Problem statement :
Drug-Related Question Risk Assessment

Dvora Goncharok
&
Arbel Shifman

Use Case:

- Patients ask medication-related questions through **chatbots/virtual assistants** that vary in risk. Proper risk classification can save lives by preventing harmful self-medication decisions.
- Current LLMs treat all questions alike, — which may lead to **unsafe answers**.
- We aim to build a system that **classifies question risk levels** to improve safety in drug-related QA.
- Accurate risk classification can **help prevent harmful self-medication and even save lives.**

The need:

Risk-aware question classifier

Problem statement :
Drug-Related Question Risk Assessment

Problem Definition:

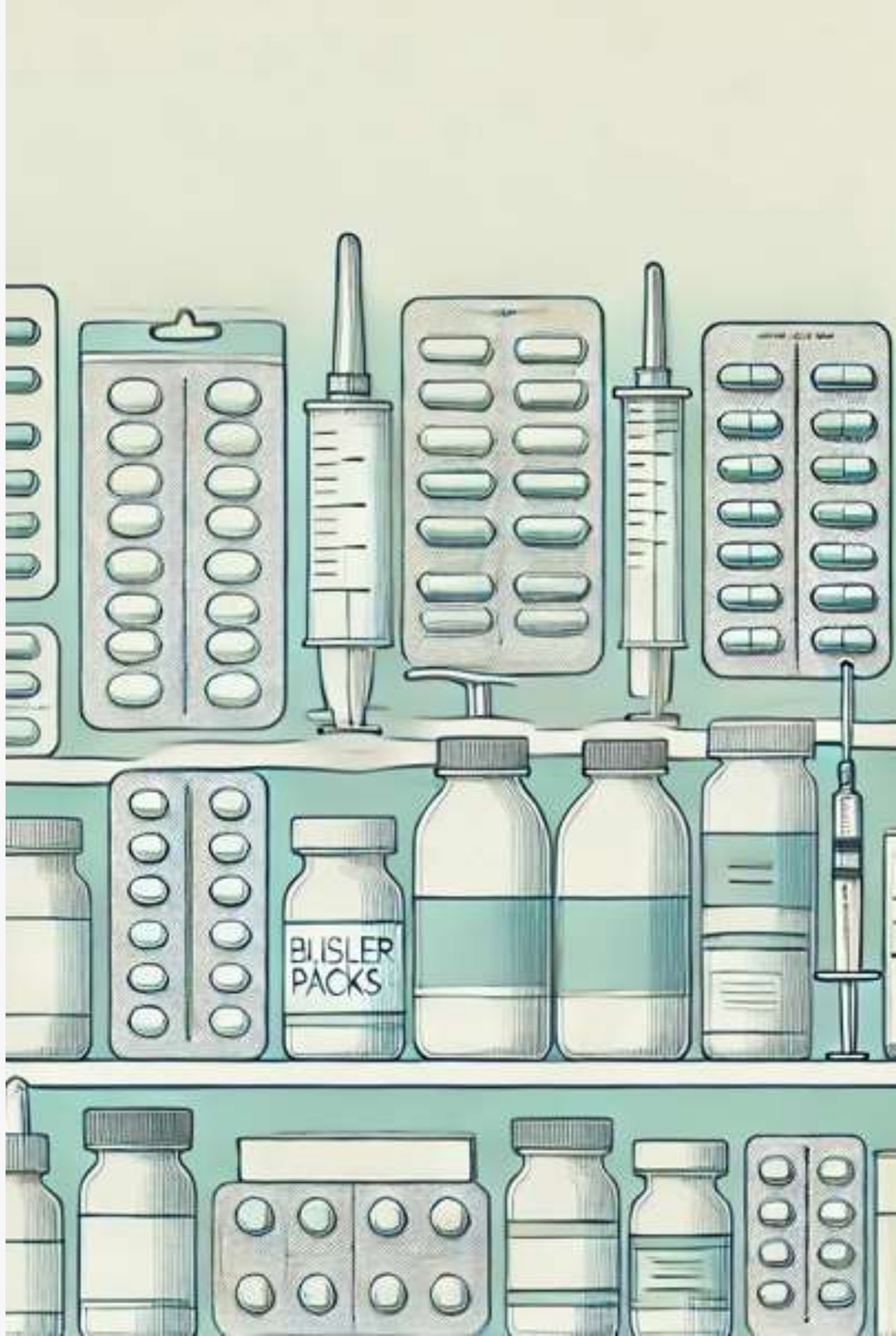
- Input: Free-text drug question about dosage, side effects, interactions etc.
- Output: **Risk level classification** (General/Personal/Critical).
- Additional info:

  Drug names (via NER or existing labels) help contextualize the question for better risk assessment

Problem statement :
Drug-Related Question Risk Assessment

Challenges:

- Ambiguous, non-expert language.
- Same topic, different risk levels depending on context.
- Brand-name medications may not be recognized by the model, making accurate understanding more difficult.
- **High precision needed to avoid clinical harm.**
- Few critical-risk examples in the dataset may challenge model training.



Training and test data

Data type and labels:

- Labeled medication questions from the publicly available **MedInfo2019-QA-Medications dataset**.
- We manually add a new annotation layer with:

  Risk Level: General (Safe) / Personal / Critical (Dangerous)

Data source:

- Drug QA sheet from the MedInfo2019 dataset.
- Data is openly accessible on GitHub.
- **We enrich data with new risk-level labels.**
- ~700 real-world patient questions and expert answers.
- The dataset already includes:
  - Question Type (e.g., Dosage, Side effects, Usage)
  - Focus (Drug) – main medication mentioned
  - Section Title, Answer, and Source URL.

Realistic examples:

Drug	Question Type	Risk Level	Input Question
Ibuprofen, Aspirin	Interaction	General	"?Can I take ibuprofen and aspirin together"
Flagyl	Interaction	Critical	"?I took Flagyl and drank wine – what do I do"
Prozac	Dosage	Personal	"I've been on Prozac a week and still anxious"

Evaluation Metrics

Metrics

- Accuracy,
- F1-score
- Confusion Matrix – to measure classification quality.
- Per-class performance (especially for Critical risk level).

Evaluation Method

- Manual labeling of ~700 examples.
- Train/Test split: 80/20.
- k-fold cross-validation to improve robustness due to small dataset.

Baseline and Comparison

Baseline:

Traditional classifier (e.g., Naïve Bayes or Logistic Regression) **using TF-IDF**.

- Purpose: compare against a simple method.

LLM-based classifier:

- **Fine-tuned DistilBERT model** (efficient BERT variant) **or prompted LLM** (e.g., GPT) for comparison.
- Purpose: Test whether LLM-based methods improve classification performance over traditional baselines.
- Comparison of results across models using same split and metrics.
- Purpose: ensure consistent evaluation conditions.