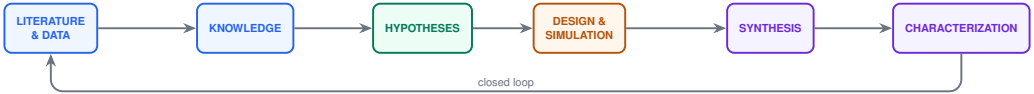
The AI/ML & LLM Perspective

Knowledge Extraction · Hypotheses Formation · The AI Scientist

Subhankar Mishra
School of Computer Sciences, NISER Bhubaneswar

Formulating a National Research Roadmap for Accelerated Materials Design:
From Generative AI to Self-Driving Labs
ICTS Bengaluru · August 31 – September 1, 2026

Scope: the intelligence layer of the loop

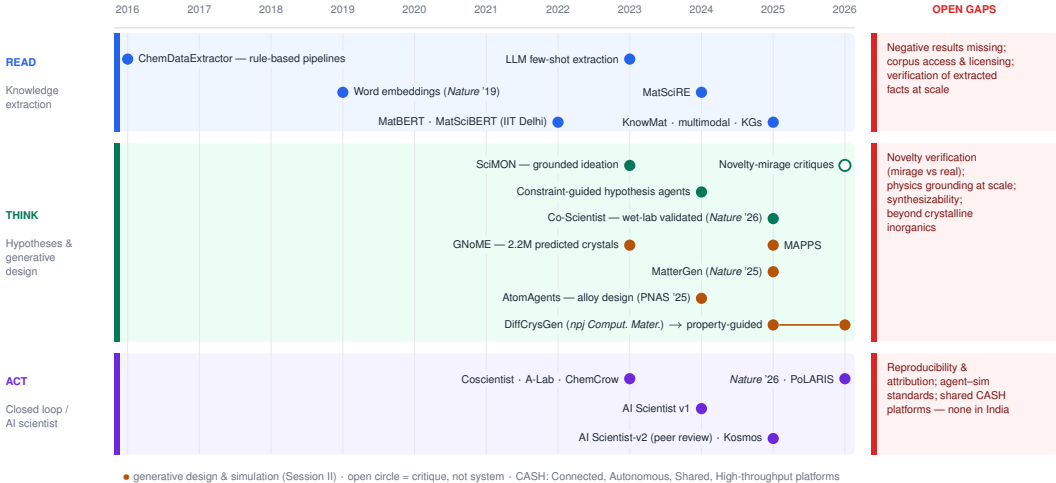


- Robotics automates the *hands* (Sessions IV–V); simulation automates the *calculations* (Session III)
- This talk: automating the *reading, reasoning, and deciding*

Part	Task	Technical question
1 · READ	Knowledge extraction	Literature → structured, queryable records?
2 · THINK	Hypotheses formation	Novel + grounded + testable proposals?
3 · ACT	The AI scientist	Full cycle, closed loop, verified output?

Generative design & simulation — Session II's territory — rides along in the THINK lane of the field map (orange), visited here only through a brief case study.

The field at a glance



● generative design & simulation (Session II) · open circle = critique, not system · CASH: Connected, Autonomous, Shared, High-throughput platforms

Part 1 · READ

Knowledge Extraction

Task: literature → structured records of composition, processing, property.

READ · The extraction problem

- Millions of materials papers, thousands more weekly; data in **prose, tables, figures, SI**
- Curated databases capture a thin slice — synthesis routes and processing conditions mostly absent
- **Negative results are not in the record at all** — the class balance ML needs is missing

What “solved” would look like: any paper $\rightarrow$ machine-readable (composition, process, property, conditions, provenance) tuples, at corpus scale, with quantified error.

READ · What has been done

- **ChemDataExtractor** (2016–): rule/grammar pipelines; auto-built databases, e.g. ∼290k-record battery-materials DB from ∼230k papers
- **Word embeddings** (Tshitoyan et al., *Nature* 2019): unsupervised word2vec on 3.3M abstracts ranked thermoelectrics *years before* their experimental report
- **Domain BERTs**: MatBERT; **MatSciBERT (IIT Delhi, 2022)** — pretrained on ∼150k materials papers; outperforms SciBERT on all tasks — NER, relation & abstract classification (SOTA at release)
- **MatSciRE** (2024): pointer networks, entity–relation F1 0.771 vs 0.716 for ChemDataExtractor on battery materials
- **LLM era** (2023–): zero-/few-shot extraction to schema-constrained JSON (KnowMat); multimodal extraction from **tables and figures**; RAG + knowledge-graph corpora

READ · What has not been done

- **?** **Verification at scale**: a hallucinated value is typographically identical to a real one; no accepted method to certify extracted numbers without human spot-checks
- **?** **Context fidelity**: units, measurement conditions, sample-specific caveats $\rightarrow$ silent errors that propagate into training sets
- **?** **Corpus access**: text-and-data-mining rights are licensed per publisher, per institution; no country-scale materials corpus exists anywhere
- **?** **Negative results**: cannot be extracted — they were never published

READ · Where India stands

Done

- MatSciBERT (IIT Delhi) — internationally used domain model for materials NLP
- Strong NLP/IR groups; sovereign LLM stack exists (Sarvam, BharatGen)

Not done

- No national materials corpus, knowledge graph, or shared extraction pipeline
- No nationally negotiated text-mining access; every group re-solves licensing alone
- No mechanism to capture negative results from our own labs (links to Session VI: Materials data)

Part 2 · THINK

Hypotheses Formation

Task: from retrieving what is known to proposing what might be true.

THINK · What counts as a hypothesis

A machine-generated hypothesis is useful only if it is simultaneously:

- **Novel** — not a restatement of training data
- **Grounded** — consistent with known physics and chemistry
- **Testable** — maps to a runnable simulation or a feasible experiment
- **Prioritized** — ranked against the cost of testing it

LLMs are recombination engines over the written record; the empirical question is whether recombination + reasoning clears all four bars at once.

THINK · What has been done

Google AI Co-Scientist (Gemini; *Nature*, May 2026)

- Multi-agent generate → debate → Elo-rank → evolve; quality scales with test-time compute
- **Wet-lab validated**: AML drug-repurposing candidates confirmed in vitro; novel liver-fibrosis targets confirmed

Materials-specific

- **AtomAgents** (PNAS 2025): multi-agent alloy design with MD in the loop — alloys outperforming pure-metal baselines
- **MAPPS** (2025): planner + force-field foundation model + human mediator
- SciMON: literature-grounded ideation, novelty-optimized

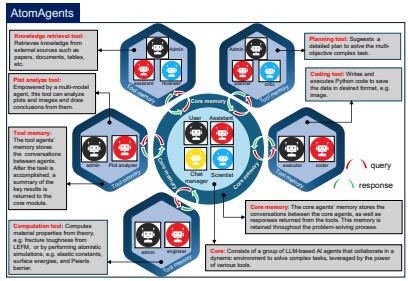


Figure: Ghafarollahi & Buehler, PNAS 122 (2025); arXiv:2407.10022

THINK · What has not been done

- **? Novelty evaluation**: in controlled human studies, LLM ideas score *higher on novelty* but *lower on feasibility* than expert ideas — and LLM-as-judge systematically inflates both
- **? No materials benchmark**: no accepted test set for hypothesis quality → system claims are not comparable
- **? Validation throughput**: each validated hypothesis still costs an expert-guided lab campaign; nobody has closed hypothesis → test automatically at scale

Working mitigation across all serious systems: **LLM proposes → physics + data verify → human disposes.**

THINK · Where India stands

Done

- Growing ML-for-materials groups; strong DFT/MD community to supply the physics oracle
- Large expert base — the human-verification layer these systems require
- Generative inverse design demonstrated at Indian institutions (brief case study next)

Not done

- No Indian hypothesis-generation system for materials — though the cost of entry is low
- Such a system's two dependencies are the two we lack: the corpus/KG (Part 1) and shared autonomous simulation (Session III)

Case study · DiffCrysGen

Mal, Ahmed, Jami, Mishra, Sen (NISER · HRI) — *npj Comput. Mater.* (2026); arXiv:2510.12329, 2607.21849

- Score-based diffusion; atom types + coordinates + lattice in **one unified process** — no task-specific priors, no decoupled modules
- Trained on 151k relaxed structures (Alexandria, $\leq$20 atoms/cell)
- $\sim$**300 structures/s** vs 0.22 (MatterGen) / 3.5 (DiffCSP) — 2–3 orders of magnitude faster sampling; S.U.N. rate 51.5 % vs 79.9 % / 59.4 % — **throughput traded for hit rate**
- Applied to **rare-earth-free magnets**: DFT-validated stable ferromagnets with high M_s and large anisotropy, plus metallic antiferromagnets
- Extended to **property-guided inverse design** (adapters + classifier-free guidance): 12.3 % / 3.9 % end-to-end success for stable magnets ($M_s \geq 1$ T) / hard materials, after MLIP + phonon validation

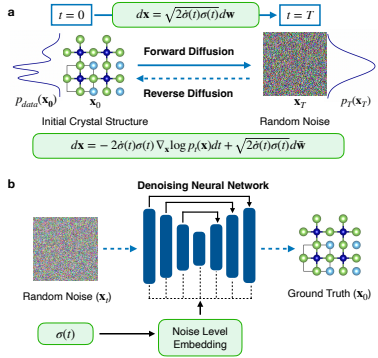


Figure: Mal et al., arXiv:2510.12329

Part 3 · ACT

The AI Scientist

Task: chain read → think → simulate → experiment, with verification.

ACT · Levels of autonomy

Level	Role	Demonstrated by
1	**Assistant** — answers, summarizes	RAG chatbots, literature QA (deployed, routine)
2	**Collaborator** — proposes, plans	Co-Scientist, ChemCrow (validated, 2023–26)
3	**Scientist** — executes full cycle	AI Scientist-v2, Kosmos, A-Lab loops (partial)

Level 2 $\rightarrow$ 3 is not a modelling gap: it requires coupling to autonomous simulation and labs — the subject of this meeting.

ACT · What has been done, 2023–24

- **Coscientist** (Boiko et al., *Nature* 2023): GPT-4 planned and *executed* Pd-catalyzed cross-couplings (Suzuki, Sonogashira) on liquid-handling hardware
- **A-Lab** (*Nature* 2023): 17-day autonomous run, claimed 41/58 novel inorganic targets synthesized — phase-identification later *disputed* (PRX Energy 2024): a lesson in autonomous verification
- **GNoME** (*Nature* 2023): GNN-predicted 2.2M crystals, $\sim$381k on the convex hull — an order of magnitude more than all prior human discovery
- **AI Scientist v1** (2024): full paper pipeline at $\sim$\$15/paper — fluent, shallow, error-prone

ACT · What has been done, 2025–26

- **AI Scientist-v2**: first fully AI-generated paper past peer review (ICLR 2025 *workshop*, avg. 6.33; withdrawn by agreement); framework in *Nature*, Mar 2026
- **Kosmos** (Edison Scientific, 2025): $\sim$1,500 papers + $\sim$42,000 lines of analysis code per run; 7 expert-validated discoveries; $\geq$79 % of statements reproducible
- **PoLARIS** (*Nat. Commun.* 2026): self-driving lab — lead-free emissive nanoplatelets in 12 h · **Industry**: Periodic Labs in talks at $\sim$\$7B valuation; Lila Sciences at $\sim$\$8.5B

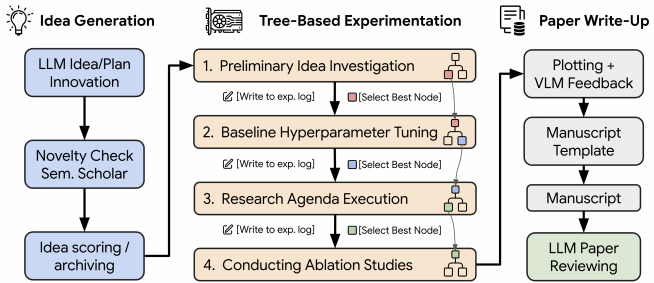


Figure: AI Scientist-v2 pipeline — Yamada et al., arXiv:2504.08066

ACT · What has not been done

- **?** No AI-led result at *main-conference / journal* standard without heavy human curation; early systems fabricated citations and numbers
- **?** **Verification, reproducibility, attribution**: unsolved as a class ($\geq$79 % reproducible $\Rightarrow$ up to 1 in 5 statements wrong)
- **?** **The loop is open**: no system yet couples literature $\rightarrow$ hypothesis $\rightarrow$ simulation $\rightarrow$ synthesis $\rightarrow$ characterization end-to-end
- **?** Physical experiment remains the rate limiter — **the lab, not the LLM**

ACT · The required stack

1. **Models** — frontier or fine-tuned domain LLMs
2. **Knowledge** — extracted, verified literature + databases
3. **Tools** — simulation codes, databases, instruments as APIs
4. **Compute** — training and inference at scale
5. **Verification** — physics checks, human review, provenance
6. **Labs** — autonomous synthesis & characterization (Sessions IV & V)

Every working system above instantiates all six layers. No single group builds all six — historically this is the profile of **national infrastructure** (synchrotrons, HPC).

ACT · Where India stands (2026)

Done

- **Compute**: IndiaAI Mission (INR 10,372 cr) — 45,000+ GPUs deployed, 237+ subsidized projects; target ∼100k by end-2026
- **Models**: Sarvam 30B/105B open-sourced (Apache 2.0, Mar 2026); BharatGen Param2 (17B, 22 languages)
- **Funding channel**: ANRF AI-SE — consortium grants to INR 30 cr; NVIDIA support for grantees
- **People & results**: DFT/MD community; MatSciBERT; generative inverse design (DiffCrysGen)

Not done

- No science/materials focus in any sovereign-model effort; layers 2, 5, 6 of the stack absent
- Capabilities exist in isolation — **no shared agent–simulation–lab infrastructure**

Fine-tune, or build our own?

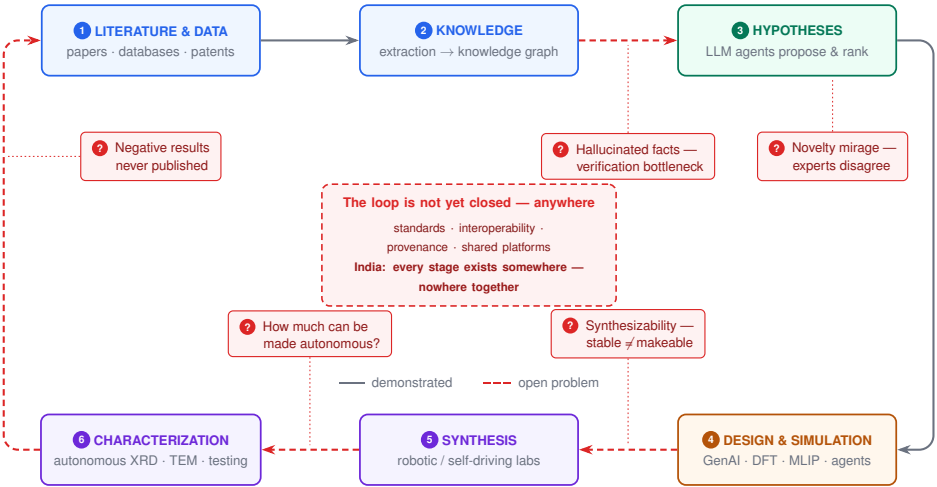
→ Discussion 1.1

A layered answer, by cost–benefit:

- **Use** frontier models via API where data sensitivity permits — fastest for agents & extraction
- **Fine-tune** open models (incl. Sarvam-class) on a national materials corpus — sovereignty + domain gain at modest cost; *same recipe our adapter fine-tuning already uses at small scale*
- **Build** from scratch only where India holds unique data leverage — national experimental archives, negative results

Models depreciate in $\sim$2 years; corpora, knowledge graphs, verification pipelines, and trained people appreciate.

Summary · The loop and its open problems



Takeaways

- **Extraction works**; verification of extracted facts does not, yet — and the corpus problem is policy, not science
- **Hypothesis generation is wet-lab real** (Co-Scientist, AtomAgents) but novelty claims need expert-referenced evaluation
- **Generative inverse design is demonstrated in India** — with honest success-rate baselines to build on
- **The AI scientist is partial**: refereed milestones exist (Nature 2026), fully closed loops do not; the lab is the rate limiter
- India has **compute, models, funding, people**; it lacks the **knowledge, verification, and lab layers** — integration is the gap

Thank you

Subhankar Mishra
smishra@niser.ac.in · School of Computer Sciences, NISER Bhubaneswar

References

READ

Swain & Cole, *J. Chem. Inf. Model.* 56, 1894 (2016) — ChemDataExtractor
Tshitoyan et al., *Nature* 571, 95 (2019) — word embeddings
Gupta et al., *npj Comput. Mater.* 8, 102 (2022) — MatSciBERT
Mullick et al., arXiv:2401.09839 (2024) — MatSciRE
KnowMat, chemRxiv (2025); *Integr. Mater. Manuf. Innov.* (2026)

THINK

Gottweis et al., arXiv:2502.18864; *Nature* (2026) — AI Co-Scientist
Si, Yang & Hashimoto, ICLR 2025 — LLM vs expert ideation study
Ghafarollahi & Buehler, *PNAS* 122 (2025) — AtomAgents
MAPPS, arXiv:2506.05616 (2025) · Wang et al., ACL 2024 — SciMON

GENERATIVE DESIGN

Mal, Mishra & Sen, *npj Comput. Mater.* (2026) — DiffCrysGen
Mal, Ahmed, Jami, Mishra & Sen, arXiv:2510.12329 (2025/26) — RE-free magnets
Mal, Mishra & Sen, arXiv:2607.21849 (2026) — property-guided inverse design

ACT

Boiko et al., *Nature* 624, 570 (2023) — Coscientist
Szymanski et al., *Nature* 624, 86 (2023) — A-Lab
Leeman et al., *PRX Energy* 3, 011002 (2024) — A-Lab critique
Merchant et al., *Nature* 624, 80 (2023) — GNoME
Zeni et al., *Nature* 639, 624 (2025) — MatterGen
Lu et al., arXiv:2408.06292 (2024) — AI Scientist v1
Yamada et al., arXiv:2504.08066; *Nature* (2026) — AI Scientist-v2
Kosmos, Edison Scientific technical report (2025)
PoLARIS, *Nat. Commun.* 17 (2026)
Bran et al., *Nat. Mach. Intell.* 6, 525 (2024) — ChemCrow

INDIA / POLICY

IndiaAI Mission, MeitY/PIB updates (2025–26)
Sarvam 30B/105B (Mar 2026); BharatGen Param2 (Feb 2026)
ANRF AI for Science & Engineering (AI-SE) call (2025)