AI Agent for Automated Literature Validation of Experimental Magnetic Structures

Megan Sun [1, 2], Cheng Peng [1, 2]

[1] SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

[2] Stanford University, Stanford, CA 94305, USA

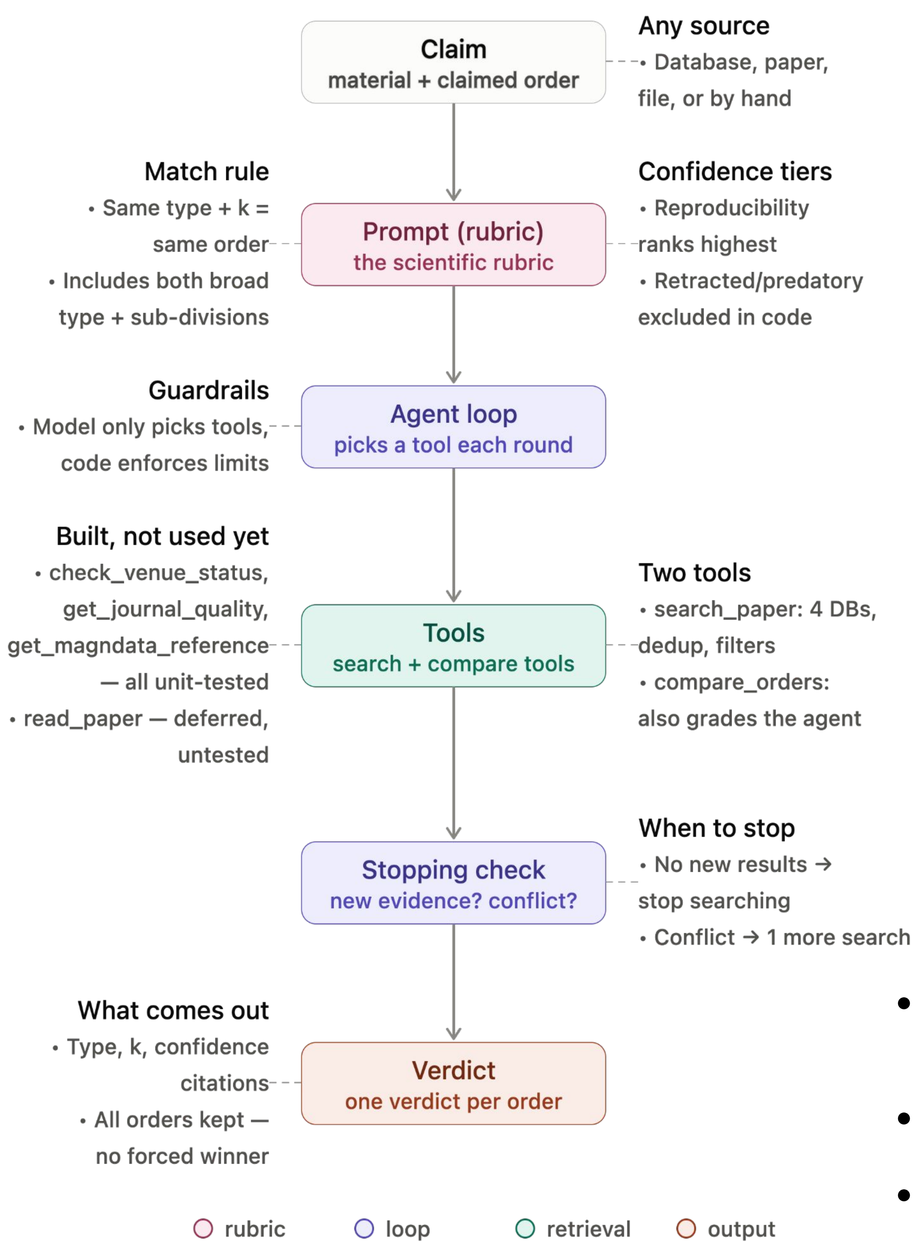
SUMMARY OF PROJECT

- Databases, papers, and online structure files all contain claims about materials' magnetic order (can become **outdated** or **wrong**)
- **Re-auditing at scale** is useful yet impractical because it requires expert review
- Agent: material + claimed order → literature search → confidence score + citations, **verifiable in minutes**
- **Domain-general**: the same approach could apply to crystal structures, phase diagrams, Tc values, protein annotations, etc.; magnetic structures are the first use case

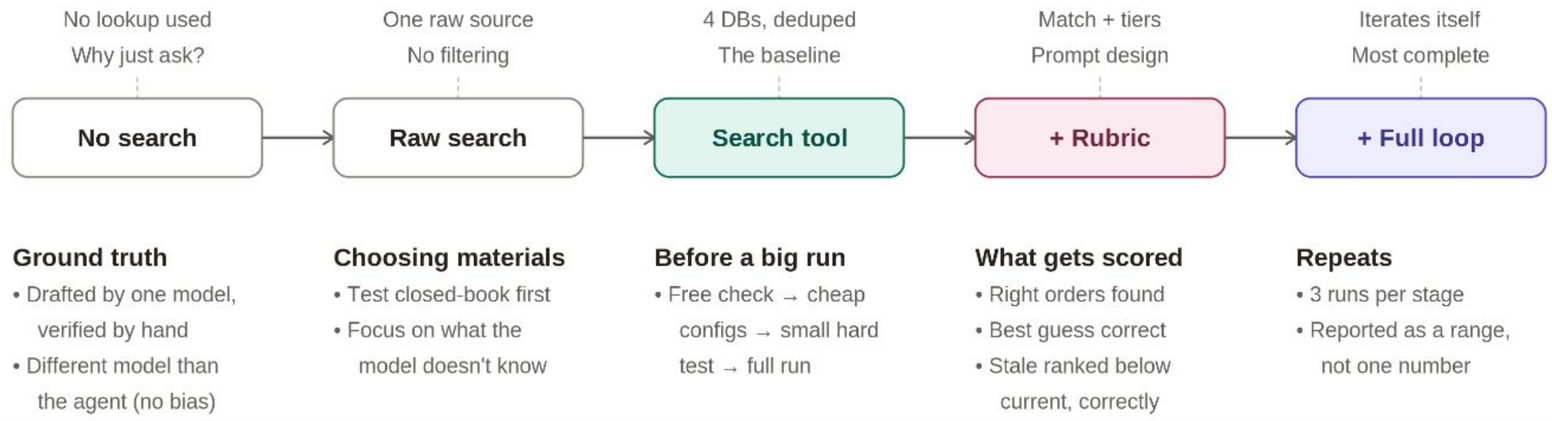
BACKGROUND & MOTIVATION

- Started as a 7-stage DFT/HPC pipeline; noticed a simulation can converge to the wrong state from a bad magnetic-structure guess (**converged ≠ correct**)
- Catching bad claims upstream **prevents wasted simulation**
- **~2 days → <30 minute**: expert magnetic-order validation is slow; the agent automates the same literature check
- Manually auditing thousands of database entries is impractical; the agent makes **large-scale validation** feasible
- Why an agent, not a pipeline or plain LLM:
  1. Requires **judgment**: weigh recency vs. replication, reconcile different descriptions, and distinguish new phases from corrections
  2. Requires **adaptive search**: what to search, how many papers to check, and when to stop depend on the evidence found
  3. Requires **grounded evidence**: every conclusion must trace to a retrieved paper (not model memory), requiring tools and code-level enforcement

METHOD / ARCHITECTURE: AGENT DESIGN



METHOD / ARCHITECTURE: EVALUATION



RESULTS

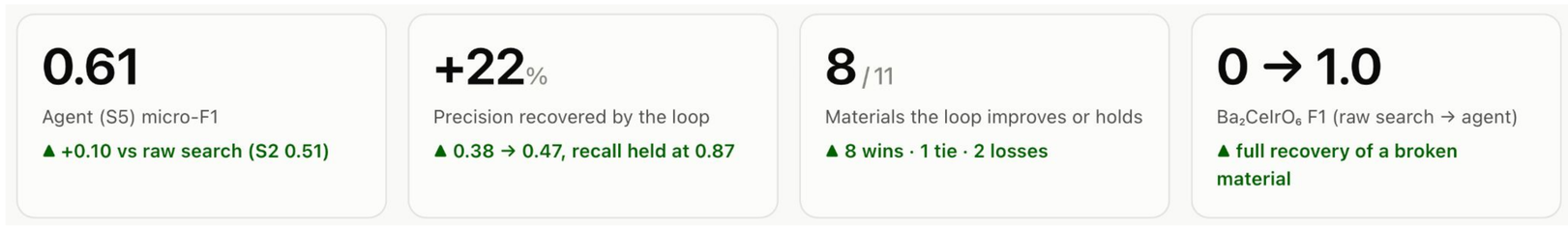


Figure 1. Headline Results

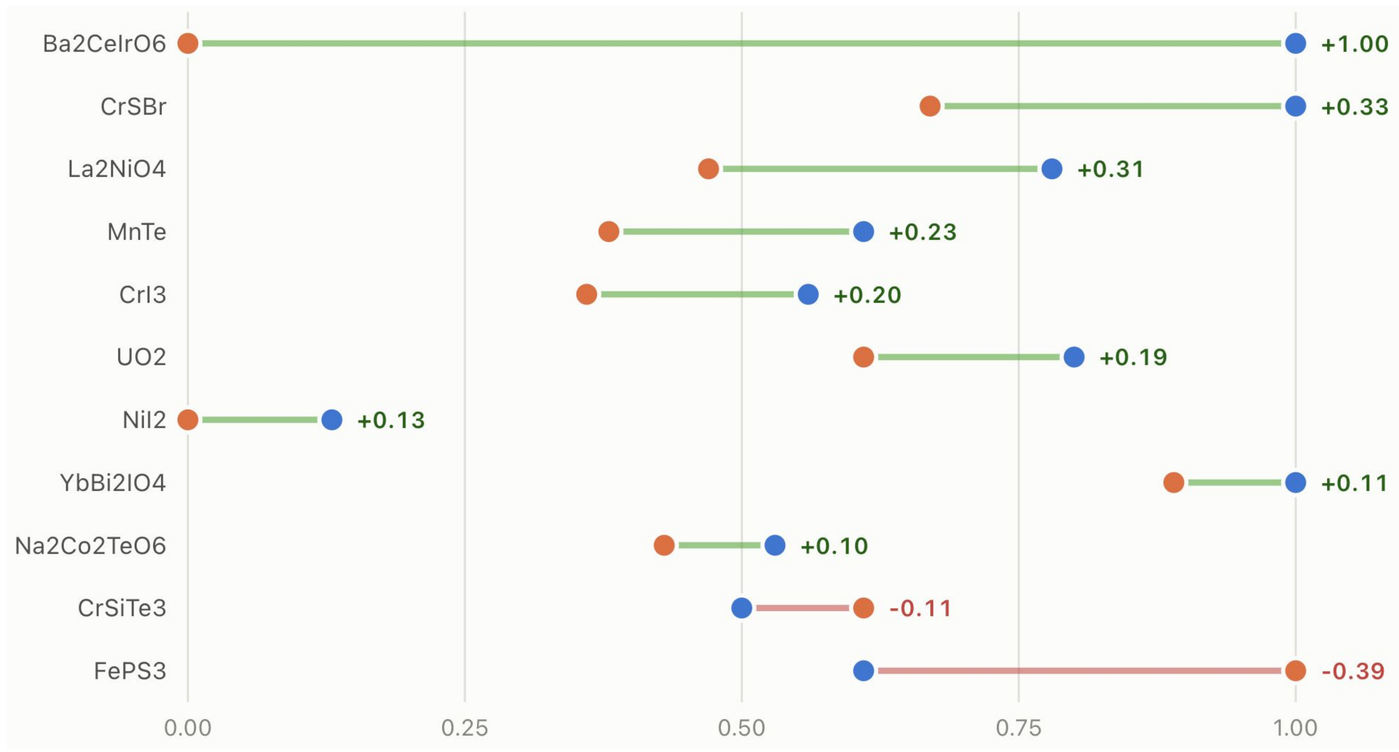


Figure 2. F1 scores for database raw search (orange) v.s. with full agent loop (blue), across all 11 materials tested.

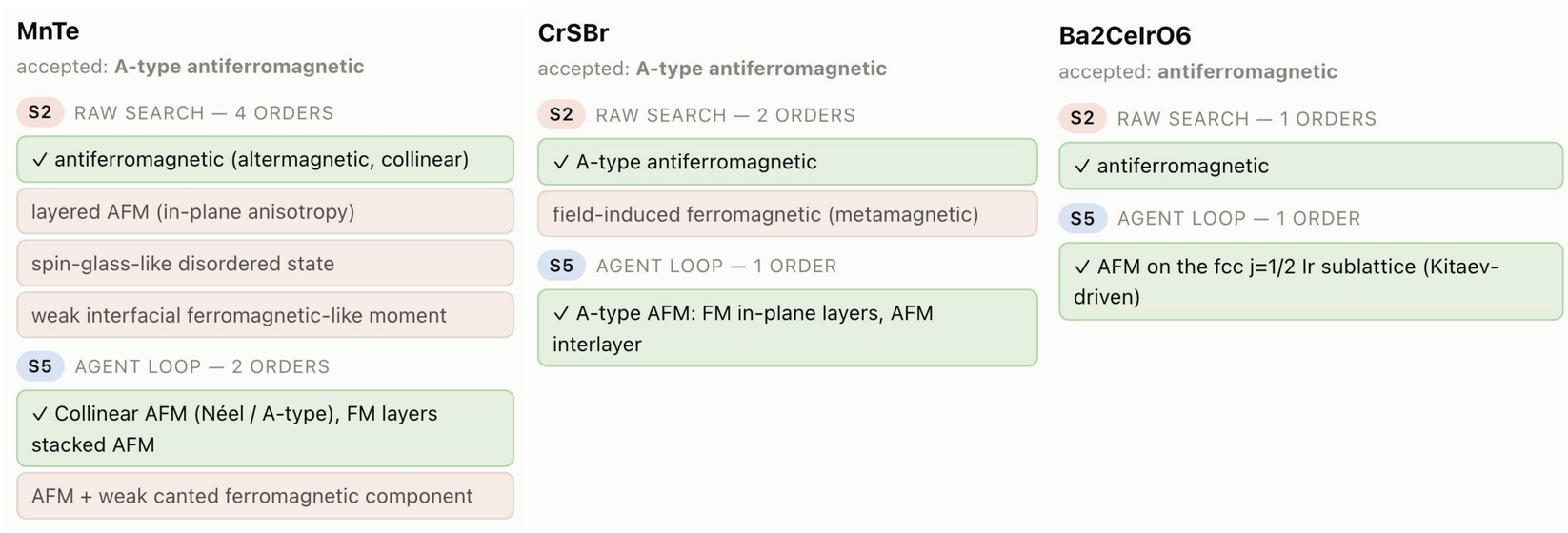


Figure 3. Example materials, raw search vs. agent results. Green = accepted structure; orange = noise filtered out by the agent.

LIMITATION & NEXT STEPS

- Eval set is small and deliberately adversarial (11 hardest / search-fragile materials) → results are a stress test, not an overall accuracy number (ex. staleness tested on only 2 materials)
- Several built tools pass unit tests but aren't yet exercised in the live loop or eval
- Ground truth had one model draft and one human review; no second reviewer
- Grow the staleness set and broaden past the adversarial 11 for a realistic typical-case benchmark
- Bring read_paper and the vetting tools into the eval staircase
- Build an accompanying front end for easier user interaction
- Add functionality for manual PDF upload, expert-judge review, and similar human-in-the-loop features

Acknowledgements: Part of this work was supported by the US Department of Energy, Office of Science, Basic Energy Sciences under the BES AI Pathfinder project MAIQMag: Multimodal AI for two-dimensional Quantum Magnets. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0034968.