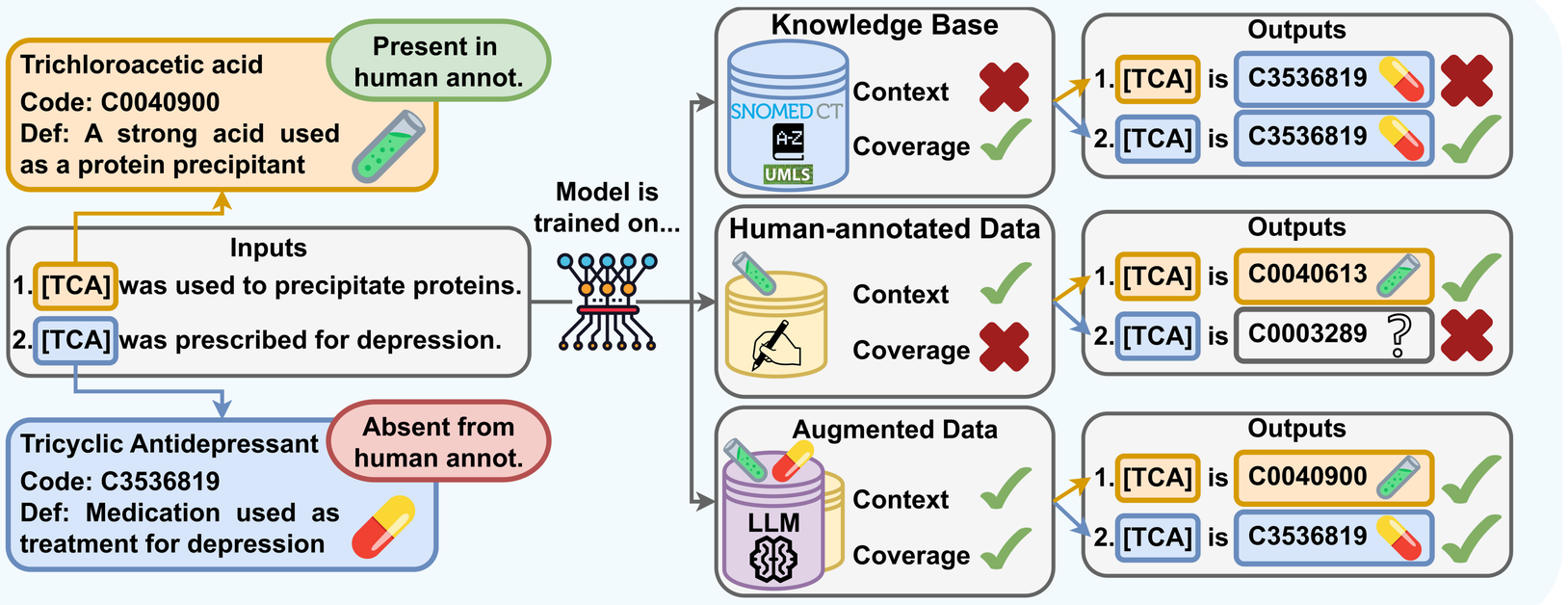
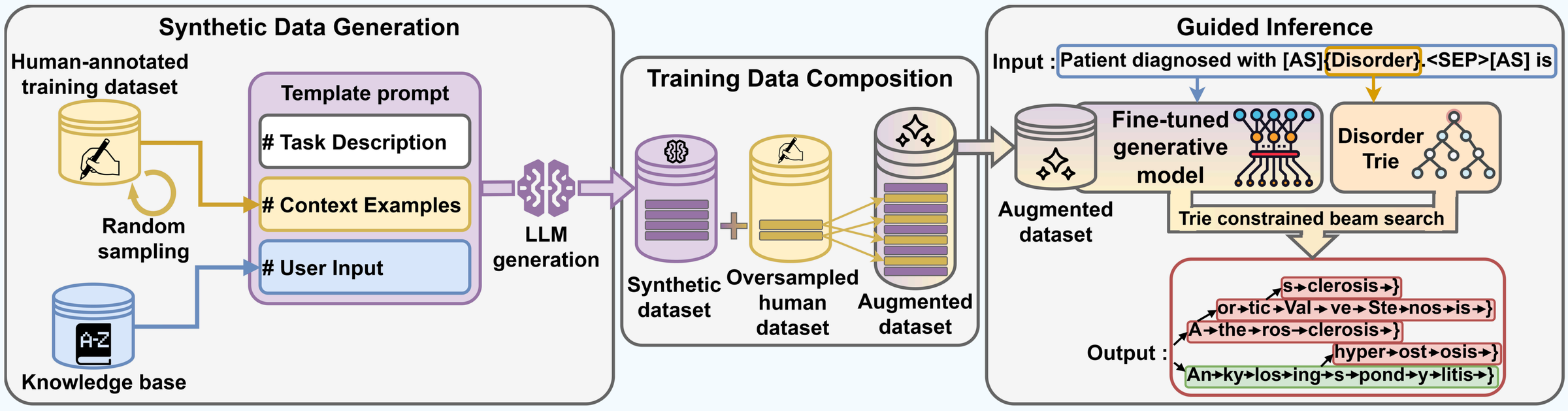
AI Generates Annotated Data When Human Annotation Is Costly



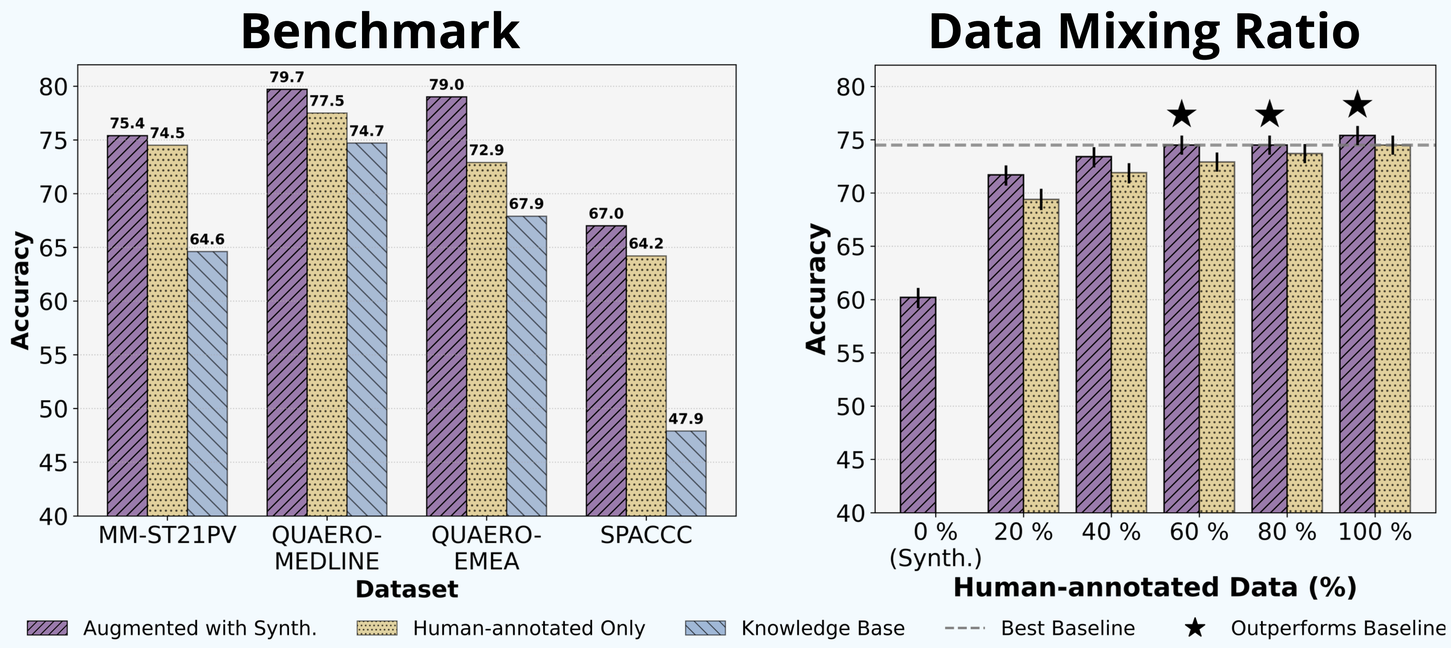
Background: Biomedical Entity Linking converts clinical notes into structured data. Rule-based systems rely on knowledge bases and ignore context, often causing errors. Machine learning systems capture context, but **require large amount of labeled data, which are scarce and costly.**



Method: We use an LLM to generate **synthetic examples**, helping our model understand a wide range of concepts. Then, we **guide its answers** so it can only pick from a valid list of medical concepts.



Results: Our trained system sets a **new state-of-the-art** on four major benchmarks across three languages, outperforming all previous methods in biomedical entity linking. Besides, our approach also **reduces the need for expensive human annotation**, reaching full data performance with only **60% of human annotations.**



Conclusion: We addressed the challenge of expert annotation scarcity in Biomedical Entity Linking by generating high-quality synthetic data, helping to **maximize the value of human annotation**.

Paper GitHub Contact

Authors: Adam Remaki, Christel Gérardin, Eulàlia Farré-Maduell, Martin Krallinger, Xavier Tannier