

Background

- High-quality, standardised metadata are essential for making data **Findable, Accessible, Interoperable and Reusable (FAIR)**.
- Current metadata curation efforts rely heavily on manual extraction of study-level information, making processes **labour-intensive, inconsistently** applied across platforms, and **difficult to scale**.
- There is a need for **scalable, reproducible** tools that can generate structured metadata from public sources and support interoperability.

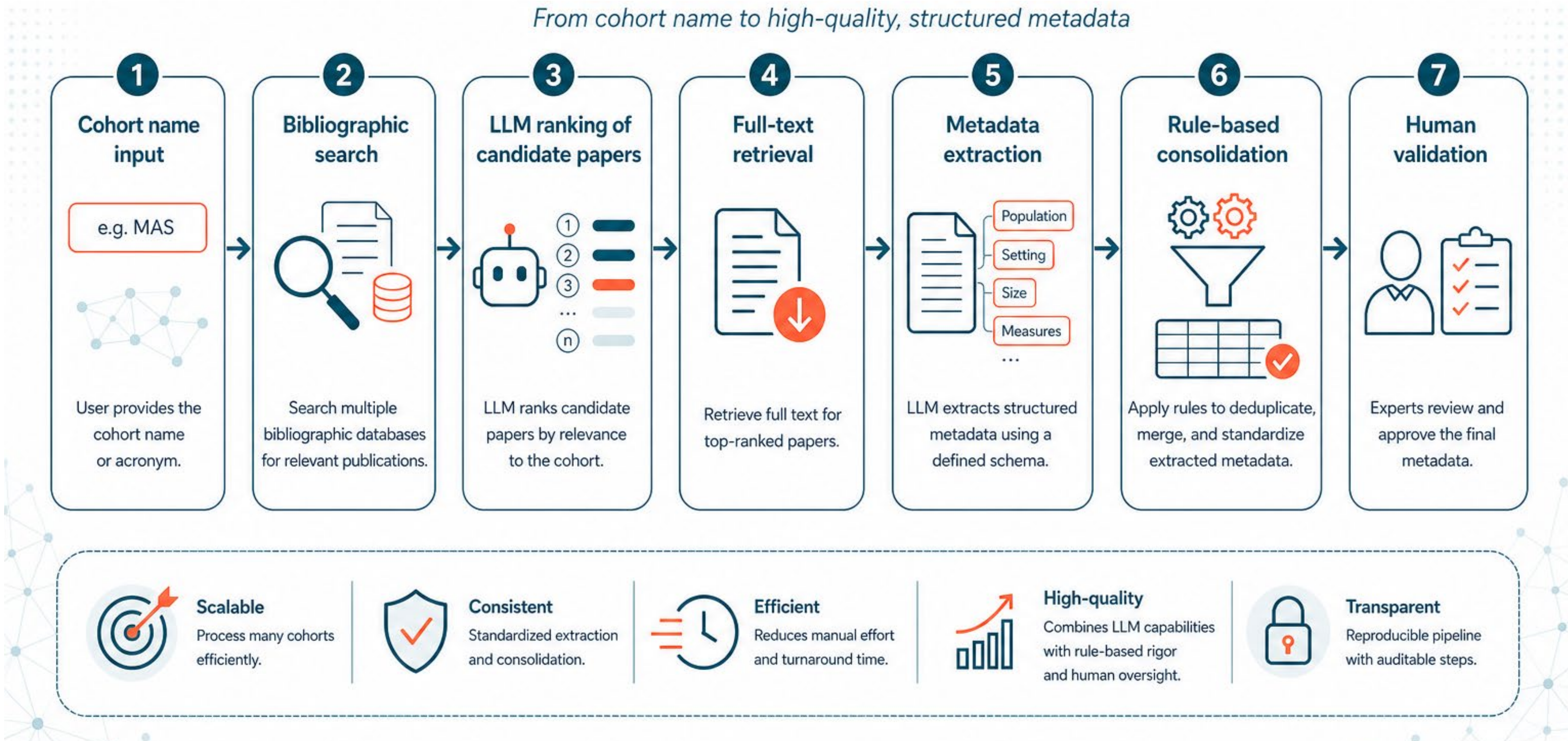
Methods

60
COSMIC cohorts

27
structured JSON fields

HITL
review

- We developed an automated, reproducible workflow that orchestrates discovery, retrieval, and **large language model (LLM)** based metadata extraction within a modular pipeline.
- Extracted metadata are produced as structured JSON and harmonised into machine-readable tables for catalogue integration.
- The proof of concept was applied to 60 members of the Cohort Studies of Memory in an International Consortium (COSMIC).
- Known inputs (cohort names) were used to construct search queries in bibliographic databases.
- Retrieved publications were rated by the LLM for relevance and quality, with top-ranked records selected for downstream extraction.
- Metadata extraction was guided by domain-specific prompts, pre-defined variable schemas, rule-based consolidation, and LLM-assisted interpretation.
- A human-in-the-loop (HITL) review step to verify accuracy.



Results

- The workflow substantially reduced the time and resources required for metadata extraction compared with manual curation.
- Level of completeness of extraction and concordance with investigator-verified metadata was high (>80%) for simple variables (e.g. study design, sample size, and selection criteria).
- Precision for more complex variables (e.g. cognitive measurement, follow-up structure) improved with human verification.
- Discrepancies** mainly arose from ambiguous, outdated or inconsistently reported study descriptions.

Conclusion

- LLM-enabled, human-validated** approach to dementia study metadata extraction is feasible.
- It provides a **scalable** solution that enhances metadata quality and supports data platforms.
- Future development will expand to variable ontologies and improve accuracy
- Explore deployment within **Dementias Platform Australia (DPAU)** and international partners to make dementia data **FAIR**.

Contacts

COSMIC: <https://www.unsw.edu.au/cheba/research-and-impact/consortia/cosmic>
 DPAU: <https://www.dementiasplatform.com.au/>
 Email: rory.chen@unsw.edu.au
 LinkedIn: <https://www.linkedin.com/in/rorychenxy>