2023 by Adrien Laurent and based in San Jose, California.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.