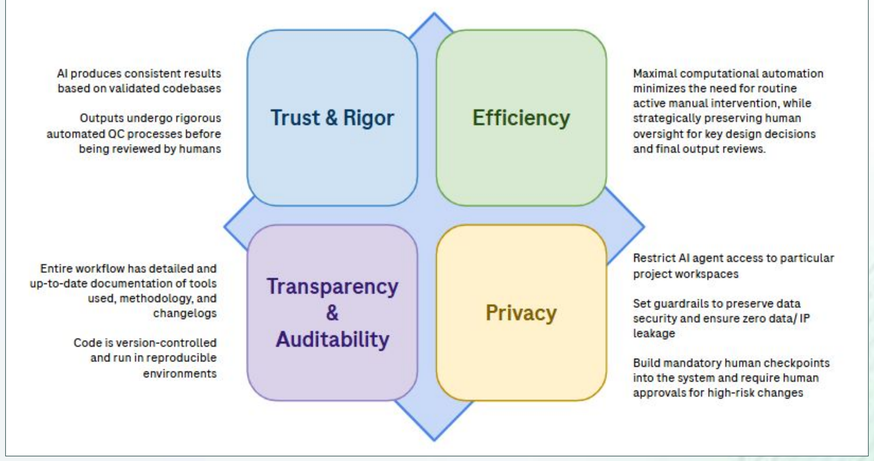
Towards Autonomous Real-World Evidence Generation: Agentic AI End-to-End Workflows

Jiying Zou, Real World Data Scientist @ Genentech

Background & Motivation

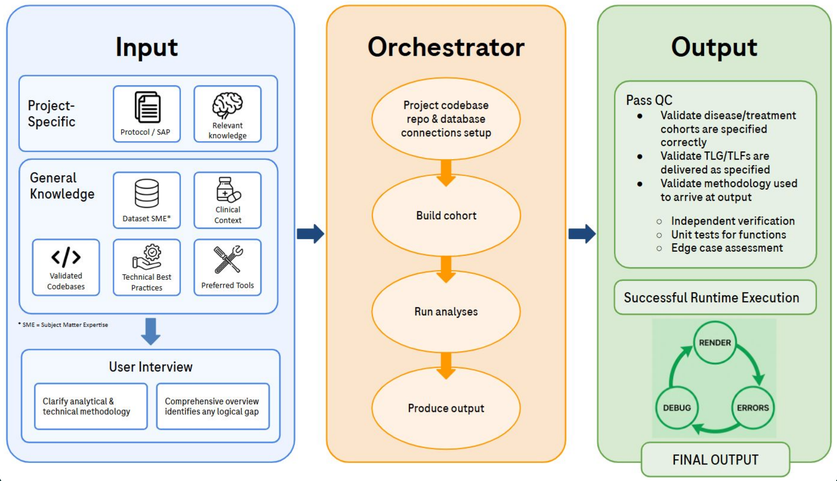
- **The Current Bottleneck:** Real-World Evidence (RWE) generation remains a heavily manual, resource-intensive endeavor.
  - Study timelines can be weeks or months due to tedious manual coding and constant technical and scientific iterations and back-and-forth.
  - Drains technical capacity through repetitive setup, runtime error debugging, and manual double-coding for quality control (QC).
- **Why Traditional Automation Fails:** Conventional non-AI approaches (e.g., static reports and dashboards) lack the dynamic flexibility required for highly nuanced, context-specific clinical projects with evolving specification requirements and complex design decisions.
- **The Agentic Opportunity:** Modern AI introduces flexible task orchestration capable of navigating complex clinical pipelines in **FAST**, **RELIABLE**, and **SECURE** ways.
  - Employs domain-specific "skills," ingests broader medical context, adapts legacy codebases to novel protocols, and autonomously self-corrects runtime errors.
- **Transforms the Clinical Workflow by:**
  - *Evolving Analytics Roles:* Shifts programmers from routine code executors to strategic study designers.
  - *Compressing Timelines:* Deliver more evidence with stronger scientific and technical rigor to drive faster, more meaningful impact for patients.

Development Framework & Principles



Design Workflow & Implementation

- **Core Functionalities:**
  - *New Dataset QC:* Automated technical checks, basic SME validations, and historical data comparisons for new datasets. Build domain-specific expertise in new datasets in a fast, scalable, reliable way.
  - *E2E Project Execution:* Move from initial protocol to final report generation quickly and accurately, keeping human-in-the-loop for key design decisions and high-stakes approvals.



CASE STUDY

The Setup:

- Two different AI workflows were tested on the same tasks (involving I/E criteria coding, cohort creation, creating an attrition table, and performing several descriptive analyses). Outputs and code are compared to the manually double-coded gold standard.

RESULTS

(Time calculated as time from first protocol receipt to final output acceptance, including output review and scientific/technical iteration with stakeholders)

Architecture	Time Spent	Token Cost	Outcome Validity
Manual (control)	1 month	$0	10/10 (gold standard, double-coded)
Skills only (no agents)	1 week (15 mins to first draft)	$9	9/10 (same methodologies but used different data source)
Prompt Chaining (agentic)	1 week (1.5 hours to first draft)	$10	9/10 (identical results to Skills Only architecture, more detailed methods log)

Conclusion

- Transitioning RWE generation from manual human-driven scripting to human-approval-based **Agentic AI Workflow** orchestration **cuts down cycle times by ~75%+** without sacrificing rigorous statistical standards.
- By grounding the system in strict engineering guardrails, **more evidence is generated with greater speed while not compromising on safety or scientific & technical rigor**.

Future Scope and Outlook

- **Automate Expanding Knowledge Ingestion:** Move from localized reference codebases □ dynamic API integrations that instantly pull database/project/codebase updates to add to knowledge repository and refine skills & agents. Unlock previously isolated data sources that external AI systems have never had access to
- **Agentic Design Improvements:** Improve architecture of agentic workflows to further increase speed while balancing cost and maintaining output integrity
- **Better Benchmarking:** More comprehensive tool performance assessment suite to test speed, reliability, and safety
- **Cross-TA Compatibility:** Expand horizontal coverage across TAs of generalizable skills to minimize reinventing the wheel

WE APPRECIATE YOUR FEEDBACK!

Tell us what resonates with you, and let's build the future of evidence generation TOGETHER