LLM-guided normalization **resolved 99.5% of Hungarian free-text drug entries into structured English language query strings**, with subsequent OMOP vocabulary lookup achieving an F1 of 0.71 — exploring the possibility of automated CDM conversion of free-text non-English drug data for future research.

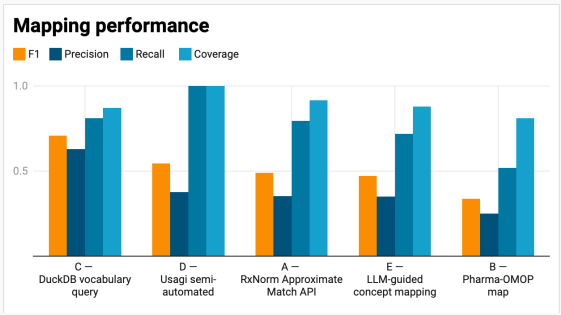
# *LLM-Guided Drug Normalization for Rapid OMOP CDM Integration of Hungarian Drug Data*

**Background:** Real-world evidence research depends on standardized clinical data, yet medication records in Hungarian hospitals are documented in heterogeneous free-text formats — brand names, local abbreviations, and unstructured dosing — with no controlled vocabulary enforced at point of entry. Mapping this data to OMOP CDM is a technical prerequisite for real world evidence research, but it remains a demanding, multi-step challenge for which rule-based approaches have historically fallen short. The rapid maturation of large language models make it timely to evaluate whether these emerging technologies can close that gap.

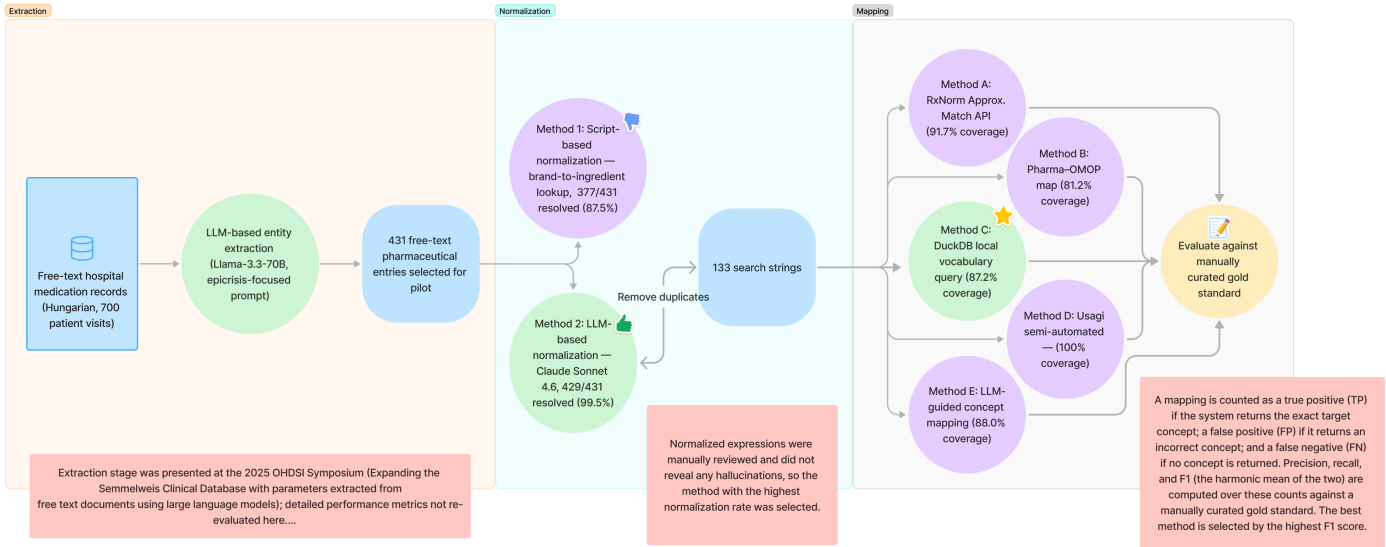
## Result 1: Comparison of script based and LLM based normalization.

| Evaluation criteria | Script based normalization | LLM based normalization |
|---|---|---|
| Development effort | High upfront | Low |
| Maintenance | Requires frequent updates, | More robust to change, just prompt fine-tuning |
| Interpretability | Fully explicit, interpretable rule set | Black-box, harder to justify |
| Risks | Misses instead of errors | Hallucinations – not detected |
| Costs | Low runtime cost | Model dependent, can be high |
| Normalized medications (total 431) | 377 (87.5%) | 429 (99.5%) |

## Result 2: Evaluating the mapping performance of 5 different methods




# Methods




Limitations: This study is exploratory, using binary classification metrics that do not reward ranked or partially correct mappings. The dataset is limited to a single source with a diabetic profile covering Hungarian pharmaceuticals, restricting generalizability, and runtime API dependencies prevent high-volume production deployment and scalability.

Orsolya Bali [1], Loretta Kiss [1], Eszter Kővári [1], Ágota Mészáros [1], Mónika Hujter [2], Zsófia Práger [2], Tibor Héja [1], Csaba Nemes [1], Zsolt Bagyura [1]
[1]Institute for Clinical Data Management, Semmelweis University
[2]Hiflylabs Zrt