🧩 1. Sumber Dataset Publik (Legal dan Bebas Lisensi)

Karena data medis bersifat sensitif, pengembangan sistem **AI Symptom Checker + Smart Triage** harus menggunakan dataset yang legal, terbuka, dan tidak mengandung identitas pasien.
Berikut beberapa sumber dataset publik yang bisa digunakan untuk melatih model klasifikasi gejala–penyakit:

Dataset	Deskripsi	Format	Link
SymCAT (Symptom–Disease Associations)	Basis data berisi lebih dari 600 penyakit dan 1000 gejala beserta probabilitas keterkaitannya. Sangat berguna untuk membuat peta *symptom-to-disease mapping*.	JSON / CSV	https://www.symcat.com/diseases
Human Symptoms Dataset (Kaggle)	Dataset 13.000 entri keluhan medis umum, cocok untuk melatih model klasifikasi teks dasar.	CSV	https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset
Medical Dialogue Dataset (MedDialog-EN)	Dataset percakapan pasien–dokter dalam bahasa Inggris, bisa digunakan untuk ekstraksi keluhan awal pasien.	JSON	https://github.com/UCSD-AI4H/Medical-Dialogue-System
ClinicalBERT Datasets (MIMIC-III Derived)	Dataset klinis hasil de-identifikasi untuk penelitian NLP medis. Dapat digunakan untuk fine-tuning model bahasa medis.	SQL / CSV	https://physionet.org/content/mimiciii-demo/

> ⚖️ ***Catatan Etika & Legalitas***
>
> *Gunakan hanya data yang* ***tidak mengandung informasi pribadi (PII)****.*
>
> *Jangan melakukan web scraping ke situs medis tanpa izin eksplisit.*
>
> *Dataset ini cocok untuk tahap riset, prototipe, dan edukasi,* ***bukan*** *untuk keputusan medis klinis langsung.*