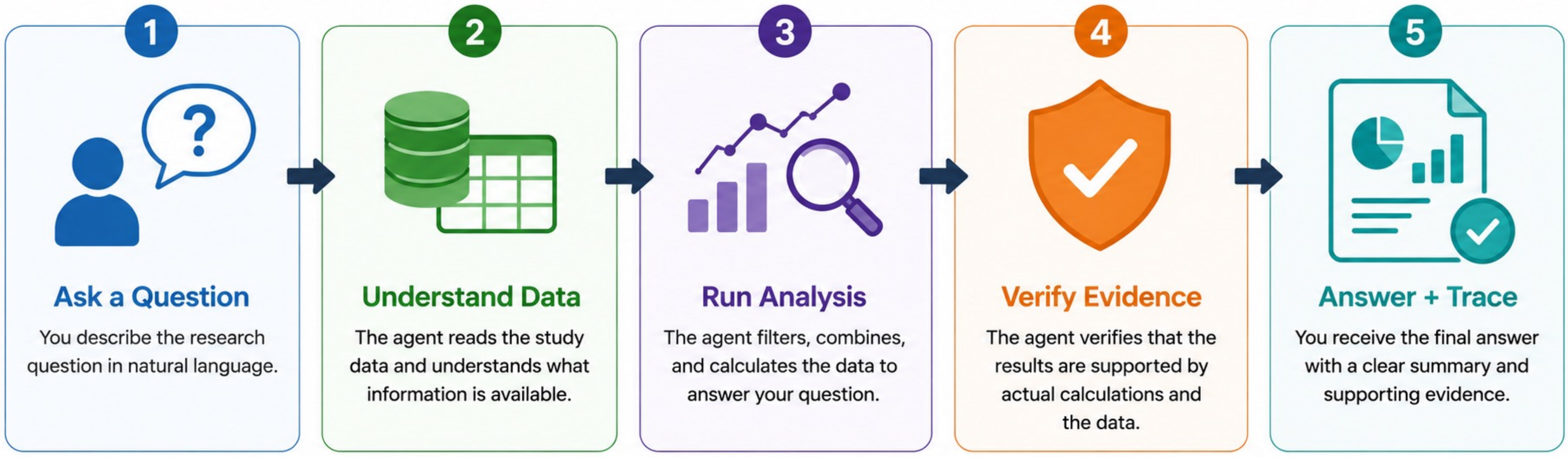
A Multi-Agent QA System for Verifiable Clinical Study Data Analysis

Siyao Wang, Florian Guitton, Kai Sun, Shuojie Fu, Mark Kennedy

Imperial College London



Introduction

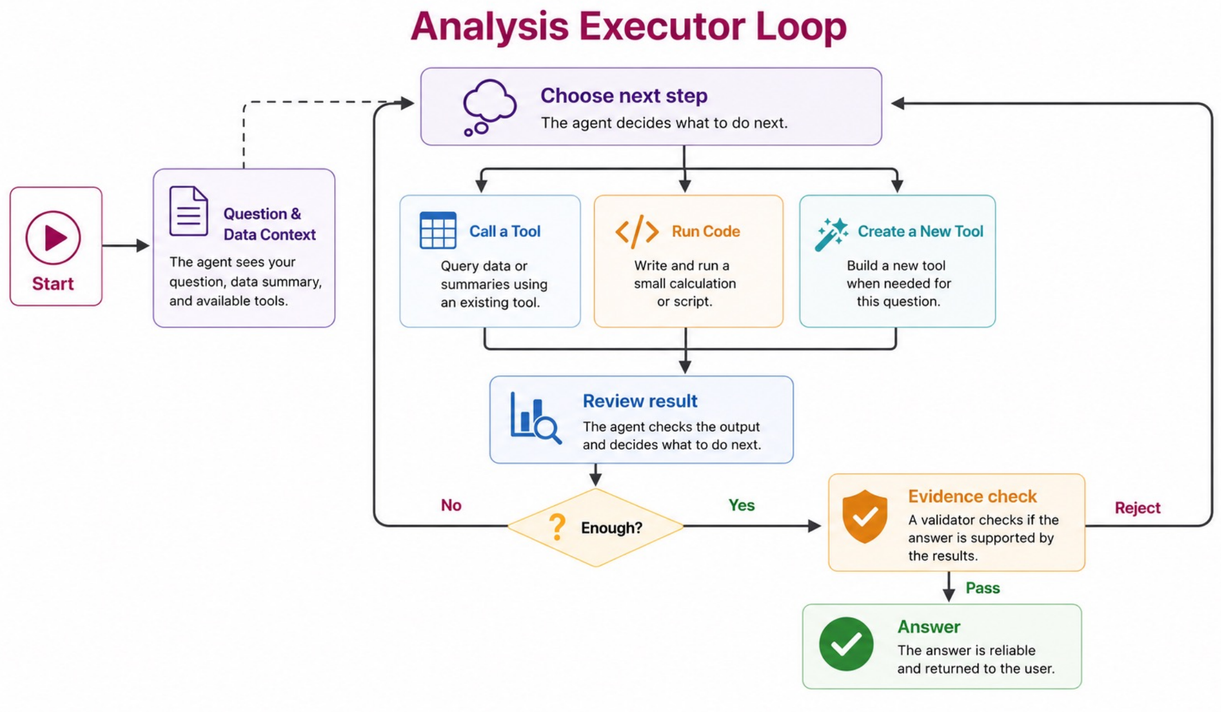
- The IDEA-FAST Data Management Platform supports governed access to complex clinical and digital health study data. These data are multi-table and analysis-ready, but simple research questions still often require joining, filtering, aggregation, and validation.
- Direct language-model answers are unreliable for numerical analysis unless results are grounded in executed calculations.

Objective

- Enable natural-language analytical QA over structured clinical research data while preserving reliability, traceability, and reproducibility.

Methodology

- We build a multi-agent research analyst, **MARA (Multi-Agent Research Analyst)**, to handle structured clinical study questions. MARA is composed of four cooperating agents
  - **Question Interpreter —** Understands the research question and translates it into a structured analytical task
  - **Data Inspector —** Reads the relevant tables and column descriptions to identify required variables and constraints.
  - **Analysis Executor —** Runs the necessary operations (filtering, joining, grouping, diary aggregation, visit comparisons) in iterative steps.
  - **Result Validator — Ensures that numerical outputs are supported by the executed calculations before they are reported.**



Experiments & Results

- **Question types:** cohort summaries, clinical assessments, PRO summaries, diary aggregation, missingness checks, visit-based comparisons, and multi-table analyses.
- **Evaluation:** 336 analytical questions with pandas-generated reference answers; best archived run achieved **83.3% accuracy**.
- **Example QA:**
  - Q: Among participants with Calprotectin severity data at both Visit 1 and Visit 7, how many changed from Active at Visit 1 to Inactive at Visit 7?
    - *Agent: 17 participants changed from Active at Visit 1 to Inactive at Visit 7 (17/186, 9.14%).*
    - *Agent: The total number of participants with Calprotectin severity data at both Visit 1 and Visit 7 was 186.*
  - Q: In participants using the high-frequency eDiary, (1) does self-reported physical fatigue differ between morning and evening on the same day, and (2) how is moment-to-moment eDiary fatigue associated with visit-level FACIT-Fatigue scores?
    - *Agent: Physical fatigue ratings were higher in the evening than in the morning (evening mean ≈ 2.58 vs morning ≈ 2.17 on the eDiary scale; Welch two-sample t-test $p \approx 5.6 \times 10^{-25}$), consistent with a clear diurnal pattern (paper: $P < .001$).*
    - *Agent: After linking participants (n = 108 with both summaries), mean eDiary physical fatigue was inversely associated with baseline FACIT-F total score (Spearman □ ≈ □ 0.52, two-sided $p \approx 9.9 \times 10^{-9}$): higher momentary diary fatigue aligned with lower FACIT-F (less fatigue on the questionnaire scale). Direction matches the paper's significant association; magnitude differs from the published □ ≈ 0.46 because of sample and aggregation choices.*