CERTARA

AI in Evidence Synthesis: Have We Reached the Promised Land or Are We Still Wandering the Desert?

Authors: Grys M[1], Kallmes K[2], Johnson J[2], Casciano R[3]

[1]Certara, Cracow office, Poland
[2]Nested Knowledge, USA
[3]Certara, USA

Background and Objective

- Evidence synthesis is a cornerstone of health technology assessment (HTA), market access, and regulatory decision-making. Traditional literature reviews (LRs) are resource-intensive, requiring thousands of hours for screening, data extraction, and quality assessment. Artificial Intelligence (AI) has transformed paradigms in LR, offering potential to automate and accelerate previously manual steps.
- This study evaluates the use of a commercially available AI-enhanced LR tool (Nested Knowledge – AutoLit), adopted by Certara in early 2025. We aimed to assess the tool's effectiveness in reducing workload, improving accuracy, and enabling flexible applications—including both built-in functionalities and innovative 'off-label' uses developed by our team during live projects.

Methods

- Significant AI updates were implemented over the past year (since July 2024) in the AI LR tool, AutoLit (Nested Knowledge), and tested in real-world projects.
- We systematically assessed Nested Knowledge across several live projects performed in 2025. The evaluation focused on four major domains of literature review work: abstract screening, full-text screening, data extraction, critical appraisal, and reporting. Both baseline (built-in) functionalities and novel extensions (off-label functionalities developed in-house) were considered.
- Performance was benchmarked against manual processes, with outcomes including time savings, accuracy of screening and extracted data, and user flexibility.
- Tool capabilities tested included:
  - Fully automated abstract and full-text screening using Criteria-Based Screening (with Smart Screener).
  - Abstract screening using Robot Screener (AI as a second reviewer).
  - Data extraction via Adaptive Smart Tags (ASTs).
  - Automated Critical Appraisal using validated templates (RoB2, JBI, AMSTAR2, NHLBI).
  - Off-label extensions, including table summarization, safety data monitoring (including Development Safety Update Report [DSUR] preparation), and automated registry/database identification.

Results

- The most impactful features were Criteria-Based Screening for automated abstracts and full-text screening and the automated extraction of text and tables from publications via Adaptive Smart Tags.
- Criteria-Based Screening in Nested Knowledge provides a structured alternative to traditional exclusion-reason-based screening by aligning decisions with protocol-defined Yes/No criteria. Reviewers can assess abstracts and full texts by answering predefined questions, ensuring consistency, transparency, and traceability throughout the review process, including assessing each criteria on each study rather than solely assigning a final exclusion reason. This functionality is particularly well-suited for protocol-driven systematic reviews that require reproducibility and auditability.
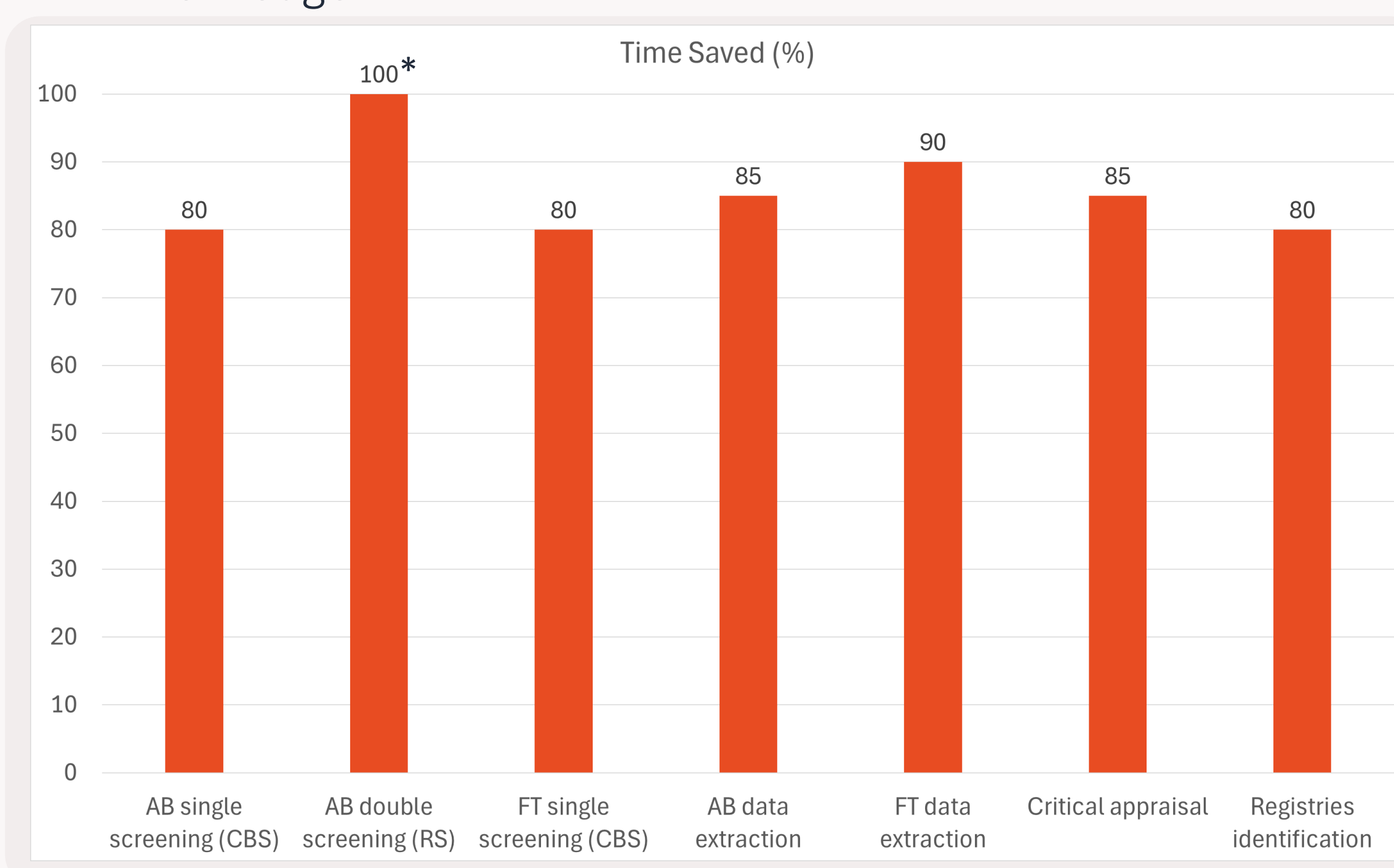
Results (*cont.*)

- Adaptive Smart Tags enabled PICOs and other key data to be accurately extracted into user-defined templates. Additionally, AutoLit enabled automated literature searches by inputting a research question into "Smart Search," eliminating the need for manual search strategy development.
- An AI Exploration feature, which derives Populations, Interventions/Comparators, and Outcomes (PICOs) from abstracts, facilitated rapid screening and proved highly effective.
- The tool also demonstrated flexibility in off-label applications: automated Critical Appraisal using RoB2 and JBI tools, and the generation of result summaries all showed high accuracy, outperforming human reviewers in some instances.
- Key results demonstrated both efficiency gains and accuracy improvements (Figures 1-3):
  - Abstract and full-text screening (automatic, single-screening): Criteria-Based Screening achieved ~95% accuracy while saving ~80% of time.
  - Abstract screening (AI-enhanced, double-screening): AI Robot Screener reduced time by ~50% across projects (~100% for one screener), with the largest exceeding 10,000 abstracts.
  - Automatic data extraction: Adaptive Smart Tags enabled extraction of patient characteristics, interventions, and outcomes with 90 - 100% accuracy for structured data; overall time savings ranged from 80 - 90% for both abstracts and the full texts.
  - Critical appraisal: Automated appraisal using JBI achieved 98–100% accuracy and saved ~100 hours in a 280-study project. RoB2 automation achieved >85% accuracy, with detailed justifications supporting rapid quality control.
  - Database/registry identification: Accuracy >90%, saving 80–85% of time compared to manual methods.
- Flexible framework of the tool enabled development of variety of „off-label" functionalities (Figure 4), for example:
  - SWOT Analysis – Automatic identification of treatment's strengths, weaknesses, opportunities, and threats from provided documents and literature to support data-driven decision-making and strategy development.
  - Proposals/Contracts – Rapid assessment of the inclusion rate based solely on the search strategy enabled precise budget calculations.

Conclusions

- The adoption of Nested Knowledge greatly improved the efficiency, quality, and reproducibility of literature reviews in our organization. Automation reduced workload by an average of 85% while maintaining or exceeding human accuracy. Key breakthroughs - Adaptive Smart Tags and Criteria-Based Screening - enabled automation of screening and data extraction.
- A major strength is flexibility: our team developed off-label functions (e.g., safety monitoring, automated summaries, registry mapping, critical appraisal), demonstrating the platform's adaptability to project needs.
- Although some limitations remain, current AI performance already surpasses manual methods in speed and reliability. These advancements can accelerate evidence generation and decision-making, supporting faster patient access to effective medical technologies.

Figure 2. Results snapshot

Figure 1. Average Time Savings Achieved with Nested Knowledge AI

Time Saved (%)

Critical Appraisal - measured on JBI

* - 100% of screening time saved for one of the reviewer

Legend

- AB – Abstract
- CBS - Criteria-Based Screening
- DSUR - Development Safety Update Report
- FT – Full text
- RS – Robot Screener
- SWOT - Strengths, Weaknesses, Opportunities, and Threats

Automation reduced workload by an average of

85%

in key steps, while maintaining or surpassing human-level accuracy.

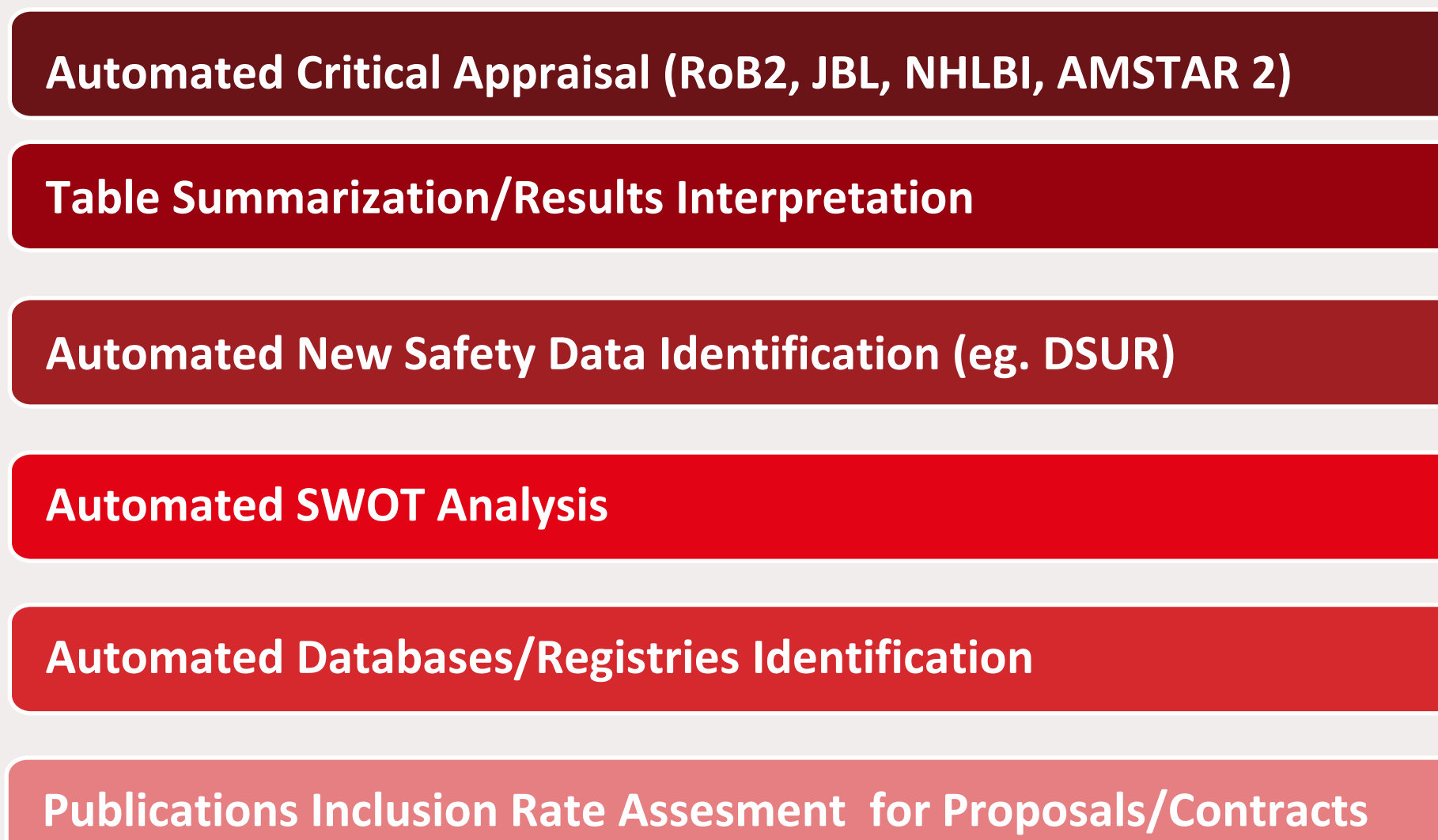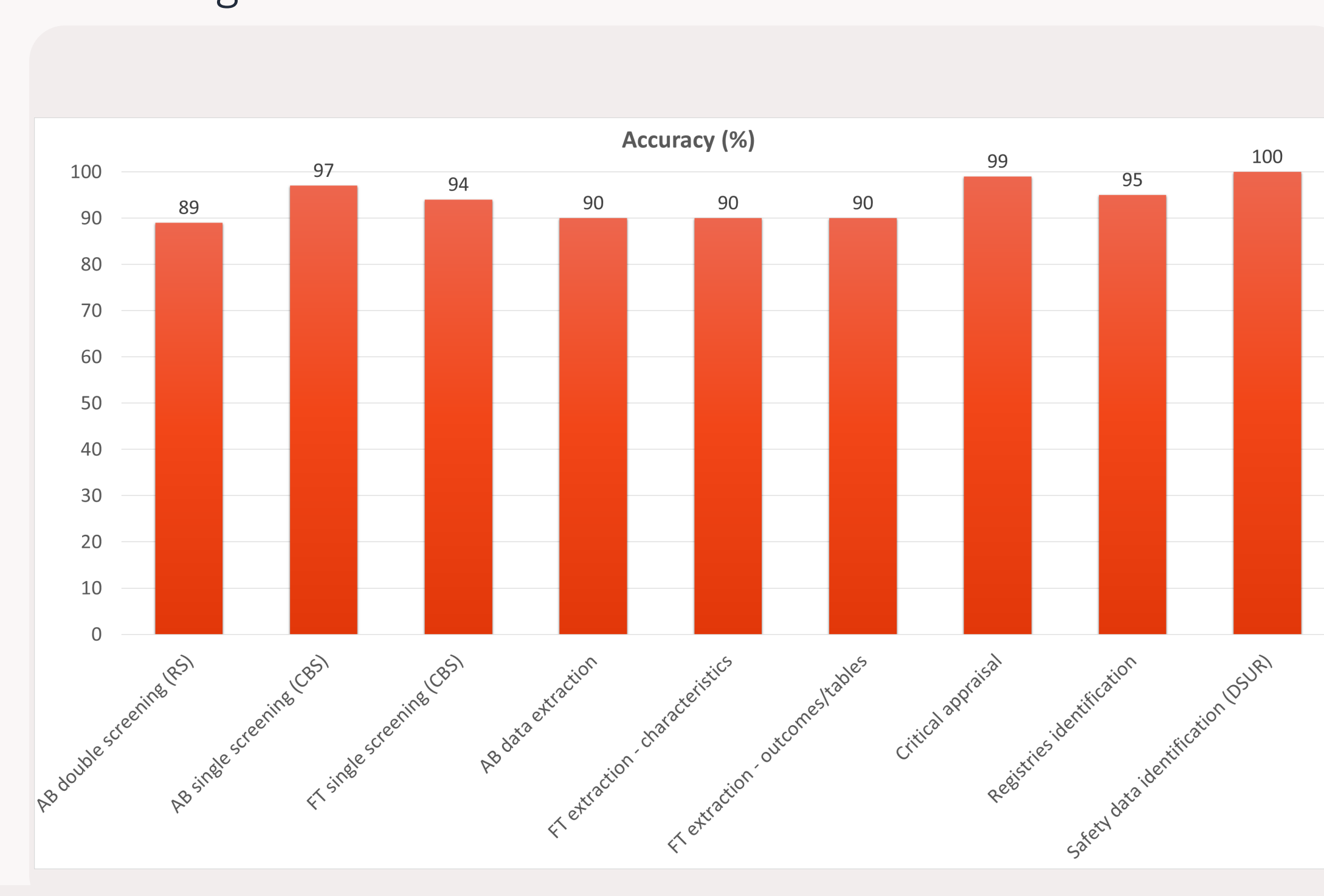Figure 4. Flexibility of Nested Knowledge Tool (off-label custom functionalities)

Automated Critical Appraisal (RoB2, JBL, NHLBI, AMSTAR 2)

Table Summarization/Results Interpretation

Automated New Safety Data Identification (eg. DSUR)

Automated SWOT Analysis

Automated Databases/Registries Identification

Publications Inclusion Rate Assesment for Proposals/Contracts

Figure 3. Accuracy of Automated AI Tasks in Nested Knowledge

Accuracy (%)

Critical Appraisal - measured on JBI