Available Datasets

The following datasets will be provided, pre-loaded into BigQuery where applicable for participants to use during the hackathon.

Data Source Name	URL	Description	Use it for
GPQA Diamond Set	https://github.com/idavidrein/gpqa	Expert-level graduate biology, physics, and chemistry questions with verified answers.	Benchmarking AI reasoning capabilities, training question-answering models, or evaluating scientific understanding.
PubMed QA	https://pubmedqa.github.io/	Question answering dataset derived from PubMed abstracts with expert annotations.	Training biomedical question answering systems, evaluating reasoning over scientific text, or building evidence-based answer systems.
BioASQ	https://participants-area.bioasq.org/datasets/	Biomedical semantic indexing and question answering challenge with curated datasets and benchmarks.	Training biomedical QA models, evaluating information retrieval systems, or benchmarking NLP performance on medical text.
bioRxiv & medRxiv	https://www.biorxiv.org/ https://www.medrxiv.org/	Preprint servers containing early-stage research papers in biological and medical sciences before peer review.	Accessing cutting-edge research findings, tracking emerging trends, or analyzing pre-publication scientific discourse.
PubMed Central	https://pmc.ncbi.nlm.nih.gov/	Free full-text archive of biomedical and life sciences journal literature.	Accessing complete research articles, extracting detailed methods, or conducting comprehensive literature reviews.
PubMed Abstracts	https://pubmed.ncbi.nlm.nih.gov/download/	Downloadable collection of biomedical literature abstracts and citations from MEDLINE.	Large-scale text mining, trend analysis in biomedical research, or training NLP models on scientific abstracts.
PubTator 3.0	https://www.ncbi.nlm.nih.gov/research/pubtator3/	Annotated biomedical literature with automatically extracted entities like genes, diseases, chemicals, and mutations.	Named entity recognition tasks, building knowledge extraction pipelines, or mapping relationships between biological concepts.
PubTator Web API	https://www.ncbi.nlm.nih.gov/research/pubtator3/api	Programmatic access to PubTator's entity annotations and literature mining capabilities.	Integrating automated entity extraction into workflows or building real-time literature analysis applications.
ChEMBL	http://ebi.ac.uk/chembl/	Database of bioactive molecules with drug-like properties, including binding, functional and ADMET information.	Drug discovery, compound screening, target identification, or structure-activity relationship analysis.
Genome Aggregation Database (gnomAD)	https://gnomad.broadinstitute.org/	Large-scale database of human genetic variation from population sequencing studies.	Variant interpretation, identifying disease-associated mutations, or filtering common genetic variants.
Genotype-Tissue Expression (GTEx)	https://gtexportal.org/home/	Comprehensive resource of gene expression and regulation across multiple human tissues.	Understanding tissue-specific gene expression, eQTL analysis, or investigating gene regulation mechanisms.

Cell Index Database (CELLX)	https://pubmed.ncbi.nlm.nih.gov/25592564/	Database of cell line authentication and characterization data for cancer research.	Cell line validation, selecting appropriate models for experiments, or tracking cell line provenance.
ClinicalTrials.gov, available via Dimensions.ai: Comprehensive Dataset for Research & Innovation	https://clinicaltrials.gov/	Registry of clinical studies with trial design, outcomes, and recruitment information worldwide.	Analyzing trial success factors, identifying recruitment opportunities, or tracking therapeutic development progress.
Human Protein Atlas	https://www.proteinatlas.org/	Comprehensive resource mapping human proteins in cells, tissues and organs using antibody-based profiling.	Protein expression analysis, tissue specificity studies, subcellular localization, or biomarker discovery.
CLaRA Models (pre-trained base, instruct, and E2E versions)	https://huggingface.co/apple/CLaRa-7B-Instruct	Pre-trained language models optimized for biomedical and scientific text understanding and generation.	Question answering on scientific literature, extracting information from papers, or building domain-specific AI assistants.
ClinVar - Human Variant Annotation Datasets	https://www.ncbi.nlm.nih.gov/clinvar/	Public archive of relationships between genetic variants and human health conditions with clinical interpretations.	Clinical variant classification, genotype-phenotype mapping, or building variant interpretation pipelines.
DepMap (Cancer Dependency Map)	https://depmap.org/portal	Systematic identification of cancer vulnerabilities and dependencies through CRISPR screens across cell lines.	Target discovery for cancer therapy, understanding genetic dependencies, or identifying synthetic lethal interactions.
GWAS Catalog (OpenTargets Genetics)	http://ebi.ac.uk/gwas/	Curated collection of genome-wide association studies linking genetic variants to traits and diseases.	Identifying disease-associated variants, understanding genetic risk factors, or prioritizing targets based on genetic evidence.
GWAS Catalog (OpenTargets Platform)	http://ebi.ac.uk/gwas/	The Open Targets Platform is a comprehensive data integration tool that supports systematic identification and prioritisation of potential therapeutic drug targets.	Identifying disease-associated variants, understanding genetic risk factors, or prioritizing targets based on genetic evidence.
PubMed Knowledge Graph (PKG)	https://www.ncbi.nlm.nih.gov/research/bionlp/RESTful/pmcoa.cgi/BioC_json/	Structured representation of biomedical concepts and their relationships extracted from PubMed literature.	Graph-based queries to discover hidden connections, pathway analysis, or building knowledge-driven search interfaces.
Reactome	https://reactome.org/	Curated database of biological pathways and molecular interactions across species.	Pathway enrichment analysis, understanding disease mechanisms, or visualizing biological process networks.
CLaRA Github Implementation	https://github.com/apple/ml-clara	Open-source implementation of Apple's Contrastive Learning for Retrieval-Augmented generation model.	Building retrieval-augmented AI systems, implementing advanced NLP pipelines, or fine-tuning language models for biomedical text.
Clinical Genome Resources (ClinGen)	https://clinicalgenome.org/	Authoritative resource defining clinical validity of gene-disease relationships and variant pathogenicity.	Gene-disease relationship validation, variant classification, or clinical genomics decision support.

Clinical Interpretations of Variants in Cancer (CIViC)	https://civicdb.org/	Community-curated database of clinical interpretations of cancer variants with evidence and therapeutic implications.	Cancer variant interpretation, precision oncology, identifying actionable mutations, or understanding resistance mechanisms.
Open Research KG (ORKG)	https://orkg.org/	Structured, machine-actionable representation of scientific contributions and their comparisons.	Comparing research findings across studies, building structured literature summaries, or meta-analysis automation.
Pathway Commons	https://www.pathwaycommons.org/	Integrated resource aggregating biological pathway and interaction data from multiple public databases.	Multi-pathway analysis, integrative network analysis, or comprehensive pathway enrichment across data sources.
Semantic Scholar (S2ORC)	https://api.semanticscholar.org/	Citation graph and paper metadata with AI-powered insights and recommendations.	Network analysis of research communities, identifying influential papers, or building intelligent literature discovery tools.
STRING	https://string-db.org/	Database of known and predicted protein-protein interactions including physical and functional associations.	Network analysis of protein interactions, identifying functional modules, or predicting protein function.