Medical Instruction Datasets

Public Medical Instruction Datasets

High-quality data for medical LLM training

📚 Major Datasets

- **MedInstruct:** 100K clinical instructions (diagnosis, treatment)
- **HealthCareMagic:** 200K patient-doctor conversations
- **MedQA:** 60K medical exam questions with explanations
- **PubMedQA:** 1K expert-annotated biomedical QA pairs
- **ChatDoctor:** 100K patient-physician dialogues

40%
Diagnosis

30%
Treatment

30%
Education

✅ Quality Metrics

- Expert validation rate: 85-95%
- Average response length: 150-300 tokens
- Domain coverage: 50+ medical specialties
- Language diversity: English, Chinese, Spanish