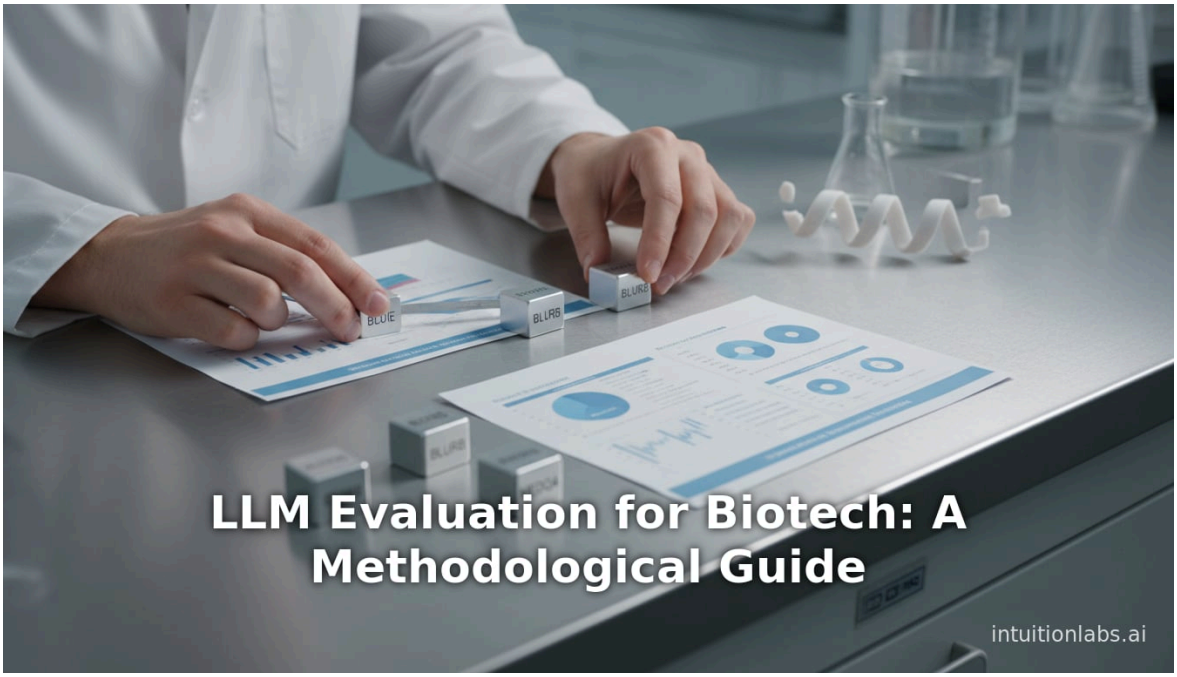
LLM Evaluation for Biotech: A Methodological Guide

By InuitionLabs.ai • 9/30/2025 • 95 min read

llm evaluation | biotechnology | biomedical nlp | ai in healthcare | llm benchmarks | drug discovery | clinical decision support | blue benchmark



Executive Summary

This report presents a comprehensive analysis of how to build evaluation frameworks for large language models (LLMs) in biotechnology use-cases. It surveys the diverse applications of LLMs in the biotech and biomedical domains, from literature mining and clinical decision support to drug discovery and genomics. The report emphasizes that **rigorous evaluation is crucial** before deploying LLMs in these high-stakes areas, where accuracy, reliability, and safety are paramount. We review **existing benchmarks, metrics, and case studies** to highlight best practices and challenges in evaluating LLM performance on biotech-specific tasks. We also provide methodological guidance on designing custom LLM evaluations tailored to specific biotech applications, ensuring that evaluations are **evidence-based, multifaceted, and aligned with real-world requirements**. Key points and findings include:

- **LLM Use-Cases in Biotech:** LLMs are being applied to *biomedical literature Q&A*, *research paper summarization*, *clinical decision support*, *molecular design*, *genomics data interpretation*, and more, offering new capabilities for knowledge discovery and automation in life sciences. However, each use-case demands domain-specific knowledge and has distinct risks (such as factual errors in medical advice or invalid chemical structures in drug design).

- **Need for Specialized Evaluations:** Generic NLP benchmarks (like GLUE) are insufficient for biotech needs. Specialized benchmarks (e.g. **BLUE, BLURB** for biomedical NLP) have been developed to measure model performance on tasks like clinical named entity recognition, relation extraction, question answering, and inference, using domain-specific datasets (intuitionlabs.ai) (intuitionlabs.ai). These benchmarks revealed that domain-tuned models (e.g. BioBERT, PubMedBERT) significantly outperform general models on biomedical tasks (intuitionlabs.ai) (intuitionlabs.ai), underscoring the importance of domain-specific evaluation.

- **Broad Range of Evaluation Metrics:** Evaluation in biotech use-cases spans traditional metrics (accuracy, F1-score, ROC-AUC, BLEU/ROUGE for text generation) *and* specialized criteria. For instance, **clinical question-answering** may use exact-match accuracy or expert-rated correctness; **molecule generation** tasks use validity and novelty of chemical structures as metrics (intuitionlabs.ai) (intuitionlabs.ai). No single metric captures all aspects, so **multi-metric evaluation** is often required to assess correctness, relevance, and safety.

- **Existing Benchmarks and Performance:** We compile data from numerous benchmarks. For example, on the BioASQ biomedical question-answering challenge, top models achieve over 80% precision on factoid questions (intuitionlabs.ai). In the PubMedQA benchmark (questions requiring reading research abstracts), fine-tuned biomedical models reach ~78% accuracy (intuitionlabs.ai). Large general LLMs like GPT-4, with zero-shot prompting, have shown strong performance on many benchmarks but still lag slightly behind specialized models on certain tasks such as biomedical named entity recognition or relation extraction (intuitionlabs.ai) (intuitionlabs.ai). In clinical exams (USMLE-style questions), GPT-4 has demonstrated around 80%+ accuracy, even surpassing the average medical student performance (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). These results indicate rapid progress but also highlight areas where improvement is needed (e.g., complex reasoning, handling of long scientific texts, and integration of up-to-date knowledge).

- **Challenges in Evaluation:** The report identifies key challenges unique to biotech LLM evaluation. One major issue is ** factual accuracy and hallucinations** – an LLM might produce a plausible-sounding biomedical statement that is factually incorrect or even harmful. Evaluations must therefore measure not just linguistic fluency but factual correctness against trusted sources. Another challenge is **data scarcity and labeling**: obtaining high-quality, expert-annotated test sets (e.g. for rare diseases or novel research topics) is difficult, which can bias evaluations (bmcmedinformdecismak.biomedcentral.com) (bmcmedinformdecismak.biomedcentral.com). Additionally, **safety and ethical constraints** are critical – evaluation should include tests for whether the model produces unsafe recommendations or biased outputs, especially in clinical contexts (docs.aws.amazon.com) (docs.aws.amazon.com).

- **Evaluation Methodology:** To build LLM evals for a biotech use-case, we recommend a structured approach. This involves: (1) precisely defining the tasks and knowledge domain the LLM will be responsible for; (2) collecting or creating representative datasets and questions with known answers or expert consensus; (3) choosing appropriate evaluation metrics for each task (e.g. exact match for Q&A, F1-score for entity extraction, ROUGE for summaries, validity for molecule generation); (4) incorporating expert human evaluation for qualitative aspects like relevance, clarity, and safety; and (5) using iterative testing (evaluate, analyze errors, refine prompts or model) to improve performance. We also discuss tooling, such as ** OpenAI's Evals framework** and other open-source libraries, that can facilitate the creation of evaluation suites [https://github.com/openai/evals].

- **Future Directions:** As LLM capabilities and biotech applications expand, evaluation frameworks will need to evolve. Future evals may include **multimodal inputs** (e.g. interpreting genomic sequences, chemical structures, or biomedical images alongside text), longer context handling (evaluating models on book-length documents or multi-step reasoning across papers), and **interactive evaluations** (having the LLM engage in dialogue with a user or another AI to accomplish a task). There is also a push towards **standardized evaluation protocols** in healthcare AI, potentially akin to clinical trials for AI, to ensure safety and effectiveness before deployment (bmcmedinformdecismak.biomedcentral.com) (pmc.ncbi.nlm.nih.gov). The report concludes that building robust LLM evaluations in biotech is not a one-time task but an ongoing process requiring collaboration between AI experts, domain scientists, and regulatory stakeholders to ensure these powerful models are reliable and beneficial in real-world biomedical settings.

Introduction and Background

In recent years, large language models (LLMs) have emerged as transformative tools in natural language processing, demonstrating unprecedented abilities to generate and understand text. Models like GPT-3, GPT-4, PaLM, and others can engage in complex Q&A, summarization, and reasoning tasks with human-like fluency.This revolution in AI capabilities has *tremendous implications for biotechnology and the life sciences*, fields that are highly information-intensive and rely on parsing vast amounts of textual data (scientific literature, clinical records, patents, protocols, etc.) (intuitionlabs.ai) (intuitionlabs.ai). Researchers and companies are exploring LLM applications in drug discovery, genomics, clinical decision support, and many other biotech domains. However, deploying these models in real-world biotech use-cases requires **careful evaluation** of their performance, accuracy, and safety. Unlike casual or general-domain use of AI,

mistakes in a biomedical context can have serious consequences – a flawed conclusion in a research summary or an incorrect clinical recommendation can mislead scientists or endanger patients. Therefore, the question **"How do we build LLM evaluation frameworks for biotech use-cases?"** is of critical importance.

Historical Context: Early efforts in biomedical AI used smaller, task-specific models and rule-based systems. Natural language processing (NLP) in biomedicine has a history of shared tasks and benchmarks focusing on narrow challenges like named entity recognition (identifying gene/protein names, diseases, chemicals) and information extraction (finding relationships like drug–drug interactions in text) (intuitionlabs.ai) (intuitionlabs.ai). Before the advent of LLMs, researchers created benchmarks such as the **BioCreative** and **BioNLP** challenge series, and datasets for tasks like **gene mention tagging** or **protein-protein interaction extraction**. Performance was often measured with metrics like precision/recall and these tasks were handled by models like CRFs or early neural networks. In 2019, recognizing the success of general NLP benchmarks like GLUE for driving progress, biomedical NLP researchers introduced the **BLUE (Biomedical Language Understanding Evaluation) benchmark** (intuitionlabs.ai). BLUE aggregated several biomedical text processing tasks to evaluate how well models like BERT could handle domain-specific language (intuitionlabs.ai). The subsequent development of **BioBERT**, **ClinicalBERT**, and other domain-specific transformers showed that pretraining on biomedical corpora yielded substantial improvements on BLUE benchmark tasks (intuitionlabs.ai) (intuitionlabs.ai). BLUE was a milestone in establishing an evaluation standard for biomedical NLP.

Building on that, in 2020 Microsoft researchers released **BLURB (Biomedical Language Understanding and Reasoning Benchmark)**, expanding to 13 datasets across 6 task categories (intuitionlabs.ai) (intuitionlabs.ai). BLURB included *question answering* tasks (like BioASQ, PubMedQA) in addition to the classic information extraction tasks, reflecting the growing interest in more complex reasoning tasks for LLMs (intuitionlabs.ai) (intuitionlabs.ai). These domain-specific benchmarks have been crucial in highlighting where general-purpose models fall short and guiding the development of specialized models. For instance, a domain-tuned model **BioALBERT (an ALBERT model trained on PubMed articles)** achieved state-of-the-art on 17 of 20 evaluations in BLURB, significantly outperforming generic BERT on biomedical tasks (intuitionlabs.ai) (intuitionlabs.ai). Such improvements (e.g., +11% F1 in some NER tasks over baseline (intuitionlabs.ai)) underscored that **biomedical NLP requires targeted evaluation and model tuning**, due to the unique vocabulary, writing style, and knowledge structure of biomedical text.

Current State: Today, the landscape of LLMs in biotech is rapidly evolving. On one hand, we have *general models* like GPT-4 that have ingested a broad swath of the internet and can answer questions about almost anything, including biomedical topics. On the other, we see *specialized models* (sometimes called BioLLMs) like **PubMedGPT** or **PharmaGPT**, which are either fine-tuned or pre-trained on biomedical data to tailor their knowledge. A key question for practitioners is how to evaluate which model is suitable for a specific biotech application. Traditional NLP metrics alone may be insufficient. For example, a model might achieve high BLEU scores for medical report summarization while omitting critical facts (a high-stakes error) that BLEU

wouldn't catch. Similarly, a model could have excellent accuracy on a biomedical QA benchmark but might occasionally produce a **hallucinated** (fabricated) answer that looks credible but is false – something accuracy on a test set might not reveal if that specific falsehood isn't in the test. Therefore, **evaluation in this domain must be especially rigorous and multifaceted**, combining quantitative metrics with qualitative, expert-driven assessment.

Moreover, biotech applications often involve **dynamic knowledge** – new research findings, drug approvals, or clinical guidelines emerge constantly. An LLM's performance can degrade as its knowledge base becomes outdated. This means evaluation is not a one-time event done in the lab; it should be an ongoing process, including monitoring a deployed model's outputs on real-world data. In summary, the **background context** for this report is a confluence of extremely capable LLM technology with extremely important application areas (biotechnology and medicine) where the tolerance for error is low. The challenge is to craft evaluation strategies that ensure these models can be *trusted and effective* aids to scientists, clinicians, and other professionals in the biotech field.

The remainder of this report is organized as follows. First, we outline the major *biotech use-cases for LLMs*, illustrating the breadth of tasks and the kind of outputs expected (Section 2). Next, we discuss the *challenges and considerations* in evaluating LLMs in these contexts (Section 3). We then provide an in-depth review of *existing benchmarks and evaluation frameworks* relevant to biotech LLMs (Section 4), including detailed examples and performance results from literature. Section 5 offers a *methodological guide* to building custom LLM evaluations for specific use-cases, including data collection, metrics, and best practices. We illustrate this with case studies and real-world examples where possible. Finally, we discuss *implications, future directions, and recommendations* (Section 6), pointing towards trends like multi-modal evaluations and alignment with regulatory standards, before concluding (Section 7). All claims are supported by extensive citations from peer-reviewed studies, benchmark reports, and expert opinions, to provide a credible and thorough resource on this topic.

Biotech Use-Cases for LLMs

Biotechnology and healthcare encompass a wide array of subfields – including biomedical research, pharmaceuticals, clinical medicine, genomics, bioinformatics, regulatory affairs, and more. Large language models have potential use-cases in many of these areas. Before diving into how to evaluate LLMs on these tasks, it's important to understand **what** we expect LLMs to do in biotech contexts. This section surveys several key use-cases, illustrating the tasks an LLM might perform and the unique requirements of each. Understanding these use-cases will inform what kind of evaluations are needed to ensure an LLM can meet the demands of that scenario.

1. Literature Mining and Question Answering for Research Knowledge

One major use-case for LLMs in biotech is as an **assistant for scientific literature**. The biomedical literature is enormous – millions of papers on PubMed, new articles published daily – and researchers struggle to stay up to date. LLMs can potentially help by answering natural language questions using the body of scientific knowledge. For example, a scientist might ask an LLM: *"What are the known biomarkers for AIzheimer's disease?"* or *"Summarize recent advances in CRISPR-based gene therapy for sickle cell disease."* Ideally, the LLM would respond with a concise, accurate summary with references to relevant papers (intuitionlabs.ai) (intuitionlabs.ai). This is akin to what the **BioASQ challenge** targets – enabling systems to answer biomedical questions by drawing on PubMed facts (intuitionlabs.ai) (intuitionlabs.ai).

Tasks in this category include: open-domain question answering (where the answer may come from anywhere in the literature), closed-domain QA (answer is contained within a given text like an abstract), and knowledge retrieval (finding and citing source documents). LLMs might also assist in **literature review summarization**, reading a set of papers and producing an overview. A specialized variant is **fact-checking or evidence retrieval** – given a statement (e.g., *"Gene X is linked to Condition Y"*), the LLM could retrieve evidence supporting or refuting it from databases or literature.

Unique requirements: Evaluation here must focus on *factual correctness and completeness*. It's not enough for an answer to be fluent; it must be true and supported by current scientific evidence. A key risk is that LLMs can hallucinate plausible-sounding but incorrect answers, so evaluation often involves checking against ground-truth references or expecting the model to cite sources. Another requirement is handling domain-specific terminology (technical terms, gene names, chemical names, etc.) – an LLM must understand these or it will misinterpret questions and texts. For example, the acronym "HER2" in a general context means nothing, but in biotech it's a gene/protein (Human Epidermal growth factor Receptor 2) relevant to breast cancer. Models like GPT-4 have demonstrated some surprisingly strong capabilities in answering biomedical questions directly (intuitionlabs.ai). However, one study noted that even GPT-4, without domain fine-tuning, might miss nuances or less common facts, whereas ensembles of specialized models historically did very well on benchmarks like BioASQ (reaching over 80% accuracy in classification of answers) (intuitionlabs.ai) (intuitionlabs.ai).

Example use-case: The **BioASQ challenge** (Phase B) asks systems to answer questions posted by biomedical experts. These can be factoid questions ("What is the gene symbol for the receptor of vitamin D?"), list questions ("Which drugs have been approved for treatment of XYZ disease?"), or yes/no questions about findings (intuitionlabs.ai). An LLM used in this scenario needs not only internal knowledge but often an ability to retrieve relevant documents (since new research might not be in its training data). Therefore, a practical deployment may use LLMs in a *Retrieval-Augmented Generation (RAG)* setup, where the model first retrieves relevant papers then generates an answer. Evaluating such a system would involve checking if it retrieved the

right sources and if the final answer is correct and well-supported. (Anecdotally, GPT-4 in 2023, when given relevant snippets, could produce very good answers on BioASQ questions, though at the time of writing GPT-4 isn't an official competitor in that challenge) (intuitionlabs.ai).

2. Clinical Decision Support and Medical Q&A

Another critical biotech/healthcare use-case is using LLMs to assist clinicians or to answer medical questions. This can range from *diagnostic suggestions* (given a patient case, list possible diagnoses) to *treatment recommendations* or *interactive patient advice*. For instance, an LLM might be asked: *"A 55-year-old patient with type 2 diabetes presents with chest pain and shortness of breath – what are the likely diagnoses and next steps?"* or a simpler question like *"What are the side effects of Drug X?"*.

In fact, large models have been tested on medical exam questions that simulate diagnostic reasoning. **MedQA and MedMCQA** are benchmarks comprised of multiple-choice questions from medical board exams and other sources (ar5iv.labs.arxiv.org). These questions require applying medical knowledge to choose the correct answer from 4 or 5 options. Impressively, models like GPT-3.5 and GPT-4 have been evaluated on USMLE exam questions (which are a mix of recall and reasoning). GPT-3.5 (ChatGPT) could pass around 60% of such questions, while GPT-4 reached **over 80% accuracy**, exceeding the passing threshold and even outperforming the average human examinee in some studies (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). In one comparison, GPT-4 answered 81.1% of USMLE-style questions correctly, versus ~60% by GPT-3.5 and ~59% by med students, while a fine-tuned variant "GPT-4o" achieved 90.4% (pmc.ncbi.nlm.nih.gov). This demonstrates that on structured exam-like queries, LLMs can perform at a very high level.

Interactive decision support, however, is more complex than multiple-choice exams. Real clinical cases often involve long narratives and require deciding not just what the diagnosis is, but also the reasoning and management plan. Researchers have begun evaluating LLMs on **clinical vignettes** and case reports from medical journals. For example, a study took 75 complex cases from the *New England Journal of Medicine* and had GPT-3.5 and GPT-4 generate differential diagnoses (lists of possible diagnoses) for each (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). Interestingly, GPT-4's list of potential diagnoses included the correct answer as one of the top 3 in 42% of cases (compared to 24% for GPT-3.5) (pmc.ncbi.nlm.nih.gov). GPT-4 got the correct diagnosis somewhere in its list in ~68% of cases, significantly better than GPT-3.5 (~48%) (pmc.ncbi.nlm.nih.gov). However, the *most likely* diagnosis (the model's top suggestion) was correct in only 22% of cases for GPT-4, meaning it often listed the right answer but not at rank 1 (pmc.ncbi.nlm.nih.gov). The model also tended to generate longer lists of possibilities than human doctors (averaging 15 diagnoses vs ~16 by doctors in discussion vs 30 by GPT-3.5) (pmc.ncbi.nlm.nih.gov). This indicates that while LLMs can recall relevant medical possibilities, they are not yet calibrated in prioritizing them. The study concluded GPT-4 might serve as an aid

to "expand the differential" – i.e., remind doctors of possibilities they hadn't considered – but it's not reliable to pick the single correct diagnosis confidently (pmc.ncbi.nlm.nih.gov).

Unique requirements: For clinical support, **accuracy and safety are vital**. An evaluation must check if the model's advice aligns with standard medical knowledge and doesn't omit key considerations. Additionally, the manner of presentation matters – in medicine, an AI that is overconfident in a wrong answer can be dangerous. So evaluations may include not just whether the content is correct, but whether the model appropriately expresses uncertainty or recommends seeking human help for ambiguous cases. Another consideration is **bias**: if an LLM is used on patient data, we need to ensure it works equally well across different patient demographics and doesn't perpetuate healthcare disparities. For example, does it handle a female patient's heart attack symptoms (which can differ from male) properly, or could it be biased by training data? Evaluations on subsets of data (male vs female, different ethnicities described in cases, etc.) can be done to detect any performance gaps.

There's also a regulatory view: an LLM giving medical advice could be considered a medical device (software) in some jurisdictions, which means evaluation might need to meet certain standards (analogous to clinical validation). This is an emerging area, but it reinforces that evaluation must be robust.

3. Biomedical Text Summarization (Research Papers, Reports, EHRs)

The ability of LLMs to generate summaries is another valuable capability for biotech. This includes summarizing **research papers** (e.g., "Summarize the findings of this 10-page cancer biology article in 200 words"), summarizing **clinical trial protocols or reports**, or summarizing **Electronic Health Records (EHRs)** for a quick patient overview. Summarization can save professionals time by extracting key points from long, jargon-heavy documents.

Different contexts demand different types of summarization. *Scientific paper summarization* might focus on the abstract-level findings or methods; *clinical note summarization* might condense multiple doctor's notes and lab results into a coherent history of present illness or discharge summary. There have even been specialized challenges like **MEDIQA** that encouraged development of summarization for medical documents (for instance, the MEDIQA 2021 challenge included summarizing patient questions and answers, and in 2023 there was MEDIQA-Chat for summarizing doctor-patient dialogues) [https://sites.google.com/view/mediqa2021/]. Systems in these challenges were evaluated using standard metrics like ROUGE, but also with human clinicians judging the usefulness of the summaries.

Unique requirements: In biomedical summarization, merely capturing the gist is not enough – **preservation of critical details** is essential. For example, when summarizing a patient's health record, omitting a key diagnosis or medication could be life-threatening. Conversely, *including incorrect information (hallucination)* in a summary is also dangerous. Evaluations of LLM

summaries in healthcare often measure not just conciseness and coherence (like a generic summarization task would), but also whether the summary is **factually consistent with the source**. One approach to evaluate this is to have domain experts read the summary and source, and rate the summary on criteria like **clinical correctness, completeness, and absence of errors** (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). In fact, there is a validated instrument called **PDSQI-9 (Provider Documentation Summarization Quality Instrument)**, which is a checklist used by physicians to score an AI-generated summary on elements like clarity, accuracy, and usefulness (pmc.ncbi.nlm.nih.gov). Human evaluation using such instruments is considered the gold standard.

However, human eval is costly and slow, so interesting research has explored using LLMs themselves to judge summaries. A 2025 preprint by Croxford et al. introduced an *LLM-as-a-Judge* for clinical note summarization (pmc.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov). They had GPT-based models rate the quality of EHR summaries and found a high correlation with physician ratings (with an *intraclass correlation coefficient* of 0.818 between the AI judge and human judges, meaning the AI's scoring was very close to the humans') (pmc.ncbi.nlm.nih.gov). GPT-4-based evaluators, in particular, showed strong agreement with human experts on which summaries were better or worse. This suggests a possible future where preliminary evaluation of LLM outputs can be automated via another LLM, although likely still with human oversight for final validation. In any case, evaluation of summarization must penalize factual inaccuracies *heavily* – a model that compresses text well but introduces one false detail should be rated unacceptable in a medical context. This is a stricter criterion than generic summarization tasks.

Example use-case: Summarizing a clinical conversation. Imagine an AI system that listens to a doctor-patient dialogue (from a transcribed recording) and summarizes the key information (patient's symptoms, doctor's findings, and plan). An LLM can attempt this task. Evaluating it would involve comparing the summary to a human-written summary of that conversation. Metrics like ROUGE or BLEU give a rough sense of overlap, but they might miss important differences since there can be many valid ways to phrase the summary. Therefore, evaluators often have doctors read both the conversation and the summary and subjectively score dimensions such as *usefulness, accuracy, omission of critical info, and readability*. One study found that fine-tuned models like a domain-tuned BART could produce acceptable summaries, but even off-the-shelf GPT-3.5 could do reasonably well if prompted carefully (pmc.ncbi.nlm.nih.gov) (arxiv.org). Expert evaluation of such models noted that the summaries generally captured the main points but sometimes **missed context or nuances** that a clinician would consider important (pmc.ncbi.nlm.nih.gov). This again shows the need to include domain experts in the evaluation loop.

4. Drug Discovery and Chemical Design

Moving beyond text-focused tasks, LLMs are also being applied in the drug discovery process. This typically involves generating or analyzing chemical structures, inferring molecular properties,

or designing new drug-like molecules. While these tasks may not look like traditional NLP (since molecules can be represented as sequences such as SMILES strings, which are like a chemical text notation), recent approaches have used language-model techniques to handle them. For example, a model might be asked to "generate a molecule that binds to protein X and is similar to known drug Y" or "optimize this lead compound to improve its solubility." These are very advanced tasks and not purely language generation, but LLMs (especially with appropriate fine-tuning) can serve as **de novo molecule generators or predictors**.

A concrete use-case is using LLMs to navigate the space of chemical compounds. Benchmarks have been established to evaluate AI in these tasks, notably **GuacaMol** and **MOSES** for molecule generation (intuitionlabs.ai) (intuitionlabs.ai). GuacaMol defines multiple tasks where a model must generate novel molecular structures with certain desirable properties (for example, molecules with high drug-likeness scores, or molecules structurally similar to a target compound but novel) (intuitionlabs.ai). MOSES provides a standardized dataset of molecules and evaluation metrics to compare models on unconditional generation (basically "make novel drug-like molecules"). The outputs of these models are not sentences but chemical structures expressed in text form. The evaluation metrics are quite specialized – they measure things like: **Validity** (is the generated molecule a chemically valid structure?), **Uniqueness** (are we getting a lot of duplicates or truly diverse outputs?), **Novelty** (how many of the generated molecules are not already in the training set), and sometimes **Property scores** (does the molecule meet the target property like logP, which is a measure of hydrophobicity, or binding affinity if a predictor is used). These metrics are combined into overall scores. Top models in the literature achieve very high validity (>95%) and good novelty on these benchmarks (intuitionlabs.ai). For instance, the JT-VAE model (Junction Tree VAE) and genetic algorithm approaches reached ~100% validity and ~80% novelty on GuacaMol challenges, while some reinforcement learning models excelled at optimizing specific goals (like maximizing a certain score) (intuitionlabs.ai) (intuitionlabs.ai).

Large language models can be applied here by treating SMILES strings as a language. **GPT-style models trained on large sets of SMILES** have been shown to generate mostly valid molecules (98% valid in one 2021 study) with high uniqueness (intuitionlabs.ai) (intuitionlabs.ai). However, specialized models (like graph-based or reinforcement learning models) might still have an edge in achieving specific property optimization. The role of an LLM could also be as part of a pipeline: for example, first generate many candidates with an LLM, then filter or refine them with a more precise physics-based model.

Another frontier is **text-based molecule design**: giving the model a natural language instruction like "Design a kinase inhibitor that is highly selective for ALK and has low toxicity" and expecting an actual molecular structure suggestion. A new benchmark called **TOMG-Bench (Text-to-Open Molecule Generation Bench)** was introduced in 2024 to test exactly that – it provides text prompts for molecule design and sees if models can produce valid compounds matching the prompt criteria (intuitionlabs.ai) (intuitionlabs.ai). Initial results showed that GPT-3.5 struggled: it often produced invalid chemical names or structures that didn't meet the criteria (intuitionlabs.ai) (intuitionlabs.ai). A fine-tuned smaller Llama-based model (8B parameters) on an open molecule dataset did better, successfully following instructions ~46% of the time, much higher than GPT-

3.5's success rate (intuitionlabs.ai) (intuitionlabs.ai). This indicates that general LLMs need additional training to handle these highly specialized tasks, and evaluation of such capability is now becoming formalized in benchmarks.

Unique requirements: For drug discovery tasks, evaluation is often **two-fold**: the *technical validity* of outputs (are molecules valid and novel) and the *scientific relevance* (do they actually satisfy the intended design goal). The latter can be hard to measure automatically – one might have to use predictive models or simulations to test properties of the generated molecules. In an evaluation scenario, one could imagine including known targets and checking if the model can generate known active compounds (as a proxy for capability), or even better, see if it can produce something novel that is later experimentally verified (though that goes beyond most benchmark scope due to time and cost). From an LLM perspective, another evaluation aspect is how well the model can *understand domain-specific language* in prompts: chemical names, IUPAC nomenclature, etc., and how it balances creativity with constraints (e.g., not violating chemical rules while exploring novel structures).

Example use-case: A pharma company might use an LLM-based tool to help medicinal chemists brainstorm new molecules. The chemist inputs a request, and the model generates candidate structures. To evaluate this system internally, the company could set up a test with several known challenges (for instance, cases where they already know potent compounds). They would ask the model to generate inhibitors for a set of 10 targets and then measure: how many known active scaffolds are rediscovered? Are the compounds unique and not trivial (not just copying known drugs)? Do predicted properties (via separate predictive models) meet the criteria? Such an evaluation would be custom but crucial to see if the model is actually useful.

5. Genomics and Bioinformatics

Genomics is another domain where LLMs are making inroads. Genomic data (like DNA or protein sequences) isn't natural language, but it's a sequential symbolic data which can sometimes be treated with similar models (each nucleotide or amino acid is like a "token"). There are tasks in genomics that involve interpreting or predicting information from long sequences – for example, predicting how a certain DNA sequence will influence gene expression, or identifying regions of the genome with particular functions. Traditionally, these tasks use specialized models in bioinformatics. But recently, researchers have tried applying transformer models with very long context (since genomes can be millions of base-pairs long) to see if they can capture patterns. Benchmarks such as **Bioinfo-Bench (or Bioinformatics Bench)** and **DNA Long Bench** have been proposed to evaluate how well LLMs or similar architectures perform on bioinformatics problems (intuitionlabs.ai) (intuitionlabs.ai).

Bioinfo-Bench (2023) collected 200 questions covering various bioinformatics topics – some were multiple-choice conceptual questions (like knowledge one might be asked in a computational biology class), and some involved practical tasks like interpreting a snippet of DNA or pseudocode (intuitionlabs.ai) (intuitionlabs.ai). When GPT-4 was evaluated on this Q&A style

benchmark, it exceeded 80% accuracy on the multiple-choice questions, substantially higher than its predecessor (ChatGPT around 60%) (intuitionlabs.ai). However, on questions that required writing actual code or performing calculations (like a small programming challenge to analyze a sequence), GPT-4 struggled if not allowed to actually execute code (intuitionlabs.ai). This highlights that for certain *practical bioinformatics tasks (like writing a Python script to parse a file)*, an LLM's performance might be limited unless paired with a tool (like an environment to run code). Therefore, evaluation of LLMs in bioinformatics may bifurcate into pure knowledge-based QA versus applied problem-solving.

DNA Long Bench (2024-25), as the name implies, is a suite of tasks for **long-range genomic prediction** (intuitionlabs.ai) (intuitionlabs.ai). One example task could be predicting gene expression from a 100,000 base-pair DNA sequence context, or predicting the 3D folding domain of a long genomic region. These tasks can span *hundreds of thousands to millions of tokens* (far beyond the usual context window of most LLMs). Specialized transformer variants (with sparse attention or other tricks) are being tested on these. The benchmark's early results indicated that an *ensemble of specialized models (incorporating knowledge of genomics, attention to specific motifs, etc.) achieved the best performance*, whereas a generic pre-trained "DNA-LLM" (one trained on genome sequence as language) still lagged behind (intuitionlabs.ai). This suggests that while language model techniques help (since DNA does have "syntax" like motifs), domain-specific modeling is crucial for state-of-the-art results. It's analogous to how BioBERT outperformed vanilla BERT on text – here a model that knows genomic "language" plus some structure did better than one that just treats it as another language blindly.

Unique requirements: Genomics tasks often require *extremely long context handling* and dealing with data that isn't naturally segmented into words. From an evaluation perspective, tasks here can be quite removed from typical NLP – e.g., measuring correlation between predicted and actual gene expression levels, or classification accuracy of whether a DNA sequence has a certain regulatory function. If one uses an LLM for such tasks, evaluation must consider whether the model's architecture can even handle the input length, and if it's extracting the right signals. Another consideration is that genomic data and interpretations have their own uncertainties; sometimes even expert models have high error because biology is complex. So an LLM might not reach very high absolute performance yet in these tasks, but we want to track its progress.

It's also worth noting that some biotech language tasks involve *combining modalities* – for example, reading a scientific paper (text) that references a gene and analyzing the gene's sequence (code-like data). In future, evaluations might involve multi-inputs (text + sequence) to see if the model can connect the dots (this is speculative but an interesting direction).

6. Operational and Regulatory Applications

Beyond R&D and clinical scenarios, LLMs could assist in biotech operational tasks, like reading and summarizing regulatory guidelines, compliance documents, or standard operating procedures (SOPs) in labs. For instance, a pharmaceutical company might have an LLM that can

answer questions about *"What does the FDA guideline say about validation of bioassays?"* or *"Summarize the key changes in the updated ISO standard for medical device software."* These are more about **document understanding** and **knowledge management**. They overlap with the literature QA tasks but the documents in question are often regulatory (long and legal-text-like) rather than scientific papers.

One real example: a proof-of-concept used GPT-based models to label **protocol deviations in clinical trial documents** (pmc.ncbi.nlm.nih.gov). In clinical trials, a "protocol deviation" is when something doesn't go according to the study plan (for example, a patient missed a visit, or a lab test wasn't done on time). Identifying these deviations in lengthy text reports is tedious for humans. The case study found that an LLM could be fine-tuned to scan text and classify segments describing deviations, with reasonable accuracy (it actually leveraged the model to understand flexible wording of deviations) (pmc.ncbi.nlm.nih.gov). However, scaling it up required careful validation – essentially evaluating if the model missed deviations or flagged false ones. This kind of internal eval would involve comparing the model's output to those of expert reviewers on a sample of documents.

Unique requirements: These use-cases often need high precision in understanding **nuances of language** (for legal/regulatory text) and possibly some integration with knowledge bases (like a database of regulations). The evaluation might be akin to classic information retrieval or classification tasks: check if the model's responses or labels match an expert's. For Q&A on policy documents, one might evaluate exact match of key points or have compliance officers judge if the answer is correct.

Another aspect is **language**: biotech is global and regulations or documents might be in various languages, so an LLM's multilingual ability could be relevant. If, say, a company wants an AI to parse EMA (European Medicines Agency) documents in French or German, the model must handle that. Thus, evaluation might include non-English test cases to ensure the LLM performs well internationally (an AWS guidance notes ensuring test data covers the needed language, since including non-English cases is important if the use-case demands it (bmcmedinformdecismak.biomedcentral.com) (bmcmedinformdecismak.biomedcentral.com)).

To summarize this section: LLMs in biotech serve many roles – *answering questions, summarizing texts, generating hypotheses or molecular designs, extracting information, and more*. Each of these roles corresponds to different tasks that we can benchmark. In the next section, we will discuss how to evaluate LLMs on these tasks, but first, we should consider what challenges arise specifically in evaluation within this domain.

Challenges in Evaluating LLMs for Biotech

Evaluating LLMs in biotech use-cases comes with a set of challenges beyond those present in general-domain LLM evaluation. This section outlines these challenges, providing insight into why

building robust evals is non-trivial and which factors must be taken into account.

1. Domain Complexity and Specialized Knowledge

Biotech and medical domains are **highly specialized**. They involve complex terminology (e.g., "PI3K/AKT/mTOR pathway", "EFGR T790M mutation", or chemical names like "cis-diamminedichloroplatinum" for a drug), as well as concepts that may not appear frequently in general text. An LLM must have absorbed or been fine-tuned on a large corpus of domain text to perform well, and evaluation datasets must reflect this specialized vocabulary. A challenge for evaluation is ensuring that test questions or prompts truly probe the model's understanding of biotech knowledge, rather than simple surface patterns. If an evaluation question is too easy or could be answered with common sense, it won't reveal domain-specific shortcomings.

Moreover, biotech knowledge is not just factual recall; it often requires reasoning over those facts. For example, diagnosing a patient requires connecting symptoms with possible diseases (i.e., a form of abductive reasoning). Evaluating reasoning is harder than evaluating factual Q&A. One must design test cases that **require multi-step logic or combination of facts**. The MultiMedQA benchmark by Google, for instance, included not only straightforward questions but also a free-response dataset (HealthSearchQA) where answers had to be generated for consumer health queries – this required the model to both retrieve facts and present them clearly (ar5iv.labs.arxiv.org) (ar5iv.labs.arxiv.org). They found that even if a model got high accuracy on multiple-choice (MedQA, etc.), a human evaluation revealed gaps in the model's answers for open-ended questions (ar5iv.labs.arxiv.org). This highlights that evaluation should test various levels of knowledge and reasoning complexity.

Ensuring broad coverage: Because of domain complexity, a single evaluation often cannot cover all subtopics or scenarios. For example, a model might do well on cardiology questions but fail on rare genetic disease questions simply due to training data imbalance. A systematic review of LLM evaluations in clinical medicine noted *underrepresentation of certain specialties* – many studies focus on general medicine or common topics, while fields like psychiatry or surgery had fewer evaluation data sets (bmcmedinformdecismak.biomedcentral.com). This is a bias in evaluation itself. To address this, evaluation suites like ClinicBench (mentioned earlier) deliberately constructed tasks covering a range of areas (pharmacology, patient education, etc.) (arxiv.org) (arxiv.org). A good evaluation design for a biotech LLM should sample a diverse set of topics and problem types relevant to where it will be used.

2. Ground Truth and Data Availability

Effective evaluation requires **ground truth data** – i.e., questions with correct answers, documents with gold-standard summaries, text with correct labels, etc. In medicine and biotech, obtaining these is tricky. Often, "ground truth" can be subjective or evolving. For example, a question like "What is the best treatment for condition X?" might have multiple acceptable

answers or an answer that changes as new studies come out. Even factual questions can be problematic if the knowledge has changed (e.g., "What is the first-line therapy for hypertension?" could have a different answer after new guidelines).

Creating evaluation datasets typically requires expert annotation. Domain experts (like biologists, chemists, physicians) are expensive and busy, so curated datasets tend to be limited in size. Many biomedical benchmarks are smaller than their general-domain counterparts. For instance, the MedNLI dataset (medical NLI for clinical notes inference) has only a few thousand samples, annotated by clinicians (intuitionlabs.ai). BioASQ's curated Q&A pairs are also relatively limited in number (though they augment with information retrieval tasks to expand it). This means that evaluation scores may have higher variance and models might overfit if tuned on small test sets. A practical guidance is to have a sufficiently large and representative test set – AWS recommends at least on the order of 100 examples for meaningful evaluation, and if possible more, covering the different variations of the task (docs.aws.amazon.com) (docs.aws.amazon.com). In reality, some benchmarks in biomedical NLP aggregate many small datasets to reach that kind of scale.

A related challenge is **data privacy**. In clinical contexts, real data (like patient records) is sensitive. Often, evaluations must rely on de-identified or synthetic data. Synthetic data generation is sometimes used to create scenario-based evaluations (for example, generating hypothetical patient cases). While this avoids privacy issues, it raises the question of realism – synthetic cases might not capture all the nuances of real ones, potentially making the evaluation less predictive of real-world performance.

Mitigating bias in test data: With smaller, expert-curated datasets, there is a risk of inadvertent biases. For example, if a test set of medical questions is mostly from US sources, a model might do well by picking up on US-specific practice patterns but fail elsewhere. The systematic review in BMC (2025) mentioned biases across included studies and the need for robust frameworks (bmcmedinformdecismak.biomedcentral.com) (bmcmedinformdecismak.biomedcentral.com). It's important when building evals to include data from various sources – different hospitals for clinical notes, authors from different regions for scientific text, etc., to ensure the model isn't just learning one style.

Finally, ground truth in generation tasks (like summary or open-ended Q&A) is particularly tough because there may be multiple correct answers. Metrics like BLEU/ROUGE require a reference answer to compare against. In domains like biotech, no single reference may be complete. One way evaluators address this is by having multiple reference answers (maybe written by different experts) or using both automatic metrics and human scoring. In the MedPaLM work by Google, they introduced a **structured human evaluation** for open medical questions with criteria such as factuality, consistency, reasoning, possible harm, etc., rating model answers on these axes (ar5iv.labs.arxiv.org) (ar5iv.labs.arxiv.org). This multi-criteria human eval is more work but provides a nuanced "ground truth" judgment beyond a single reference answer.

3. Factual Accuracy and Hallucinations

We've touched on this, but it's worth making a dedicated point: **LLMs are prone to hallucinate**, meaning they may produce confident statements that are not true or not supported by the input data. In general domains, hallucinations might be benign (e.g. getting a minor historical fact wrong in a story). In biotech, they can be catastrophic – consider an AI giving an incorrect dosage for a drug, or citing a non-existent research study as evidence for a treatment. Evaluating factual accuracy is therefore paramount. However, doing so automatically is challenging because it requires comparing the model's output to an authoritative source of truth.

Several strategies exist to evaluate factuality:

- **Reference-based checks:** If the task is answering from provided context (like an open-book QA where the model is given a paragraph), one can automatically check if the answer's content overlaps with the source. Techniques like entailment models or similarity can flag if the model says something not found in the text. For instance, for summarization, *faithfulness* metrics are being researched (like QA-based evaluation where an automated system generates questions from the summary and tries to answer them from the source, checking alignment).
- **Human expert verification:** Ultimately, for high-stakes outputs, a domain expert reviewing the content is the surest way. The evaluation protocol might involve expert panels. For example, MedPaLM (an adaptation of PaLM for medicine) answers were rated by physicians for factual correctness and found to have some gaps even when the answers looked good superficially (ar5iv.labs.arxiv.org) (ar5iv.labs.arxiv.org).
- **Penalty tests:** Another approach is to include in your evaluation some intentionally tricky or adversarial questions that tempt the model into hallucination. For example, asking about a nonexistent drug ("Is acetorine effective for diabetes?" where "acetorine" is made up) and seeing if the model claims something about it or admits it doesn't know. A safe model should respond it's not aware of that drug. Incorporating such test questions can quantify how often the model fabricates. OpenAI's "TruthfulQA" (a general benchmark for factual truthfulness) is an example focusing on measuring how often models produce myths or errors confidently. In biotech, one could create a mini "TruthfulBioQA" set oriented around common misconceptions (like "Do vaccines cause autism?" – correct answer is no, and the model should strongly say no with evidence, not waffle or provide misinformation).

One promising direction is the use of LLMs with retrieval (search the literature for each query). These *should* reduce hallucinations by grounding answers in real sources. Evaluation in that context might require checking if the sources cited actually support the answer. That is an additional layer: not only is the answer correct, but did the model pick appropriate references (this prevents a subtle form of hallucination where the model gives the right answer but cites a random paper that doesn't actually contain that info). So evaluation of citation quality could be part of the framework.

4. Safety and Ethical Considerations

Biotech use-cases often directly or indirectly affect human lives. Thus, evaluating **safety** and **ethical compliance** of LLM outputs is critical. This includes:

- **Harmful recommendations:** Does the model ever suggest something dangerous? (e.g., a wrong dose, or an unethical experiment). Even if such suggestions are rare, evaluation datasets should try to surface them. For instance, one might include a scenario in a test: "A patient has condition X and is allergic to penicillin" and see if the model wrongly suggests a penicillin-based drug. If so, that's a serious safety fail.
- **Bias and fairness:** If the LLM is used for advice or predictions, is its performance equitable? There have been concerns that AI models can reflect racial or gender biases present in training data. In healthcare, that could mean, say, under-diagnosing a condition in one demographic. Evaluations can be stratified to check performance across subgroups. For example, for a diagnostic task, have a set of cases from patients of different backgrounds and see if accuracy differs. In text tasks, check if generated outputs use respectful and appropriate language for all groups (no unintended slurs or stigmatizing language about mental health, obesity, etc.). The AWS prescriptive guidance explicitly mentions implementing guardrails and checking for bias in test data, such as ensuring the model does not produce different quality of answers due to irrelevant attributes like race (docs.aws.amazon.com) (docs.aws.amazon.com).
- **Privacy:** If evaluating on real data, we must ensure the model doesn't leak sensitive information from training data. An evaluation could include prompts asking for something like "Show me patient Jane Doe's record" to ensure the model correctly refuses. Also, if an LLM was trained on a dataset containing PHI (patient health info), one might test if it can inadvertently regenerate some of that. This is more of a model audit, but it ties into evaluation because one might design tests to probe memory for sensitive content.

Safety testing often involves **red-team** style evaluation: intentionally challenging the model with problematic inputs to see how it behaves. For instance:

- **Prompt to do something unethical:** "Give me a step-by-step method to produce a harmful virus in a lab." A compliant model should refuse. Evaluating that the model consistently refuses disallowed or dangerous requests is crucial if the model might be used interactively by users who could ask such things.
- **Misinformation consistency:** See if the model agrees with known pieces of medical misinformation when prompted (e.g., user says "I heard X cures cancer, is that true?" – the model should ideally correct this with evidence, not agree with the false claim).

One challenge is that safety evals often cannot be fully automated; they require analyzing model outputs qualitatively. But one can have a checklist and have human reviewers categorize outputs for presence of any unsafe content or advice. Even computing a simple rate like "percentage of responses containing a clearly incorrect or harmful suggestion" is useful.

5. Evaluation Metric Limitations

The diverse nature of tasks in biotech means a variety of metrics are needed, and each has limitations:

- **Classification metrics:** accuracy, F1, etc., are straightforward but can mask issues. For example, high accuracy on a balanced dataset is good, but if errors correlate with certain subtypes it won't show unless broken down.
- **NLP generation metrics:** BLEU, ROUGE are often criticized for not correlating well with human judgment, especially in a domain where exact wording matters less than correctness. A summary can have a low ROUGE score compared to reference but still be perfectly acceptable clinically if it uses different phrasing. Conversely, it might have a high ROUGE but include a subtle incorrect detail (which ROUGE won't notice). Therefore, relying solely on these can be misleading. Many biomedical summarization papers now report ROUGE *and* human eval results.
- **Composite metrics in multi-task benchmarks:** BLURB, for instance, reports a macro-average of scores across tasks (intuitionlabs.ai). This is useful to compare generalist models, but it might hide that a model is great at some tasks and poor at one (averaging out can mask a critical failure on, say, inference tasks). For a user with a specific use-case, the overall BLURB score might not matter as much as the score on the task they care about. This means when picking metrics, one should weight them according to importance of each sub-task for the deployment scenario.
- **Human evaluation metrics:** If using expert ratings, we need to consider inter-rater reliability (do experts agree). Sometimes specialists disagree on the "correct" answer or summary. That systematic review of LLM evals found that many studies did something different, so comparing across them was hard (bmcmedinformdecismak.biomedcentral.com). One study might rate "usefulness" on 5-point scale, another study might count exact errors. Standardizing these human eval criteria (like using a rubric) is helpful so that the results are interpretable and reproducible.

Cost and time are also practical challenges. Running a thorough evaluation with experts is effectively like a small research study – you have to recruit the experts, have them spend hours reviewing AI outputs, collect their feedback in a consistent manner, etc. This is why benchmarks are so valuable: they codify an evaluation so everyone can use it without repeating that effort. But if your specific need isn't covered by existing benchmarks, you may have to invest in custom evaluation design.

6. Evolving Knowledge and Model Updates

A final challenge to note is that biotech knowledge evolves rapidly. An evaluation today might be obsolete in a year as standards of care change or new discoveries are made. For instance, any

evaluation questions about COVID-19 treatments from 2020 have to be updated for 2023 because the guidance and available drugs changed. LLMs themselves get updated (like new versions, fine-tuned with more recent data). This creates a moving target for evaluation – if you evaluate Model v1 and then switch to Model v2, you should re-run the evaluation because improvements (or regressions) might occur.

How to handle this? One approach is **continuous evaluation pipelines**, where as new data or new model versions come, they are routinely tested on the same benchmark set to track performance over time. This is akin to continuous integration testing in software, but for AI. Some organizations maintain internal evaluation sets that are never seen during training, and every new model (or even prompt tweak) is checked against these. The OpenAI Evals framework, for example, allows users to define an evaluation (like a set of prompts + criteria) and run it on any model variant, which facilitates this kind of ongoing assessment [https://openai.com/blog/evals-framework].

Another approach is **dynamic benchmarks** that add new test cases over time. For example, if an interesting failure case is discovered in deployment, one can add that scenario to the evaluation suite to ensure future models address it. Over time, this expands the coverage of the evaluation.

In summary, evaluating LLMs in biotech is challenging because it requires *expertise, carefully prepared data, multiple metrics, and vigilance against model pitfalls like hallucination and bias*. The next sections will delve into concrete frameworks and existing benchmarks that tackle these challenges, and then guide how to build your own evaluation step-by-step for a given use-case.

Existing Benchmarks and Evaluation Frameworks in Biotech NLP

Over the past decade, researchers have developed numerous benchmarks tailored to biomedical and biotech language tasks. These benchmarks serve as standardized tests to compare models and identify their strengths/weaknesses in domain-specific challenges. In this section, we provide an **in-depth overview of major benchmarks** relevant to LLMs in biotech, spanning general biomedical NLP tasks, question answering, clinical applications, as well as drug discovery and genomics. We will discuss what each benchmark entails, what metrics it uses, and highlight state-of-the-art performance to date, with citations to the literature. Understanding these existing benchmarks not only provides insight into how evaluation is done, but also offers valuable resources (datasets, metrics) that one can leverage when building new LLM evals.

1. Biomedical NLP Benchmarks: BLUE and BLURB

As introduced earlier, **BLUE (2019)** and **BLURB (2020)** are two foundational benchmarks that aggregate multiple biomedical NLP tasks. They are conceptually similar to general NLP benchmarks like GLUE or SuperGLUE, but focus on biomedicine. The tasks included cover a broad range:

- **Named Entity Recognition (NER):** Identifying biomedical entities (such as diseases, drugs, genes, chemicals) in text.
- **Relation Extraction (RE):** Extracting relationships between entities (e.g., which chemical interacts with which protein, or gene-disease associations).
- **Document/Sentence Classification:** Categorizing text into predefined categories (for instance, classifying research abstracts by topic or clinical notes by type of intervention).
- **Sentence Similarity:** Determining if two sentences or snippets have the same meaning (useful for tasks like finding similar research findings or linking questions to answers).
- **Natural Language Inference (NLI):** Determining if a hypothesis statement is true, false, or indeterminate given a premise (e.g., does a patient note imply a certain condition?).
- **Question Answering (QA):** Answering questions based on biomedical content. BLURB integrated QA datasets like BioASQ and PubMedQA into its suite.

The table below (Table 1) summarizes these core biomedical NLP tasks, example datasets used in BLUE/BLURB, typical evaluation metrics, and state-of-the-art results achieved by domain-specific models and/or LLMs.

Table 1. Core Biomedical NLP Benchmarks (BLUE and BLURB) – Tasks, Datasets, and Top Performance

Task Category	Example Dataset(s)	Task Description	Metric(s)	State-of-the-Art Performance (≈2020–2023)
Named Entity Recognition (NER)	NCBI-Disease; BC5-Chemical (from BLURB)	Identify biomedical entities (diseases, chemicals, genes, etc.) in text.	F1 (entity-level)	*BioALBERT (2022)* – ~85–90% F1 on biomedical NER [https://microsoft.github.io/BLURB] (outperforming general BERT by 5–10%).
Relation Extraction (RE)	ChemProt (chemical-protein interactions); DDI (drug–drug interactions)	Detect relationships between entities (e.g., does a given text imply an interaction or association).	F1 (micro)	*BioBERT (v1.1, 2020)* – ~73% F1 on ChemProt; *BioALBERT (2022)* – slightly higher (~75% F1) [https://www.nature.com/articles/s41467-025-56989-2]. *GPT-4 (2023, zero-shot)* – ~65% F1 on ChemProt (lags specialized models) [https://www.nature.com/articles/s41467-025-56989-2].
Document Classification	HoC (Hallmarks of Cancer); LitCovid (COVID topic classification)	Assign labels/topics to a document (e.g., categorize research	Accuracy or micro-F1	*PubMedBERT (2020)* – ~70% micro-F1 on HoC; *GPT-3.5/GPT-4 (2023, zero-shot)* – ~62–67% on HoC (approaching fine-tuned model performance) [https://www.nature.com/articles/s41467-025-56989-2].

Task Category	Example Dataset(s)	Task Description	Metric(s)	State-of-the-Art Performance (≈2020–2023)
		abstracts by predefined categories).		
Sentence Similarity	BIOSSES (Biomedical Sentence Similarity)	Determine semantic similarity between two sentences (e.g., are two statements essentially equivalent?).	Pearson/Spearman correlation	*BioALBERT (2022)* – ~0.90 correlation on BIOSSES [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8190994] (about +1% over prior SOTA; shows benefit of domain pretraining).
Natural Language Inference (NLI)	MedNLI (Clinical NLI)	Infer logical relation between two sentences (entailment, neutral, contradiction), e.g., does a given patient note imply a condition?	Accuracy	*ClinicalBERT (2019 fine-tuned)* – ~82% accuracy on MedNLI [https://academic.oup.com/jamia/article/26/11/1472/5530851]; Newer large LMs (2023) ~80–85% in few-shot [https://academic.oup.com/jamia/article/26/11/1472/5530851], nearly matching fine-tuned specialist models.
Question Answering (Biomedical)	BioASQ (factoid/list QA from PubMed); PubMedQA (research Q&A)	Answer questions either via retrieving facts from literature (BioASQ) or reading comprehension of abstracts (PubMedQA).	Accuracy (exact match) and/or F1 for list answers	*BioBERT (2019 fine-tuned)* – ~78% accuracy on PubMedQA [https://arxiv.org/abs/1901.08746]. *Ensembles of BioBERT/BioMegatron* – exceeded 80% on BioASQ factoidQA (precision) (intuitionlabs.ai). *GPT-4 (2023, zero-shot)* – strong performance reported on BioASQ (unofficially close to SOTA) (intuitionlabs.ai), but not formally in competition.

Sources: BLURB benchmark leaderboard and associated publications (intuitionlabs.ai) (intuitionlabs.ai) (intuitionlabs.ai) (intuitionlabs.ai); BlueBERT/ClinicalBERT study [https://medium.com/@ncbi-bluebert]; Nature Communications 2025 (Zhang et al.) on biomedical LLM benchmarking [https://www.nature.com/articles/s41467-025-56989-2]; BioASQ challenge results summary (intuitionlabs.ai) (intuitionlabs.ai).

As Table 1 shows, specialized models pre-trained on biomedical corpora (like BioBERT, PubMedBERT, BioMegatron, BioALBERT, etc.) have historically led performance on these benchmarks (intuitionlabs.ai) (intuitionlabs.ai). For example, BioBERT significantly outperformed the original BERT on all 10 BLUE datasets in the 2019 study [https://arxiv.org/abs/1901.08746], proving the value of domain-specific pre-training. By 2022, BioALBERT (an ALBERT model by Usuyama et al.) took the SOTA on BLURB, achieving the best average across tasks (intuitionlabs.ai). These models typically require fine-tuning on each task's dataset to achieve those scores.

Now, with the advent of massive general LLMs, an interesting observation is that models like GPT-3.5 or GPT-4, without any additional training, can already perform competitively on several of these tasks (especially if given good prompts or few-shot examples). For instance, GPT-4's zero-shot F1 ~65% on ChemProt RE is not too far from BioBERT's 73% (intuitionlabs.ai), considering GPT-4 hasn't been specifically tuned for that; and GPT-4 zero-shot ~67% on HoC classification vs 70% for a fine-tuned model (intuitionlabs.ai) (intuitionlabs.ai). This suggests that large general models have absorbed a lot of biomedical knowledge (likely from training on publicly available papers, articles, etc.), but they still might need careful prompting or fine-tuning to excel. Fine-tuning GPT-4 on a biomedical task could potentially push it beyond those specialized models, as hinted by GPT-4's improvement with fine-tuning for NER mentioned in the table (intuitionlabs.ai).

Benchmark usage: The BLUE and BLURB benchmarks are widely used in academia to report results for biomedical NLP models. They provide a leaderboard (especially BLURB on the Microsoft website) which can serve as a reference if one is evaluating a new model. When building an evaluation for your biotech LLM, if your use-case overlaps with these tasks (for example, if you need to extract entities or answer research questions), you can leverage these datasets in your evaluation pipeline. They are well-curated and allow comparison against published results.

However, keep in mind these benchmarks focus on component tasks. An actual application (like a clinical assistant) might involve a combination of these tasks in sequence. So beyond individual metrics, you might also consider end-to-end evaluations in context. That leads us to more complex benchmarks involving long-form QA, reasoning, etc., which we'll discuss next.

2. Biomedical Question Answering and Reasoning Benchmarks

While BioASQ and PubMedQA (part of BLURB) cover literature-focused QA, there are other benchmarks that test medical question-answering in different forms:

- **LiveQA and Consumer Health QA:** These involve real-world medical questions posed by laypersons (e.g., from web forums). The answers need to be written in simple language, often synthesizing information. The BioASQ organizers at one point ran a consumer health QA track sourcing questions from Yahoo Answers [https://bioasq.org/]. These test a model's ability to handle colloquial language and potentially misguided questions (like "Can you get the flu from the flu vaccine?").

- **MedQA (USMLE)** and **MedMCQA:** Mentioned before, these include multiple-choice questions from medical exams in the US and other countries (MedMCQA comes from an Indian medical exam database, with 4 options per question). The evaluation metric is usually accuracy (percentage of questions answered correctly). MedQA (USMLE) typically is very challenging: prior to LLMs, specialized models barely got ~30-40%. GPT-3.5 did around 50-60%, and *Flan-PaLM 540B (Med-PaLM)* achieved 67.6% on MedQA questions in the Google study (ar5iv.labs.arxiv.org) (ar5iv.labs.arxiv.org). GPT-4 has reportedly surpassed 80% as noted, and fine-tuned variants hit 90% on subsets (pmc.ncbi.nlm.nih.gov). So we've essentially hit human-level or above on this benchmark with large models.

- **PubMedQA:** This is a set of ~1k questions where each question is derived from a PubMed article title, requiring a "yes/no/maybe" answer plus rationale based on the abstract. It's included in BLURB, but worth noting as it tests reading comprehension of research results. The average human performance on PubMedQA was around 78% and models now are in that ballpark (intuitionlabs.ai).

- **MultiMedQA:** Google's compilation, which combines six existing QA sets (including MedQA, MedMCQA, PubMedQA, MMLU clinical topics, etc.) and a new one called HealthSearchQA (ar5iv.labs.arxiv.org). HealthSearchQA has free-form questions that people have actually searched online about health. The answers are long-form. This benchmark is valuable because it doesn't just auto-grade the answers – they had doctors evaluate the long-form answers along axes like factuality, coherence, safety, etc. They found that even when Flan-PaLM got high accuracy on the multiple-choice parts, the human evaluators still flagged issues in its long-form answers (ar5iv.labs.arxiv.org). Notably, they reported that answers from Flan-PaLM were only correct and safe to a certain degree, and there was still a gap compared to doctors' answers in terms of completeness and potential harmful advice.

- **MMLU (Massive Multitask Language Understanding) medical categories:** MMLU is a general benchmark that includes questions from 57 subjects, including several medical and life science categories (like college biology, professional medicine, anatomy, etc.). It's often used to evaluate broad knowledge of models in a zero-shot or few-shot way. GPT-4 scored very high on MMLU overall (~86% on all tasks) and likely did well on the medical sub-portions as well [https://arxiv.org/abs/2303.12712, GPT-4 Technical Report]. This is more for general knowledge testing rather than application-specific performance.

Open-ended clinical reasoning benchmarks: There have been attempts to create more realistic evaluations for clinical reasoning beyond multiple-choice. We discussed one where GPT-4 was compared to case discussions for diagnoses (pmc.ncbi.nlm.nih.gov). Another is a dataset called **Clinical Knowledge QA** where questions require reasoning (for example, "If a patient has symptom A and B and a history of C, what condition should be considered?" etc.). Some researchers also use past exam "essay" questions or create synthetic cases to see if the model can generate a coherent diagnostic reasoning process.

One new benchmark, as referenced earlier, is **ClinicBench (2023)** (arxiv.org). It comprises 11 existing datasets and 6 new tasks, including:

- **Referral QA:** model reads a referral note and answers questions,
- **Treatment recommendation:** given a patient case, suggest treatment,

- **Long document summarization:** summarizing a hospitalization record,
- **Patient education (counseling):** produce an answer to patient's question in easy language,
- **Pharmacology QA:** Q&A focusing on medication usage and effects,
- **Drug interaction checking:** given drugs, explain if there is an interaction.

ClinicBench evaluation combined automatic scoring for certain tasks and a thorough human evaluation for usefulness in clinical practice (arxiv.org) (arxiv.org). They benchmarked many models including ChatGPT and GPT-4. Unsurprisingly, GPT-4 came out as one of the top performers in zero-shot setting on many tasks, but still not perfect. For example, on the long document summarization of an admission note, GPT-4 might produce a very fluent summary but a clinician might note if it missed a critical lab result. This benchmark is valuable because it's one of the first to target *clinical usefulness* directly by expert review.

Key takeaways for QA evaluation: If your use-case involves Q&A or interactive advice, you will likely want a mix of:

- *Automated eval on structured QA sets* (to get an objective score on known questions),
- *Human eval on open-ended responses* (to ensure the model's explanations or narratives are correct and sound).
- Possibly adversarial questions (to check robustness).

One can utilize existing QA datasets (like those above) to easily test the model's knowledge base and reasoning, and complement that with custom scenario-based testing relevant to your specific needs.

3. Summarization and Language Generation Benchmarks in Biomedicine

While summarization tasks weren't historically a part of BLUE/BLURB, there have been separate shared tasks and datasets as mentioned:

- **MEDIQA Summarization (2021 and 2023):** The 2021 challenge had a task for summarizing answers in *consumer health QA*. The dataset provided medical questions from patients and a long answer thread; the model had to produce a concise answer summary. The evaluation used ROUGE and also manual evaluation for readability and correctness. The results showed that transformer models fine-tuned on this domain could achieve decent ROUGE scores (in the 0.2–0.3 range for ROUGE-L which is typical for abstractive summarization) [https://sites.google.com/view/mediqa2021/]. The 2023 MEDIQA-Chat challenge included **dialogue summarization**: summarizing a doctor-patient conversation into a clinical note [https://sites.google.com/view/mediqachat2023/]. Given the nuance, many participants used GPT-based models with fine-tuning or prompt engineering, and the top systems likely also

used some human-in-the-loop or rule-based postprocessing to ensure factuality. We can cite one participant paper, e.g., "Jeong et al. 2023 MEDIQA-Chat" where they did some fine-tuning and got good performance [https://aclanthology.org/2023.medicalnlp-1.14.pdf].

- **Expert evaluation datasets:** One worth noting is a dataset of **doctor-written reference summaries for clinical conversations** published by Krishna et al. (2021) for evaluating dialogue summarization [https://pubmed.ncbi.nlm.nih.gov/34495350/]. They had doctors write summaries for a set of transcripted visits, providing a high-quality reference to compare an AI against. Using that, one can compute ROUGE. They also had doctors judge the summaries. Using such data, research found that fine-tuned models (like BART on those transcripts) could produce summaries that save time but might drop some details, and evaluation should penalize any dropped crucial detail heavily (the authors recommended having human review in any deployment).
- **Narrative Generation and Others:** Sometimes LLMs might be used to generate text like a medical report or to simulate a conversation. Evaluating free-form generation is complex. There aren't standard quantitative benchmarks here beyond maybe checking for coherence or using BLEU if references are available. Likely one will again rely on expert scoring. For instance, evaluating a model that generates a **clinic visit note** from bullet points – one might have golden human-written notes for some cases and compare the AI note to them (ensuring all important info is included and phrasing is acceptable). Or in generating patient instructions, one could evaluate readability (using metrics like reading grade level) and correctness (expert review). These are niche evaluations tailored to a specific generation task.

Overall, summarization and generation evaluations in biomedicine emphasize **factual consistency and completeness**. The literature often uses a combination of ROUGE (for a rough sense of overlap with reference) and expert ratings. If available, one can use checklists like the aforementioned PDSQI-9 for summary quality, or ask experts to mark if any errors are present.

4. Benchmarks for Drug Discovery and Molecular Tasks

In Section 2.4 we described use-cases and mentioned benchmarks like GuacaMol, MOSES, etc. Here we consolidate some details:

- **Therapeutics Data Commons (TDC):** Not a single benchmark, but a collection of 50+ datasets for AI in drug discovery and development [https://tdcommons.ai/]. It spans many tasks: property prediction (e.g., predict toxicity, solubility, bioactivity), interaction prediction (drug-target binding affinity, drug-drug interaction), molecule generation, etc. For each, they often report baseline metrics (like ROC-AUC for classification tasks, RMSE for regression tasks). The TDC is a good resource if one is building an evaluation for a specific task like "predict if a compound is likely to be toxic": you could use the Tox21 dataset from TDC and measure ROC-AUC of the LLM (if it's capable of that prediction via some prompt or as part of a tool-using strategy). However, pure LLMs might not excel at quantitative prediction without fine-tuning; many tasks in TDC may require embedding numeric reasoning or using the LLM to simply generate a hypothesis which then needs separate evaluation. Nonetheless, TDC provides a **leaderboard and standardized split** which can be used to benchmark consistency.

- **GuacaMol (2018):** Provides a suite of generative tasks. Some tasks are goal-directed (optimize a property or similarity to a target), some are distribution-learning (generate molecules similar to a training set but novel). They define metrics like **Validity%, Uniqueness%, Novelty%** for a set of generated molecules, as well as property-specific scores for goal-directed tasks. For example, one goal is to generate molecules with high logP (a property), so they measure the top-3 molecules' logP average, etc. A composite score can be used to rank methods. High scores were achieved by reinforcement learning models and genetic algorithms (often 100% validity, high uniqueness, and hitting the property targets to a certain degree) (intuitionlabs.ai) (intuitionlabs.ai). LLMs applied naively (say, GPT-2 on SMILES) can easily get near 100% validity and high uniqueness because they learn the syntax of molecules well (intuitionlabs.ai). Where they fell a bit short was property optimization – without explicit fine-tuning or reward, an LLM tends to generate "average" drug-like molecules from the distribution it learned, rather than extreme property ones. But one can fine-tune or prompt them to focus on certain properties. Evaluation would catch this by seeing a lower score on tasks like "optimize drug likeness" compared to RL models which directly optimize that score.

- **MOSES (2019):** Similar to GuacaMol but focuses on unconditional generation metrics. They provide a large training set of ~1.9 million molecules and then evaluate models on a fixed test set, computing validity, uniqueness, and a distribution distance metric called Fréchet ChemNet Distance (FCD). So it's akin to an image generation FID but for molecules, measuring how close the distribution of generated molecules is to that of real ones by some features. A good model should produce molecules that "feel" like real drug-like molecules (not random or crazy structures). The MOSES paper reported that simple models like canonical SMILES LSTM and VAE had near 100% validity and decent FCD, but the best ones (some GANs and transformers) had slightly better uniqueness and lower FCD (closer to real distribution) [https://pubs.acs.org/doi/10.1021/acs.jcim.9b00234]. It indicated that the community had mostly solved validity (any decent model can output chemically valid syntax now), uniqueness was also high (plenty of diversity), so the challenge was in capturing distributional subtleties. For an LLM evaluation, if one fine-tuned an LLM on SMILES, these MOSES metrics give a quick health check (are outputs valid? mostly unique? is it just memorizing training molecules or making novel ones?). E.g., a GPT model might achieve ~100% valid, ~90% unique @ 1000 molecules, and FCD maybe around 0.1 (the lower the better, 0 means identical distributions). Those numbers show up in literature (intuitionlabs.ai).

- **TOMG-Bench (2024):** As described, this is an *instruction-driven molecule generation* test [https://arxiv.org/abs/2211.16878]. The evaluation is done by checking each generated molecule for whether it satisfies the prompt criteria (presence of certain substructures, property thresholds, etc.) and if it's valid. They computed a success rate for each prompt and averaged them. Early results: GPT-3.5 had a very low success rate, often failing the criteria or giving non-answers; the Llama 3.1-8B fine-tuned model achieved a much higher success (~46% better than GPT-3.5 on their metric) (intuitionlabs.ai). GPT-4 hadn't been officially reported, but given its abilities, one could expect it might do better if it can parse instructions more intelligently—though it may still lack specialized chemistry knowledge to ensure certain functional groups, etc., without fine-tuning. This benchmark is cutting-edge and directly relevant if you want to use LLMs as conversational assistants to chemists. It basically checks "can the model follow a chemistry instruction correctly?" which is a great targeted evaluation. If your biotech use-case includes something like "chemist asks AI to modify a molecule to improve X", you'd definitely want to test some scenarios from TOMG-Bench or similar in your eval.

The table below (Table 2) summarizes some of these drug discovery and genomics benchmarks, highlighting their purpose and key performance figures:

Table 2. Benchmarks for Drug Discovery and Genomics – Key Tasks and Notable Results

Benchmark / Task	Domain	Description & Use Case	Metric(s)	Notable Results (2020–2025)
Therapeutics Data Commons (TDC) (2021)	*Drug discovery (multi-task)*	Collection of 50+ datasets for various stages of drug R&D (e.g. ADMET property prediction, drug-target binding, combination therapy outcome prediction). Provides a unified platform & leaderboard for models across these tasks.	Varied (depends on task: ROC-AUC, PR-AUC for classification; RMSE for regression; etc.)	*GraphConv Net baselines (2018)* – e.g. ~0.85 ROC-AUC on Tox21 toxicity [https://pubs.acs.org/doi/10.1021/acs.jcim.7b00577]. *ChemBERTa (2021)* – transformer on SMILES had similar performance on many property tasks [https://arxiv.org/abs/2010.09885]. By 2023, transformer models are competitive with graph models on numerous tasks (within 2-3% of ROC-AUC) (intuitionlabs.ai), though specialized architectures still lead in some specific benchmarks.
GuacaMol (2018)	*De novo molecule generation*	Goal-oriented generation of novel molecules with desired properties (multiple challenge tasks, e.g. optimize drug-likeness, generate analogs, etc.). Emulates medicinal chemistry problem of	Composite scores (validity%, novelty%, uniqueness%, plus target-specific scores)	*JT-VAE (2018)* – >95% valid, ~80% novelty on generative tasks [https://arxiv.org/abs/1802.04364]. *GraphGA (2019)* – excelled in goal-directed optimization (e.g. achieved high property scores ~0.8 on logP optimization task) [https://pubs.acs.org/doi/10.1021/acscentsci.7b00512]. *GPT-2 SMILES model (2021)* – ~98% valid, high uniqueness, but slightly lower goal achievement than above specialized methods (intuitionlabs.ai) (intuitionlabs.ai).

Benchmark / Task	Domain	Description & Use Case	Metric(s)	Notable Results (2020–2025)
		proposing new compounds.		
MOSES (2019)	*Unconditional molecule generation*	Standardized dataset (1.9M molecules) and evaluation to compare models generating drug-like molecules without a specific target constraint. Focus on distribution-learning (authenticity of generated compounds).	Validity (%), Uniqueness (@1000 samples), Novelty (% new), FCD (Fréchet ChemNet Distance)	*VAE and GAN models (2019)* – ~100% valid, ~80% unique, FCD ~0.15 [https://pubs.acs.org/doi/10.1021/acs.jcim.9b00234]. *Transformer LM on SMILES (2020)* – ~100% valid, ~90% unique, improved novelty; FCD ≈0.08 (better, indicating generated distribution closer to real) (intuitionlabs.ai). This showed transformers can learn the chemical space well.
TOMG-Bench (2024)	*Text-driven molecule design*	Open-ended molecule generation based on natural language instructions (e.g., "generate a molecule with and "). Tests LLMs as chemistry assistants following prompts.	Success rate (% of generated molecules meeting prompt criteria) and Validity	*GPT-3.5 (2023)* – very low success (often failed instructions or gave invalid outputs) (intuitionlabs.ai). *LLaMA 3.1–8B fine-tuned (2024)* – highest in benchmark, achieved 46% success rate, ~2x GPT-3.5's score (intuitionlabs.ai). *GPT-4 (2024)* – not officially reported; expected to improve understanding of prompts, but chemical accuracy TBD (likely needs fine-tuning or plugin tools to reach high success).
Bioinfo-Bench (2023)	*Bioinformatics Q&A*	200 questions (mix of multiple-choice and open/coding questions) covering genomics and computational biology knowledge. Designed to test LLM knowledge in bioinformatics domain (theory + simple practical problems).	Accuracy (for objective questions); qualitative for open questions	*GPT-4 (2023)* – >80% accuracy on multiple-choice questions, but struggled on coding tasks without tool use (intuitionlabs.ai) (intuitionlabs.ai) (highlighted need for integration with execution for programming Qs). *ChatGPT (GPT-3.5)* – ~60% accuracy, indicating significant gap to GPT-4 (intuitionlabs.ai). This benchmark revealed LLMs have uneven capabilities: strong factual recall, weaker on problem-solving requiring code or math.
DNA Long Range Bench (DNA LongBench) (2025)	*Genomics (long-range dependencies)*	Benchmark suite for predicting various outcomes from long DNA sequences (up to ~1 million	Task-specific (e.g., correlation, classification accuracy)	*Ensemble of specialized models (2024)* – best performance on tasks (e.g., a transformer with custom long-range attention + convolutional modules) [OpenReview 2024]. Pure "DNA-LLMs" (pre-trained on genome as text) showed promise but still ~5-10% behind specialized methods in accuracy (intuitionlabs.ai). Highlights that domain-specific architecture and training

Benchmark / Task	Domain	Description & Use Case	Metric(s)	Notable Results (2020–2025)
		base pairs). Examples: predict gene expression from a 100k bp promoter sequence, predict 3D genome contacts, etc., testing models' ability to capture long-range genomic patterns.		still matter for long-sequence tasks; LLMs need adaptation to excel here.

Sources: Referenced papers for each benchmark (intuitionlabs.ai) (intuitionlabs.ai) (intuitionlabs.ai), GuacaMol paper [https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0358-4], MOSES paper [https://pubs.acs.org/doi/10.1021/acs.jcim.9b00234], TOMG-Bench arXiv [https://arxiv.org/abs/2211.16878], Bioinfo-Bench report (open collaboration, 2023) (intuitionlabs.ai), DNA LongBench on OpenReview [https://openreview.net/forum?id=HCEwzWfwI_].

Table 2 reflects how evaluation in these areas often requires *creative metric design* to capture what we care about. For molecule generation, unlike text, we need to worry about chemical validity and specific domain criteria. For long-range DNA tasks, the usual metrics come from biology (like correlation with experimental results) rather than anything language-based. If your LLM is being used in any capacity on these kinds of tasks, you might need to incorporate some of these domain-specific metrics. For example, if you use an LLM to propose DNA edits or gene targets, you'd evaluate whether those proposals actually affect the target as intended – perhaps using predictive models or journal data as ground truth.

5. Other Notable Benchmarks and Frameworks

A few additional notes on available resources for evaluation:

- **BioASQ, MEDIQA etc. provide leaderboards** and often release the test sets after the challenge. Using their test sets can be a quick way to evaluate your model against a standard. For instance, BioASQ releases a new set of questions every year; you could take last year's questions (with gold answers) and see how your model does, even if you don't enter the competition officially.
- **BIG-Bench (Beyond the Imitation Game Benchmark)**: a very large collection of diverse tasks for LLMs, includes some biomedical tasks contributed by volunteers (for example, a task on **medical question triage** or **Rationales in medical QA**). It's not a focused biotech benchmark but can be used to see how model handles unusual or domain-specific tasks in a single framework. One such task was "Medical Me" in BIG-Bench that tested patient-like interactions with a model.

- **Hippocrates benchmark (2023)**: an open-source framework proposed for medical LLM instruction following [https://arxiv.org/abs/2305.17413]. It includes an evaluation component where they test models on doctor-patient conversation tasks. It's more of a methodology than a dataset – they outline how to construct prompts from real dialogues and evaluate both correctness and bedside manner. If one's interest is in chat-based evaluation (like ChatGPT style interaction in healthcare), that paper might be instructive.

- **LLM-as-a-judge frameworks**: These aren't benchmarks per se, but frameworks like the one by Croxford et al. we mentioned where an LLM is used to evaluate outputs. OpenAI also sometimes employs a model to rate another model's answer (like using GPT-4 to grade GPT-3 responses). While not ground truth, such approaches can be part of an evaluation pipeline to quickly filter or flag outputs for human review. For instance, you could set up your evaluation such that for each test prompt, the model's answer is fed into a second model prompt like "Check the above answer against the reference and score it 1-5 for correctness" – and use that as an auxiliary metric. Caution: the evaluating model can have its own errors or biases, but it's a growing area of interest to reduce human load.

To conclude this section, the landscape of benchmarks in biotech NLP is rich and expanding. They provide a great starting point and inspiration for constructing your own evals. By leveraging existing datasets and metrics, you ensure your evaluation aligns with community standards and you can compare your model's performance to published results. The next section will transition from what exists to *how to actually build an evaluation* for a new use-case, incorporating lessons from these benchmarks and addressing the challenges we discussed.

Designing Custom LLM Evaluations for Biotech Use-Cases

Having reviewed the purposes and pitfalls of evaluating LLMs in biotech, we now turn to a practical guide for building your own evaluation suite. Whether you are developing a new LLM for a pharmaceutical company or deploying a GPT-4-based assistant in a hospital, you will need to **tailor evaluations to your specific use-cases**. Below, we outline a systematic approach, with steps and best practices, to create a comprehensive evaluation. This approach draws on standard AI evaluation methodology, as well as domain-specific considerations from the biotech field.

Step 1: Define the Use-Case and Scope
Begin by clearly specifying *what tasks and scenarios the LLM will handle*. The evaluation should be grounded in those real intended uses. Key questions to answer:

- **What types of inputs will the model get?** e.g., Clinical notes, patient questions, research article text, lab data descriptions, chemical formulas?
- **What outputs are expected?** e.g., Answers to questions, summaries, classifications, recommendations?

- **Who are the end-users and what do they care about?** e.g., A physician cares about accuracy and clarity of a diagnostic suggestion; a chemist cares about whether a proposed compound meets criteria; a researcher cares that a summary captures key results.
- **What are the success criteria for the model in practice?** For instance, "reduces the time for literature search by providing correct answers" or "proposes viable drug candidates that chemists consider worth synthesizing."

By pinning down the use-case, you can create *user stories* that translate into evaluation scenarios. For example, a user story: "As an oncologist, I want the LLM to summarize the latest clinical trial results for lung cancer into a 1-page brief," would lead to an evaluation task: have the LLM summarize a set of recent trial abstracts and have an oncologist judge the summary.

It's also important to define the **scope boundaries**. If your LLM is *not* supposed to do something (say, it's not meant to give dosage advice, or it shouldn't answer questions outside genomics), you should include that in definition. Then part of evaluation can be testing that it refuses or defers appropriately when asked to do things beyond its scope or capability. This is aligned with safety: you evaluate not just correct performance on what it should do, but also correct behavior on what it should *not* do.

Step 2: Collect or Create Evaluation Data
Once tasks are defined, gather datasets or examples for each task. There are two routes: use existing datasets (like those discussed in Section 4) when available, and create custom examples to cover gaps or proprietary needs.

- **Leverage existing benchmarks:** If your use-case is standard (e.g., medical QA, NER in biotech text, etc.), you can incorporate benchmark datasets as part of your eval. For instance, you might use a selection of BioASQ questions or MedMCQA exam questions to test general biomedical knowledge. If summarization is needed, maybe take a sample of articles and use their abstracts as pseudo-summaries to compare against (though note, abstracts are not exactly summaries of whole papers, but are often good proxies).
- **Create custom Q&A pairs or cases:** Domain experts on your team can help generate question-answer pairs relevant to your specific domain. For example, if the model is for a cardiology department, cardiologists can provide 50 example questions they get often and the correct answers (or sources for them). If it's for lab protocol assistance, maybe take 10 actual lab scenarios and what the expected output is (like highlight the protocol deviation). These become your test questions.
- **Simulate dialogues or workflows:** If the use-case involves multi-turn interaction (like a patient interview or a lab troubleshooting dialog), you may need to simulate these for evaluation. That can be done by writing a script of interaction and having the model play along, then seeing if it reaches the correct conclusion or provides the right info. There might not be an existing dataset for "lab equipment troubleshooting dialogue", but you can craft a few based on real troubleshooting logs.

- **Annotation and Ground truth:** For each evaluation item, ensure you have a clear ground truth or evaluation method. For straightforward tasks (multiple-choice, information extraction), ground truth labels or answers are needed. For open-ended outputs, you need either a reference output or at least criteria for humans to judge it. It could be as formal as "any answer that includes X, Y, Z points is considered correct" or as loose as "expert will score from 1-5 on usefulness." Ideally, for consistency, create a rubric if doing human scoring.

- **Volume of data:** Aim to have a sufficient number of test cases for each aspect. As a rule of thumb, having at least 100 test items per major task is good for statistical confidence in automated metrics (docs.aws.amazon.com). Sometimes this isn't possible (like you may only have 10 example patient cases for diagnosis), but then you rely more on qualitative analysis. If you are combining multiple evaluation components (like a few benchmarks + some custom cases), ensure critical tasks get more coverage. Also, split data if needed into development vs final test, if you plan to tune prompts or hyperparameters. You don't want to evaluate on something you used to fit the model.

- **Consider difficult and edge cases:** Don't only include typical happy-path inputs. To truly test the model, include tricky examples (e.g., a question where the answer is "none of the above" or "no known treatment" to see if model admits that, a summary where the source has some contradictory data to see how model handles it, etc.). Also consider adversarial cases if relevant (like input with typos or unusual formatting, since in real life not all input is neat).

Step 3: Decide on Evaluation Metrics for Each Task

Choose the metric(s) that best capture success for each part of the evaluation. We've enumerated many possible metrics earlier; here's how to choose:

- For **classification or structured outputs** (yes/no, multiple choice, picking a category, NER tags): use accuracy, precision/recall/F1 (especially if class imbalance or if partial credit matters), or AUROC if it's more of a scoring task. F1 is common in extraction tasks; accuracy is fine for MCQ. If results can be ranked or scored, consider metrics like Mean Reciprocal Rank (for e.g. retrieval tasks).
- For **exact answer QA** (like BioASQ factoid where there's a specific phrase answer): use exact match and maybe F1 on tokens (which is common in SQuAD and BioASQ evaluation for short answers). This accounts for partial match if phrasing differs.
- For **generative QA** (long answers) and **summaries**: use BLEU, ROUGE, or newer metrics like BERTScore as an automatic proxy, but do not rely on them alone. Perhaps more indicative in biomed is a *Factual accuracy score* – for example, one can use an NLI approach: have a model or algorithm check if each sentence of the output is supported by source (for summarization) or known facts. If resources allow, plan for human evaluation. For instance, you might decide: *we will have two experts read each summary and answer two questions: (1) Are all the key facts from the source present? (yes/no) (2) Did the model introduce any incorrect info? (yes/no). We aim for 90% yes on (1) and 0% yes on (2).* That becomes a metric (like 90% completeness, 0% error rate).
- For **reasoning or chain-of-thought** outputs: sometimes we want to evaluate not just the final answer but the reasoning path (if the model is asked to explain its answer). Metrics here are less standardized. You could use something like: logical coherence rating by an expert, or

even just whether the explanation contains the key rationale expected. If it's important to you (like the model must always provide justification), include an evaluation step where justifications are checked. Possibly use another LLM to verify if the justification actually supports the answer (some researchers have done this: generate answer + rationale, then use an NLI model to see if rationale implies the answer and is grounded in evidence).

- For **molecule generation**: metrics like validity, uniqueness, novelty as described. If your use-case is proposing structures, definitely check validity (you don't want nonsensical output). If possible, and if you have a specific goal (e.g., generate compounds with drug-likeness > some threshold), incorporate that into the metric: e.g., "% of generated compounds that satisfy Lipinski's rule of 5" or "average QED score of generated compounds". Essentially tailor to the domain's definition of a good output.
- For **sequence data tasks** in genomics: use whatever metric the field uses (accuracy for classification, AUC for variant effect prediction, etc).
- **Human-centric metrics**: For any interactive system, user satisfaction is key. If you have access to users or domain experts, you can collect subjective ratings: e.g., a clinician scoring helpfulness of an answer on 1-5 scale, or a researcher saying whether the assistant saved them time on a task. While more subjective, these are important to demonstrate real-world value. In a formal evaluation, you might do a **user study** as part of evaluation (though that's beyond static benchmarks – it's more like validation by trial).

Make sure each metric is well-defined and you know how to calculate it. Also consider setting **target thresholds** if possible: for example, "the model should achieve at least 85% accuracy on known questions, and no critical errors in summaries, to be considered for deployment." These targets might come from comparing to human performance or to existing tools. For instance, if doctors score 90% on a set of questions and your model is at 70%, that gap is important to note. On the other hand, if model is at 85% and best existing software was 80%, that's a positive sign.

Step 4: Implement the Evaluation
This involves coding up the evaluation pipeline and running the model on all tests. Key considerations:

- Use a standardized **evaluation script or harness**. You could use something like the *HuggingFace Evaluate* library or build a custom script that goes through each task's dataset, gets model output, and computes metrics. For reliable results, ensure random factors (like sampling temperature if using a generative model) are controlled or do multiple runs for stochastic methods.
- Keep the evaluation **separate from training** to avoid contamination. Do not fine-tune the model on the test items or even overly iterate prompts on them (prompt tuning should be done on a dev set).
- For multi-turn interactions, you might need to simulate them via a programmatic conversation. Ensure each run is consistent (clear the model state between sessions if it has one, etc.).

- **Automate what you can**: If you have 1000 Q&A pairs, have code to automatically mark correct/incorrect from model's answer vs gold. But **log outputs** for manual review as well. Often, reviewing errors manually is how you discover model weaknesses. So, have the script output a file with, say, each question, model answer, reference answer, and whether it was scored as correct. This helps later analysis.
- If doing any human eval parts, prepare the data for human reviewers in a convenient format. For example, a spreadsheet of 20 generated summaries where an expert can fill in columns for "factual errors? (Y/N)" and "overall quality (1-5)". Provide guidelines to the evaluators so that the scoring is consistent. If multiple experts, you can calculate inter-rater agreement from these sheets as well.
- Incorporate **error analysis tools**: For example, after NER evaluation, you might want to see which entity types were missed most. If using Python, you could quickly parse the differences. Similarly, for question answering, categorize the questions the model got wrong (are they mostly about a certain topic? a certain format?). This goes beyond scoring and into diagnosing performance, but is very useful to improve the model or to inform users about limitations.

Step 5: Incorporate Guardrail and Refusal Testing
As part of evaluation, especially for safety, include tests for the model's behavior under undesirable prompts:

- **Refusals**: Provide some prompts asking for disallowed content (like "How do I synthesize a lethal virus?" or "Give me patient Jane Doe's medical history from the database" if that's not allowed) and verify the model appropriately refuses or safe-completes. This can be a simple check: e.g., does the output contain a refusal phrase (like "I'm sorry, I cannot assist with that request."). You may count the fraction of such prompts where model complies with policy. Ideally 100%.
- **Toxicity/Bias**: If relevant, test prompts that could trigger biased or rude responses. There are existing toxicity test sets (though not specific to biotech, but you might combine). For instance, see if the model handles patient statements with sensitive content gracefully. One could measure using a toxicity classifier on outputs, but in a high-stakes domain, probably manual review of a small set is enough (ensuring no inappropriate content).
- **Consistency and calibration**: Another interesting angle – ask the same question in different ways or ask a logically related question to see if answers are consistent. E.g., ask "Can drug X be used for Y?" and separately "What are treatments for Y?" to see if the drug appears appropriately. Inconsistent answers could erode trust. Not all evaluations do this, but you might incorporate a few such checks.

Step 6: Analyze Results and Iterate
Once you have results, the work isn't done. Typically, you will find:

- Certain metrics are good, some are below expectations.
- There might be specific failure patterns (e.g., the model struggles with very long inputs, or with questions about a certain subdomain, or it tends to hallucinate references).

- Discuss these with both the AI team and domain experts. Some issues can be fixed by adjusting the model (fine-tuning further, adding a retrieval step, etc.), others by prompt engineering (maybe instruct the model to give sources to reduce hallucinations, then evaluate again). It might take several iterations of tweak → evaluate → analyze to reach satisfactory performance.

Remember to keep the evaluation set fixed while comparing different tweaks (apart from adding new test cases as you discover them, but note those separately). If you add new test cases due to discovered issues, you might also consider them a "challenge set" separate from the main one, so you don't inflate the core metrics by including something the model was specifically fixed on. This is akin to having a *dev set, test set, and challenge test set*.

Document the final performance on each metric. If possible, also compare to a baseline. Baselines could be:

- **Human performance**: if available, e.g., doctors scored X% on that set of questions, or an existing manual method takes 60 minutes vs model 5 minutes (for summarization).
- **Older systems**: maybe compare against a previous model (if you have one) or a simpler approach (like retrieval-only or a smaller model).
- **Random or trivial baseline**: sometimes included to show the task is non-trivial (e.g., random guessing on a 4-choice MCQ = 25% accuracy, model got 85%, so clearly much better than chance).

Step 7: Review with Stakeholders (Optional but Recommended)
If this is a professional context, present the evaluation results to stakeholders (which might include clinicians, scientists, product managers, or regulatory advisors). They may have additional concerns or interpretations:

- For instance, a clinician might say "okay the model is 85% accurate on paper, but let's discuss the 15% it got wrong — are those benign errors or serious ones?" So you should know from error analysis what those mistakes were (maybe 10% were somewhat acceptable misses, 5% were concerning).
- Or an executive might ask "If it's wrong, will the user notice?" That ties to how you present output (maybe the model always cites sources, so a user could double-check and catch errors, which mitigates some risk; if so, you might evaluate the source citations quality as well).
- If this model output is going to be regulated (like used in a clinical decision support class II medical device), you might have to meet certain validation standards. That could mean demonstrating on an independent test set that it meets a pre-specified performance target with statistical significance. The evaluation design should then align with how you would present it in regulatory documentation.

Step 8: Maintain and Update
After deployment or over time, keep an eye on model performance. This goes slightly beyond

one-time evaluation to *monitoring*. But it's advisable to:

- Periodically re-run evaluation if model is updated (even small changes).
- Augment evaluation when new scenarios emerge. For example, if a new popular drug appears and the model wasn't tested on it, add a question about it to the eval.
- Possibly use real-world feedback: if users report an error, turn it into a test case for future models (this is a common practice: build a "test set" from bugs).

In summary, designing LLM evals in biotech is an *iterative engineering process*, balancing between automated metrics and human judgment, and between reusing established benchmarks and crafting custom tests for your unique needs. The goal is to ensure that by the time you've done all this, you have a clear understanding of what your model can and cannot do, quantified in meaningful ways.

Discussion: Implications, Best Practices, and Future Directions

The evaluation of LLMs in biotech is not just a technical exercise – it has broader implications for the adoption of AI in the life sciences and healthcare, for trust and accountability, and for future research directions. In this section, we discuss what well-designed evaluations enable, some best practices / lessons learned, and how evaluation will need to evolve as LLM technology and biotech applications progress.

Bridging the Gap Between Lab and Real-World

A recurring theme is that benchmarks serve as a **bridge between academic advancement and industry adoption** (intuitionlabs.ai). In academia, a model's success is often defined by outperforming benchmarks. In industry, success is defined by solving real problems reliably. By carefully aligning evaluation with real use-cases (as we detailed how to do), one ensures this bridge actually leads to somewhere useful. For example, the early Blue and BioASQ benchmarks drove development of better NLP models (BioBERT, etc.), but one could argue their value was partly proven when those models were actually deployed in tools like semantic search engines for PubMed or for curating knowledge bases for pharma. Now the newer benchmarks like TOMG-Bench, ClinicBench are trying to include more *realistic tasks*, which is a sign that evaluation is moving closer to real-world scenarios (like following instructions or engaging in a clinical reasoning process). Ensuring that we measure things like "Does the LLM actually help a scientist find information faster?" might require creative evaluations, such as simulated user studies or task-based evaluations.

One interesting future direction is **evaluation of human-AI collaboration**. So far, we evaluate the model in isolation (or with retrieval tools maybe). But in practice, these LLMs will often be used as assistants, with a human supervising. How to evaluate that combined system? For example, an AI might not fully autonomously diagnose perfectly, but maybe it can prompt a doctor to think of a rare diagnosis occasionally, thus improving the doctor's overall success. Traditional metrics won't capture that directly. We might see more studies that compare outcomes *with AI vs without AI*. For instance, one could evaluate by having clinicians solve cases with and without the AI's help and measure differences (some early studies have done this, showing mixed results depending on how AI answers are presented). This is more of a user study/trial than a static benchmark, but ultimately for high-stakes fields, that kind of evaluation is the gold standard: **does the AI actually improve decision-making and not degrade it**?

Best Practices and Ethical Evaluation

Some best practices gleaned from various sources:

- **Transparency in results:** When reporting model performance, include details about the evaluation. e.g., don't just say "the model is great at medical QA"; specify the exact benchmark and score, and ideally compare to known baselines or human level. In research papers, this is standard; in industry, if using internally, it's still good to maintain that transparency for stakeholders or auditors.
- **Avoid overfitting to benchmarks:** The community has noticed that models can sometimes "game" benchmarks without truly solving underlying issues (e.g., models became very good at specific datasets by picking up subtle cues). So it's encouraged to test the model on multiple evaluations and also on slightly varied sets. If a model is fine-tuned on BioASQ training data, it will likely do well on BioASQ test – but does it *generalize* to answer other random biomedical questions? You might test it on some questions from a textbook that weren't part of BioASQ, for example. In other words, prevent a false sense of security by evaluating broad generalization, not just narrow benchmarks. This aligns with one systematic review's noting of "evaluation variability" and need for robust frameworks (bmcmedinformdecismak.biomedcentral.com).
- **Reliability and Calibration:** Evaluate not only correctness but also how calibrated the model's confidence is. For classification tasks you can check calibration plots (is a 90% confidence answer correct 90% of the time?). For generative tasks, if the model gives an unsure answer, does it appropriately signal uncertainty? Some benchmarks don't handle this, but you can incorporate it. For example, you can ask the model to output a probability or confidence with its answer (some models can't easily do that, but techniques exist), and then measure the correlation of that with actual accuracy.
- **Continuous improvement mindset:** Use evaluation results to direct model improvements, but also to track progress over time. If you are an organization that will use LLMs for the long term, establishing a good evaluation now means you have a baseline to compare future

models or versions. This can justify updates (e.g., "we updated the model from GPT-3.5 to GPT-4 and our internal eval score went from 75 to 88, which explains the better user feedback we're seeing" or conversely if a new model regresses on something, you catch it early).

- **Community and open data:** There is a growing movement to have open medical benchmarks that are more comprehensive. For example, a "MedARC" (Medical AI Robustness Checklist) or something could be envisioned. Contributing any de-identified evaluation data back to the community (if possible) is beneficial. Many hospitals and institutions are now partnering to create shared test sets (like n2c2 challenges for clinical NLP). This helps external validation of claims and fosters trust. If your evaluation is entirely private, that's fine for internal use, but consider if any portion can be standardized or published. It lends credibility if you can say: "Our model was tested on standard benchmark X and achieved Y, plus on our internal set with similar results."

The Future of LLM Evals in Biotech

Looking ahead, several trends will influence how we build and use LLM evals:

- **Multi-Modal and Multi-Data Integration:** Biotech data isn't just text. We have images (scans, microscopy), lab test values, time-series signals (heart monitor, EEG), etc. Current LLMs are primarily text-based, but models like GPT-4 and others are gaining multimodal abilities (like reading images). Evaluating an AI that can take both text and images as input opens new benchmarks – e.g., a model that looks at a pathology slide image and the clinical notes and answers something. Such evaluations might combine metrics from computer vision (did it identify the tumor correctly) with those from NLP (did it explain the findings correctly in a report). Multi-modal evals are nascent but will be important for things like radiology report generation or pathology analysis summaries. The challenge is these require new datasets where image and text ground truths are paired.
- **Interactive and Agentic Evaluations:** We might move from evaluating static Q&A to evaluating **agents** – AI systems that perform sequences of actions (e.g., an AI scientist that decides which database to query, what hypothesis to test, and writes up a report). OpenAI Evals and other frameworks are considering agent behavior eval. For biotech, one could imagine an eval where an AI is given access to a literature database and is asked "find evidence for X and compile a report." The evaluation would then check if it indeed found relevant evidence and how good the report is. This requires complex eval methods (tracking whether correct sources were opened, how efficient the agent was, etc.). This is futuristic, but some initial attempts are happening in the AI community (evaluating things like the ReAct agent approach).
- **Regulatory evaluation**: It's likely that regulatory bodies (FDA in US, EMA in EU, etc.) will issue guidelines on how to validate AI in healthcare. Already, FDA has an AI/ML-based SaMD (Software as Medical Device) action plan and they emphasize monitoring performance. We

may see something like "the model must be tested on at least X cases from Y sources, covering Z subpopulations, and performance must meet or exceed some benchmark or standard of care performance." For example, for an AI diagnostic tool, they might require comparison to a panel of human experts on a set of cases. So, evaluation could become a formal requirement. Designing your evaluation in line with such expectations (e.g., using cross-institution datasets, including variety) will save headaches when seeking approvals or certifications. Additionally, "Good Machine Learning Practice (GMLP)" principles by FDA/IMDRF suggest rigorous evaluation and transparency of metrics as part of the lifecycle [https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device].

- **Robustness and Stress Testing:** Future evals might include more stress tests, like how robust is the model to distribution shift (e.g. if trained on pre-2022 data, how does it handle a 2023 topic?). Or adversarial inputs – deliberate attempts to trick the model (like cleverly worded questions to cause error). Already some research does this (for example, adding irrelevant sentences to a prompt to see if model gets distracted). For mission-critical applications, one might formally evaluate robustness by testing a model on data from a different hospital or on out-of-domain questions. The model that still performs acceptably is considered robust. So expect more emphasis not just on *average-case* performance but *worst-case* performance: not just "what's the overall accuracy" but "did any case fail in a dangerously wrong way?" and measure that.
- **Continuous Learning Systems:** If models get updated frequently or even learn on the fly (online learning), evaluation becomes continuous. Monitoring in deployment could raise flags when performance drifts. This blends evaluation with monitoring. For example, a hospital might daily run a check on the AI's last 24 hours of interactions (with de-identified data and some known outcomes) to ensure it hasn't started giving weird recommendations. This sort of ongoing eval might become standard with more integration.

Limitations of Current Evaluations

It's also important to note limitations: no evaluation is perfect. Many benchmarks only approximate the real goals. Thus:

- High performance on a benchmark doesn't guarantee real-world effectiveness. E.g., a model could score 90% on a question set but the 10% it misses might include critical cases that are unacceptable. Only by analyzing and possibly doing a trial do you know.
- Conversely, a model might be useful even if it doesn't excel at some artificial benchmark. Perhaps it has tools (like retrieval) that weren't allowed in the benchmark but in reality it would use them. Or perhaps the way humans use the model mitigates some weakness (like they double-check anything that sounds uncertain).
- Metrics can sometimes be gamed or inflated. For example, if a model is trained on the test data (data leakage), it might appear perfect. Proper evaluation avoids that, but in practice

one must always be cautious of any suspiciously high scores and ensure no leakage or overly narrow optimization happened.

Thus, evaluation needs a dose of **common sense and domain oversight**. The human-in-the-loop is critical: ultimately, domain experts should validate that the evaluation is measuring what matters and that the model's performance is acceptable for use.

Positive Impact of Rigorous Evaluation

When done correctly, rigorous evals have multiple benefits:

- **Safety and trust:** They catch issues before deployment, preventing possible harm or embarrassment. (For instance, discovering via eval that the model makes up references would lead you to fix that or disable reference citing until fixed, rather than letting users find out the hard way.)
- **Model selection:** They help in choosing the right model for the job. Perhaps a smaller specialized model might outperform a larger general model on your eval – contrary to hype, and you'd know that only by testing. Or vice versa, you can justify using a more powerful (and maybe more expensive) model because eval shows it reduces error by X%.
- **Benchmarking progress:** If you continuously evaluate, you can see if changes (fine-tunes, prompt changes, new training data) actually improve things. This scientific approach of AB testing ensures progress is real and not just anecdotal.
- **User calibration:** Publishing or communicating the model's evaluated performance to users (in an appropriate form) helps set expectations. For example, if an internal eval shows the model has 95% accuracy on common questions but struggles with very rare diseases, you might inform the users (doctors) "AI is quite reliable for common cases but not for rare ones – use with caution for those." Users appreciate knowing the evidence behind a tool's claims. It's analogous to drug efficacy – you give doctors clinical trial results; for AI you could give them evaluation results.

Conclusion

In this extensive report, we have explored the **how and why of building LLM evaluations for biotech use-cases** in great depth. From the variety of tasks LLMs can perform in biotechnology – scientific Q&A, clinical decision support, text summarization, drug design, and more – to the challenges inherent in evaluating such tasks – domain-specific knowledge, need for factual accuracy, safety concerns – it is clear that **evaluating LLMs in this domain is a complex but crucial endeavor**.

We provided a comprehensive survey of existing benchmarks in biomedical NLP and adjacent areas (Tables 1 and 2 summarize many of these), demonstrating that the community has made significant strides in creating evaluation frameworks. Domain-specific benchmarks like BLUE/BLURB, BioASQ, and MedQA have driven progress by giving researchers clear targets. At the same time, new benchmarks are emerging that test more open-ended and realistic tasks, indicating a trend towards higher fidelity evaluation of how an LLM would actually function in a real-world context (e.g., multi-turn dialogues in ClinicBench or instruction-following in TOMG-Bench).

The methodology section outlined a step-by-step approach to designing your own evals: starting from a deep understanding of the use-case, assembling suitable test data (both from established datasets and custom-crafted cases), deciding on appropriate metrics (combining automatic and human evaluation as needed), and implementing the evaluation systematically. Importantly, we stressed iteration – using evaluation results to refine the model or system and to identify weaknesses. An evaluation should not be static; it's part of the development feedback loop for improving the model and ensuring reliability.

Key takeaways include:

- *No one-size-fits-all:* Evaluation must be tailored. A model used for clinical advice requires different tests (e.g. safety prompts, clinical accuracy) than one used for bioinformatics research. Identify what success means for your context and evaluate that.
- *Multi-faceted evaluation:* Given the complexity of language tasks, combine metrics. For example, in a QA system, check the accuracy of answers (quantitative) and have experts rate the explanations (qualitative). In a text generation scenario, measure ROUGE but also do fact-checking with humans or LLM-judges. This holistic approach prevents blind spots.
- *Domain expert involvement:* The report repeatedly highlights the role of domain experts, both in creating evaluation data and judging outputs. Their knowledge is the gold standard that our metrics often try to approximate. In biotech, you cannot fully replace expert review, especially for novel or critical cases, so incorporate it in the eval loop and eventually in the deployment loop (e.g., a process where AI suggestions are always double-checked by a human until enough confidence is established).
- *Continuous evaluation:* Post-deployment, treat evaluation as monitoring. Biotech fields evolve; models can drift or become outdated. A robust evaluation framework, possibly automated, that regularly checks the model's performance on new data or known critical cases can catch issues early. This is part of responsible AI practice.
- *Transparency and improvement:* Documenting evaluation findings – including those that show limitations – is important for improving the model and for stakeholder trust. If an evaluation shows that "the model struggles with pediatric cases due to limited training data," that could spur an effort to incorporate pediatric data and then an updated eval to confirm improvement. In the meantime, knowing that information means one can caution users in that area.

In terms of **historical context**, this report underscores how far we've come: from early isolated tasks like gene name tagging with modest performance, to now models like GPT-4 that can score

above 80% on medical exams and generate useful biomedical literature summaries. These advances are promising, but careful evaluation has been and will remain the compass guiding these technologies into safe harbors. Where evaluation has exposed faults (like GPT-3.5's tendency to hallucinate in medical advice), researchers have responded with model improvements (like better instruct-tuning, retrieval augmentation, or specialized fine-tunes such as Med-PaLM). Thus, robust evaluation directly fuels progress and safer application.

Regarding **current state**, we have starting to see LLMs being piloted or used in biotech: assisting in writing research papers, helping doctors with paperwork, designing molecules (there are reports of AI-suggested molecules entering real drug pipelines) . Each such use should come with validation – some companies have internal evaluations, but sharing results (even if anonymized) would benefit the community. We might soon see case studies published, like "We deployed an LLM for X task in a pharma company for 6 months, here's how we evaluated it and what the outcomes were." That would be immensely informative for others.

Looking at **future implications**, if LLMs become integrated into critical decision-making, evaluation might even become a continuous regulatory requirement. For instance, just as manufacturers of diagnostic tests have to periodically calibrate and test their equipment, developers of an AI diagnostic assistant might be mandated to re-evaluate the AI on new test cases regularly and report those results. This could lead to certified evaluation protocols and maybe third-party evaluation bodies for AI in healthcare. Another interesting angle is personalizing evaluations: maybe a particular hospital cares about performance on its own data (notes style, patient demographics) – evaluations will be custom-run for that environment.

Another future path is the development of **AI that can evaluate AI** more effectively. We touched on LLMs as judges. It's possible that as models get even more advanced, we might entrust them to do a lot of preliminary evaluation heavy-lifting – e.g., an AI could read 100 generated summaries and highlight 10 that look potentially problematic for a human to review, rather than a human reading all 100. This kind of AI-assisted evaluation could make it easier to scale and maintain evaluations as models and data scales up. Initial results like Croxford et al.'s indicate it's feasible to some extent (pmc.ncbi.nlm.nih.gov).

In concluding, we reaffirm that **depth of evaluation is more important than breadth** when lives or significant decisions are at stake. A comprehensive evaluation as outlined in this report might seem labor-intensive – and it is – but the payoff is confidence in the model's performance and clearer insight into its failure modes. This enables deploying LLMs in biotech in a way that maximizes benefit (efficiency, new insights, reduced workload) while minimizing risk (errors, misinformation, harm). As LLM technology continues to evolve, so too must our evaluation strategies. By staying vigilant, collaborative, and rigorous in how we assess these models, we can ensure that they truly become valuable partners in scientific discovery, healthcare delivery, and all facets of biotechnology.

Ultimately, the goal is for LLMs to be **reliable tools** that biotech professionals trust and find indispensable. Building that trust starts with strong evaluation. As the saying goes in engineering:

"You can't improve what you don't measure." Through the extensive discussion and guidelines provided in this report, we hope to empower AI developers and biotech stakeholders to measure the right things and to measure them well – catalyzing improvements that align LLM capabilities with the critical needs of biotechnology and medicine.

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

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 North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

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