Automated Extraction of Pharmaceutical Knowledge Using Artificial Intelligence for the STEP Database

Sifan Hu[1]; Sagar Uprety[1]; Sridhar Radhakrishnan[2]; Chandra Sekharan[3]; Peter Silvera[4]; Ranjana Srivastava[4]; James Cummins[5]; Kristen Porter[5]; Santosh Reddy Nella[6]; Daniel Schaufelberger[7]; Bruno Hancock[8]; Savani Deshpande[1]; Smita Salunke[1]

[1] UCL School of Pharmacy, London, UK; [2] School of Computer Science, University of Oklahoma, USA; [3] Department of Computer Science, Texas A&M University Corpus Christi, USA; [4] Advanced Bioscience Laboratories, Inc, USA; [5] National Institute of Allergy and Infectious Diseases (NIAID), USA; [6] Vincatis Technologies Private Limited, India; [7] Schaufelberger Consulting LLC, USA; [8] Aleurites LLC, USA

INTRODUCTION

GAP

Manual data integration:

Needs for a shift towards automated processes

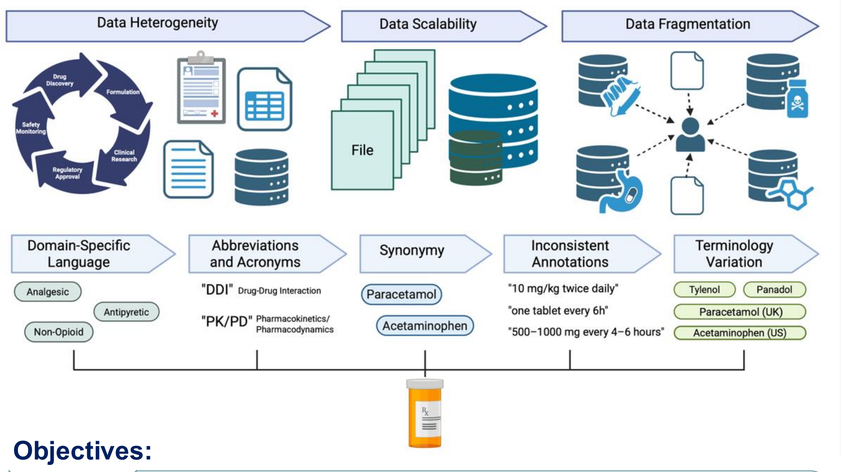
SOLUTION

Using AI to create an enhanced STEP database (STEP 2.0)

BENEFITS

A smarter, more insightful and more user-friendly system

Challenges in pharmaceutical knowledge extraction:



Objectives:

01 • To provide a comprehensive analysis of existing AI-driven methods for information extraction, highlighting their tools, techniques, and pharmaceutical applications.

02 • To critically assess the limitations, unresolved challenges, and performance gaps that constrain current information extraction systems.

03 • To explore emerging trends and potential solutions that hold promise for improving accuracy, scalability, and regulatory compliance in pharmaceutical information extraction.

SURVEY STRATEGY

Structured Information extraction

Knowledge extraction

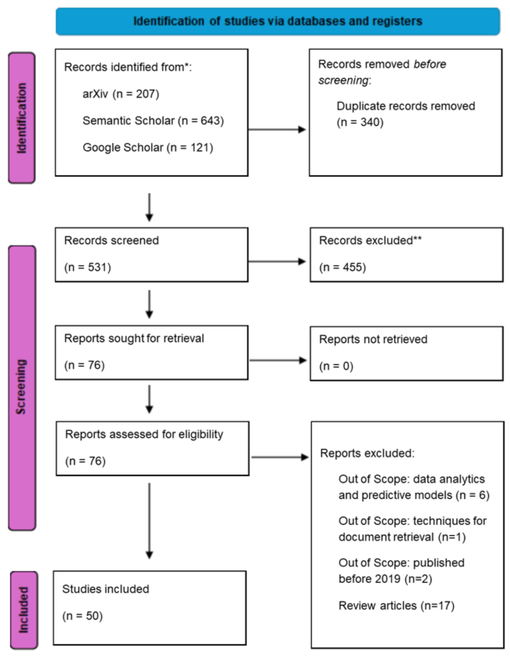
Named entity recognition

Relation extraction

Attribute value extraction

Pharmaceutical

Biomedical



CONCLUSION

The results of this review informed model selection for applying AI-enabled information extraction to the STEP database, although direct evaluation on STEP data is ongoing.

References

1. Neumann, M., et al., *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing.* ArXiv, 2019. **abs/1902.07669**.
2. Kocaman, V. and D. Talby, *Accurate Clinical and Biomedical Named Entity Recognition at Scale.* Software Impacts, 2022. **13**: p. 100373. DOI: 10.1016/j.simpa.2022.100373.
3. Sung, M., et al., *BERN2: an advanced neural biomedical named entity recognition and normalization tool.* Bioinformatics, 2022. **38**(20): p. 4837-4839. DOI: 10.1093/bioinformatics/btac598.
4. Lee, J., et al., *BioBERT: a pre-trained biomedical language representation model for biomedical text mining.* Bioinformatics, 2020. **36**(4): p. 1234-1240. DOI: 10.1093/bioinformatics/btz682.
5. Jahan, I., et al., *A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks.* Computers in Biology and Medicine, 2024. **171**: p. 108189.
6. Li, M., et al., *BiomedRAG: A retrieval augmented large language model for biomedicine.* Journal of Biomedical Informatics, 2025. **162**: p. 104769. DOI: 10.1016/j.jbi.2024.104769.
7. Chen, Q., et al., *An extensive benchmark study on biomedical text generation and mining with ChatGPT.* Bioinformatics, 2023. **39**(9). DOI: 10.1093/bioinformatics/btad557.
8. Chellagurki, P., et al. *Biomedical Relation Extraction Using LLMs and Knowledge Graphs.* in *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService).* 2024.
9. Zhang, J., et al., *A Study of Biomedical Relation Extraction Using GPT Models.* AMIA Jt Summits Transl Sci Proc, 2024. **2024**: p. 391-400.

RESULTS

Information Extraction Tasks:

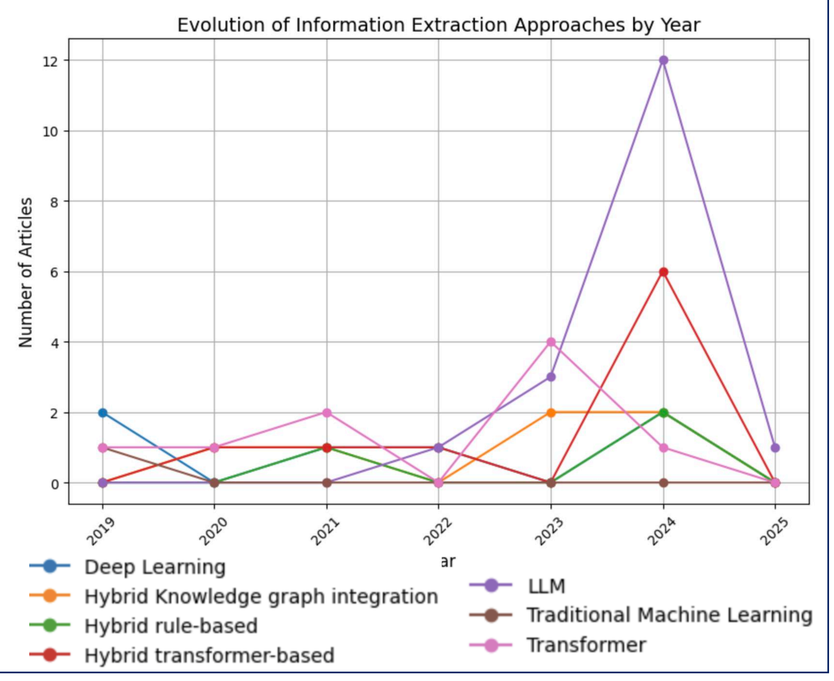
Named entity recognition | Relation extraction | Question & Answering | Attribute value extraction | Text classification | Language recognition

Datasets and Evaluation Metrics:

Chemical & Drug Dataset	Disease & Gene/Protein Datasets	Clinical Datasets
BC4CHEMD, BC5CDR, DDI, CHEMPROT, KD-DTI	NCBI-Disease, Linnaeus, BC2GM, JNLPBA, GAD	THYME, EU-ADR

There are also uses of custom datasets, which involves manual or AI-assisted annotation, with information retrieved from clinical records, public databases, and scientific literature.

Models are assessed using standard NLP metrics:
Precision: The proportion of true positive results among all predicted positives.
Recall: The proportion of true positives identified out of all actual positives.
F1-score: The harmonic mean of precision and recall, widely used as the main performance indicator in biomedical NER and RE tasks.
Accuracy: Occasionally reported, though less common for imbalanced classification tasks like NER.



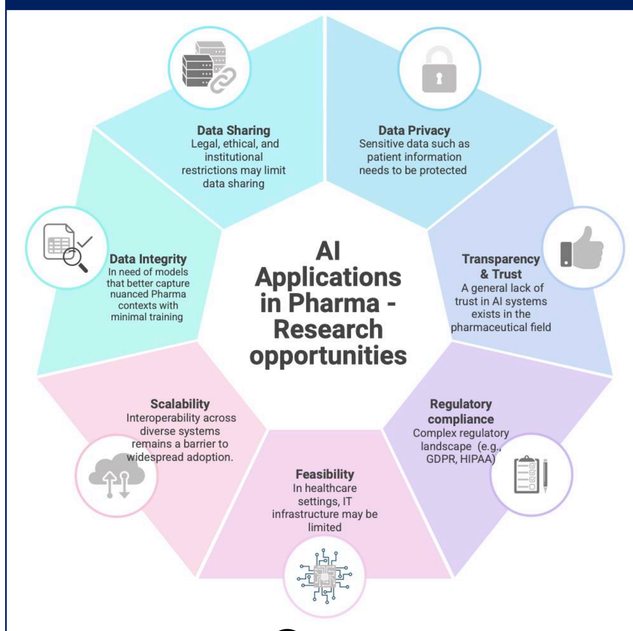
COMPARATIVE ANALYSIS & LIMITATIONS

	Strength	Limitations	Use Cases
Traditional ML	Fast, interpretable	Poor contextual modeling	Pharmaceutical production records processing
Deep Learning	Handling Structural Complexity	Compute-intensive, high data dependency	Chemical NER
Transformers	High accuracy	Expensive training	Medical/ clinical/ pharmaceutical NER
Hybrid	Balanced performance	Complex implementation	KG–based clinical/ pharmaceutical RE
Generative LLMs	Few shots adaptability (Small samples)	Hallucinations, inconsistency	Attribute extraction from unstructured texts

Representative Metrics for Different Models on NER and RE Benchmarks

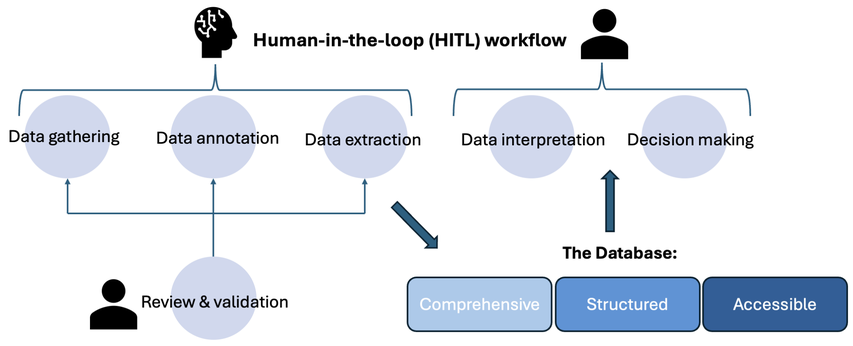
Architecture	Benchmark dataset	F1 score	References
ScispaCy sci md	BC4CHEMD	84.55	Neumann et al. (2019)[1]
BiLSTM-CNN-Char	BC4CHEMD	93.6	Kocaman et al. (2022)[2]
BERN2	BC4CHEMD	92.8	Sung et al. (2022)[3]
BioBERT	BC4CHEMD	92.36	Lee et al. (2020)[4]
GPT-3.5 (zero-shot)	BC4CHEMD	26.01	Jahan et al. (2024)[5]
LlaMA2-13b-BiomedRAG	DDI	80.00	Li et al. (2025)[6]
ChatGPT (zero-shot)	DDI	51.62	Chen et al. (2023)[7]
LlaMA2-7b (fine-tuned)	EU-ADR	93.18	Chellagurki et al. (2024)[8]
GPT-3.5-turbo (zero-shot)	EU-ADR	80.9	Zhang et al. (2024)[9]

EMERGING TRENDS & RESEARCH OPPORTUNITIES



Emerging AI solutions

- Agentic AI
- Small Language Models (SLMs)
- Privacy-Preserving Data Sharing



Author's recommendations

- Use domain-specific pretraining to enhance relevance and accuracy.
- Incorporate normalization techniques and ontology alignment to ensure consistency across varied data sources
- Validate model outputs using a human-in-the-loop process or RAG to reduce errors and improve trust
- Fine-tune models to better capture the language and structure unique to pharmaceutical data.
- Combine symbolic approaches with neural models for better coverage and interpretability
- Evaluate models using multiple, domain appropriate benchmarks or create benchmark on custom data to assess generalizability and task performance.

This project was funded by NIAID-DAIDS Task Order 75N93024F00005 entitled, "Use of Artificial Intelligence/Machine Learning tools to Enhance the STEP Pediatric Excipient Database" awarded to Advanced Bioscience Laboratories, Rockville, Maryland, USA.