LASER AI

Give Me the Numbers: Fine-Grained, Structured Data Extraction Beyond Narrative LLM Outputs

Ewa Borowiack, Ewelina Sadowska, Joanna Konieczna, Monika Opałek, Damian Stachura, Artur Nowak

Evidence Prime, Krakow, Poland

MSR179

Background

Generative large language models (LLMs) have demonstrated strong capabilities in extracting information from unstructured biomedical text. However, their outputs are typically narrative in nature, lacking the fine-grained structure required for evidence synthesis, statistical analysis, and downstream reuse. This limitation becomes particularly evident in tasks involving patient-flow data, where precise numerical values, contextual relationships, and per-arm distinctions must be consistently captured.

In traditional systematic review workflows, data extraction is often performed using spreadsheet-based tools. While sufficient for simple datasets, this approach encourages flattening complex information into wide tables and free-text fields, limiting the ability to represent relationships between variables, enforce consistency, and support scalable analysis.

Laser AI addresses these limitations by adopting a database-oriented approach to evidence extraction, in which extracted data are treated as structured, relational entities rather than rows in a spreadsheet. This enables consistent handling of repeated and hierarchical elements such as study arms, and supports direct integration with downstream analytical workflows.

A key aspect of this approach is granular data extraction. Importantly, achieving granularity requires dedicated infrastructure rather than annotation guidelines alone. Laser AI supports this through relational extraction forms, predefined field structures, controlled vocabularies, and native handling of repeated entities. This design enables both AI-assisted extraction and human validation at the level of individual data elements, improving transparency, consistency, and quality control.

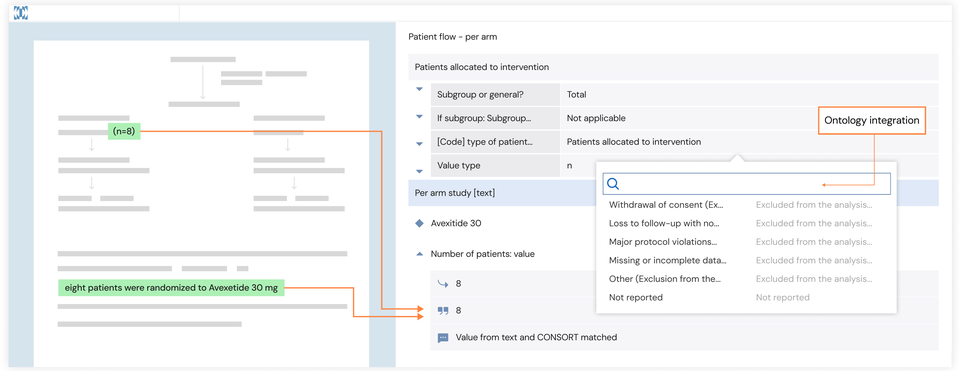
This architecture not only supports accurate extraction, but also provides a robust environment for systematic dataset development, where structured fields, relationships,s can be consistently applied.

By combining domain-specific LLM agents with a structured data model, Laser AI enables extraction of analysis-ready, vocabulary-mappable data with full traceability, including supporting text, document location, and rationale. [Fig 1] This approach bridges the gap between narrative LLM outputs and the structured requirements of evidence synthesis, supporting scalable and reusable data pipelines.

Objective

To assess the feasibility of specialized, fine-grained large language models for extracting highly contextual patient-flow data and to develop a robust framework for high-quality domain datasets, including annotation guidelines and field definitions to support later stages. This work is conducted within the EU-funded LASER-LLM project (FENG.01.01-IP.02-4479/23), which aims to enable structured, vocabulary-mappable, and analysis-ready data extraction to support scalable evidence synthesis.

Interface solutions supporting structured, fine-grained data extraction — Fig 1



Methods

Dataset development

A training dataset of primary studies (target n=50 per domain) was constructed across pharmacotherapy, drug development, and medical devices. Eligible studies included RCTs (mainly parallel-group, with selected cross-over designs) reporting patient-flow information (e.g., assessed, randomized, analyzed), supported by CONSORT diagrams or explicit descriptions.

A representative pool of studies was assembled from multiple sources, including prior systematic review datasets, Cochrane reviews, ClinicalTrials.gov, and targeted PubMed searches.

Study selection was conducted in Laser AI using a two-step process (deduplication and full-text screening) by three reviewers in iterative batches.

Data extraction was performed in Laser AI using structured, granular fields with direct linkage to source text, capturing extracted values, supporting evidence, and reviewer rationale [Fig 2]

Model development & evaluation

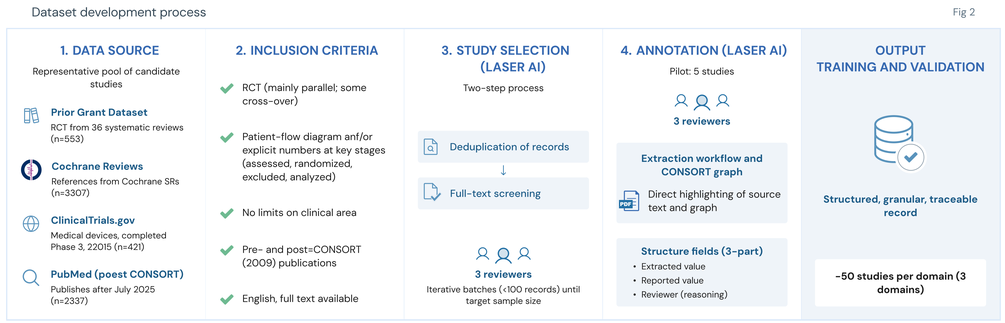
The training dataset was used to iteratively develop and refine LLM prompts and configure domain-specific extraction agents. Prompt design was guided by annotation guidelines and field definitions, ensuring alignment between model outputs and the structured data schema.

AI agents were executed as a modular extraction pipeline, generating one structured record per study (and per study arm where applicable). Each extracted variable was produced at the field level, consistent with a granular, database-oriented data model.

In addition to structured numeric outputs, agents returned an audit trail for each extracted value, including supporting text excerpts, document location, and a brief rationale. This enabled traceability, facilitated human verification, and supported systematic quality control.

Model outputs were evaluated against human-annotated reference data. Quantitative performance was assessed using F1 scores calculated on structured outputs at the field level. In parallel, qualitative error analysis was conducted to identify patterns of discrepancies, assess consistency of extracted values and supporting evidence, and inform further refinement of prompts, field definitions, and annotation guidelines.

Dataset development process — Fig 2



Results

A total of 160 studies were included, with separate training and test sets defined to ensure robust and unbiased evaluation. LLM-based agents demonstrated strong and consistent extraction performance across all patient-flow variables, achieving over 0.91 F1 score for each concept, indicating high agreement with human-annotated reference data. Performance remained stable across variables of varying complexity, suggesting the approach generalizes well to different types of patient-flow information. Also, F1 scores were similar for text level extraction (range 0.93-0.96) and text+graph extraction (range 0.91-0.98). Text+graph examples were more challenging, because often the information was present only in the graph. [Table 1]

In addition to quantitative evaluation, a detailed qualitative review was conducted to assess the correctness, internal consistency, and practical acceptability of both the extracted values and their supporting evidence, including source text quotations and accompanying rationales. This review enabled identification of systematic error patterns, such as boundary misinterpretations and context ambiguity, and provided actionable insights for iterative refinement of annotation guidelines, field definitions, and prompt design [Table 2]. Together, these findings support the reliability of structured LLM-based extraction while highlighting key areas for further optimization.

Quantitative assessment- Field-Level Extraction Performance Metrics — Table 1

Level of extraction	Name of data extraction field	F1 score (text level)	F1 score (text + graph)
Study level	Number of patients assessed for eligibility	0.96	0.94
Study arm level	Number of randomized patients	0.943	0.98
Study arm level	Number of patients that were lost to follow up	0.933	0.91

Qualitative assesment - Error Analysis — Table 2

Number of patients assessed for eligibility	Number of randomized patients	Number of patients that were lost to follow up	Error patterns	Implication
+	+	−	Model attempts to infer missing values by calculating from related data when values are not explicitly reported; reasoning is provided but increases cognitive load for users.	Requires alignment with user preferences: either default to "not reported" or support model-derived values. If the latter, improved interface design is needed to present calculations and reasoning in a clear, low-effort way.
+	−	−	Inconsistencies arise from ambiguity or misalignment between field definitions and real-world reporting, particularly when publications deviate from standard guidelines.	Refinement of field definitions and prompts through inclusion of edge cases and non-standard reporting examples to improve robustness.
−	+	+	Incomplete extraction at the study arm level, with values captured for only one arm while missing for others.	Requires improved handling of repeated entities (study arms) and stricter validation to ensure completeness across all arms

Conclusion

Specialized, fine-grained LLM agents demonstrated strong performance in extracting highly contextual patient-flow data, achieving high accuracy while producing structured, analysis-ready outputs. By operating at the level of granular fields and study arms, the approach enables consistent capture of complex, relational data that are typically difficult to extract using traditional methods.

The integration of domain-specific agents with a database-oriented extraction framework proved critical for supporting both structural and conceptual granularity. This design not only improves machine readability and downstream usability, but also enables transparent, traceable extraction through linked evidence and rationale, facilitating validation and quality control.

Future directions

The study highlights that achieving high-quality structured extraction is not solely a modeling challenge, but also depends on well-defined data schemas, annotation guidelines, and appropriate infrastructure. Observed limitations, including handling of implicit values and variability in reporting standards, underscore the need for continued refinement of prompts, field definitions, and interface design.

Future work will focus on extending the framework to multilingual and data, improving robustness to non-standard reporting, and integrating ontology-based standardization. Ultimately, this approach supports scalable, reusable, and analysis-ready data pipelines for evidence synthesis across domains, including effectiveness, safety, and health outcomes.