PROJECT 2: “Decode the Jargon”: Abbreviation Disambiguation in Clinical Notes

Business Scenario
Clinical medical texts overflow with abbreviations whose meaning changes with context (e.g., “PCP” = *Primary Care Physician* **or** *Pneumocystis Pneumonia*). Mis-interpretation derails downstream NLP tasks and decision-support systems. Your mission is to build a model that reads the surrounding words and picks the correct expansion.

Learning objectives

- Compare **static** (Word2Vec) vs. **contextual** (ClinicalBERT) embeddings for word-sense disambiguation.
- Practise **sliding-window** extraction, sequence modelling with **LSTM**, and attention-based interpretability.
- Explore **class imbalance** handling and per-abbreviation vs. multi-label training strategies.

Dataset options

Source	Contents	Access notes
UMN SHRS CUI-mapped corpus	Thousands of sentences with ambiguous abbreviations + gold CUIs	Free academic use (sign-up)
Custom mini-set (fallback)	Manually curate ~500 sentences per abbreviation from PubMed abstracts or MIMIC clinical notes	Provide a starter CSV template

Tasks

1. **Data wrangling**
    - Parse JSON/CSV, keep only sentences containing a target abbreviation.
    - Extract a *context window* (e.g., ±10 tokens) with a sliding-window routine.
    - SpaCy pipeline: tokenise, lower-case, remove PHI placeholders.
2. **Feature pipelines**
    - **Static baseline** – average Word2Vec/BioWordVec embeddings → logistic-regression classifier.
    - **Sequence model** – LSTM (or GRU) over the embedding sequence → soft-max over meanings.
3. **Training & evaluation**
    - Split by *document*, not sentence, to avoid leakage.
    - Loss = cross-entropy; metric = **accuracy** + **macro-F1** (handles class imbalance).
    - Visualise attention (or gradient*input) weights to highlight decisive context words.

Enhancements / stretch goals

- Fine-tune **ClinicalBERT** or **BioClinicalBERT** and compare sample efficiency.
- **Multi-task** setup: train one model that disambiguates *all* abbreviations jointly.
- t-SNE visualisation of hidden states for different senses.

- Add **confidence thresholding**; route low-confidence cases to “ask a human” bucket.

6 Deliverables

- Clean, documented notebook or script.
- Confusion matrix + two error-analysis paragraphs.
- Attention/IG heat-map for at least three correctly and three incorrectly predicted sentences.
- README with install and run instructions.
- Presentation deck (template attached)