AI for Theoritical Discovery

——starting from Olympiad level

Xiang Li

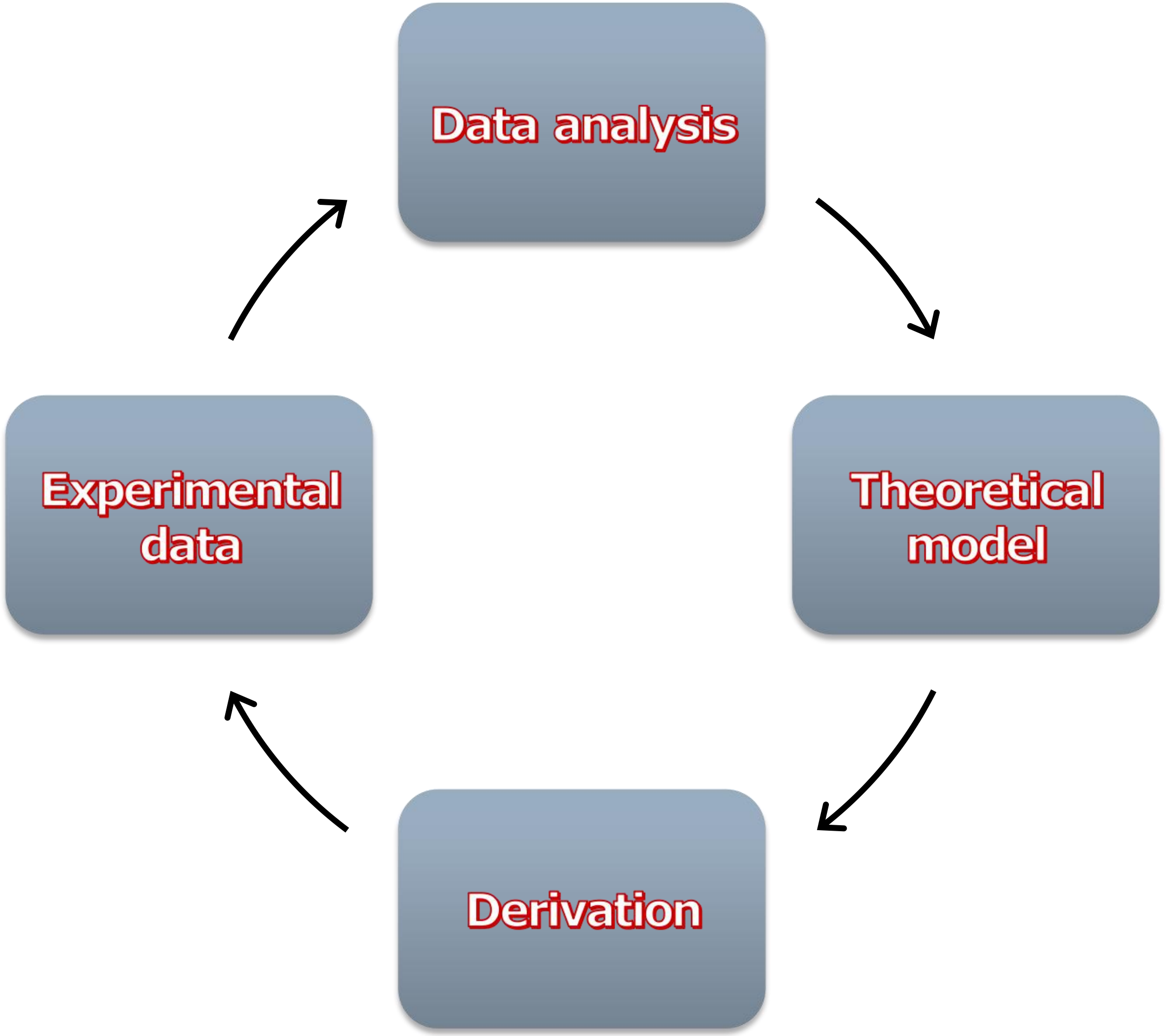
Peking University

lix-PHY@pku.edu.cn

2026/01/21

The circle of scientific discovery

From data to model

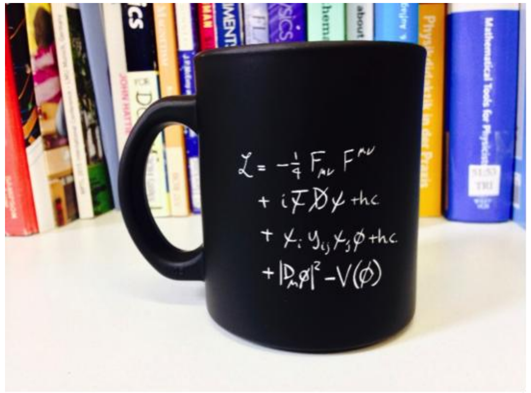


From model to data

Reflection

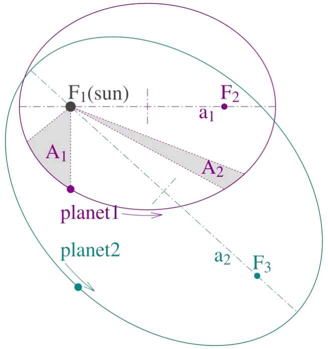
➢ Human exploration of natural laws:

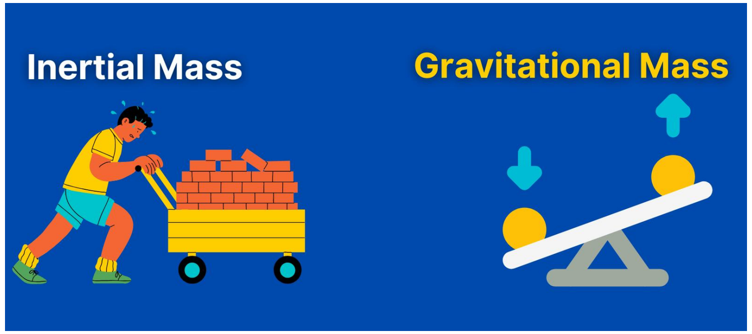
- **Advantages:** interpretability, conciseness, universality

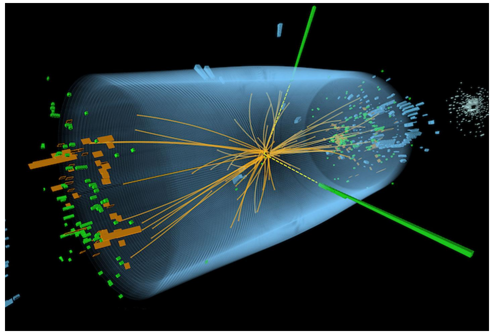

$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}$
$+ i\bar{\psi} \not{D} \psi + h.c.$
$+ \psi_i y_{ij} \psi_j \phi + h.c.$
$+ |D_\mu \phi|^2 - V(\phi)$

- **Disadvantages:** long period, preconceived notion,
insufficient ability to handle complex problems

New paradigm?

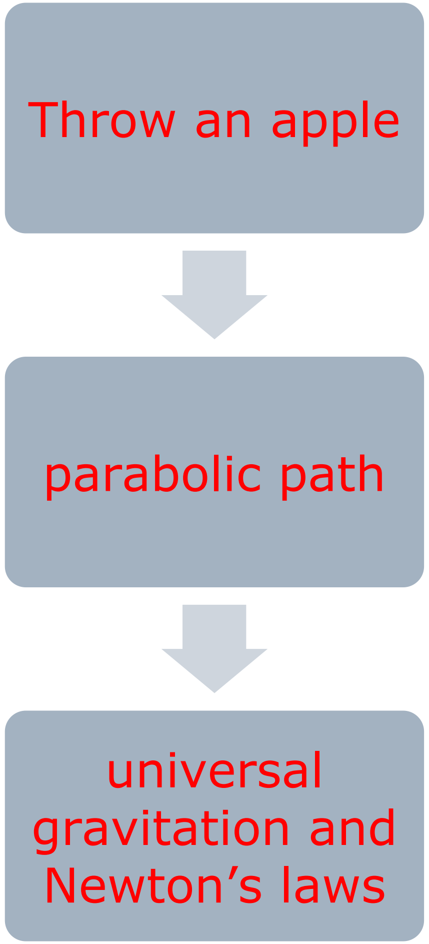






From Model to Theory

- **Specific model for one experiment**
  - Explore relations for a set of data
  - Symbolic regression, Funsearch/AlphaEvolve, …
- **General model for a large set of experiments**
  - How to define and explore relations between specific models?
- **Is it possible for AI to reproduce human’s theories?**


Throw an apple

parabolic path

universal gravitation and Newton’s laws

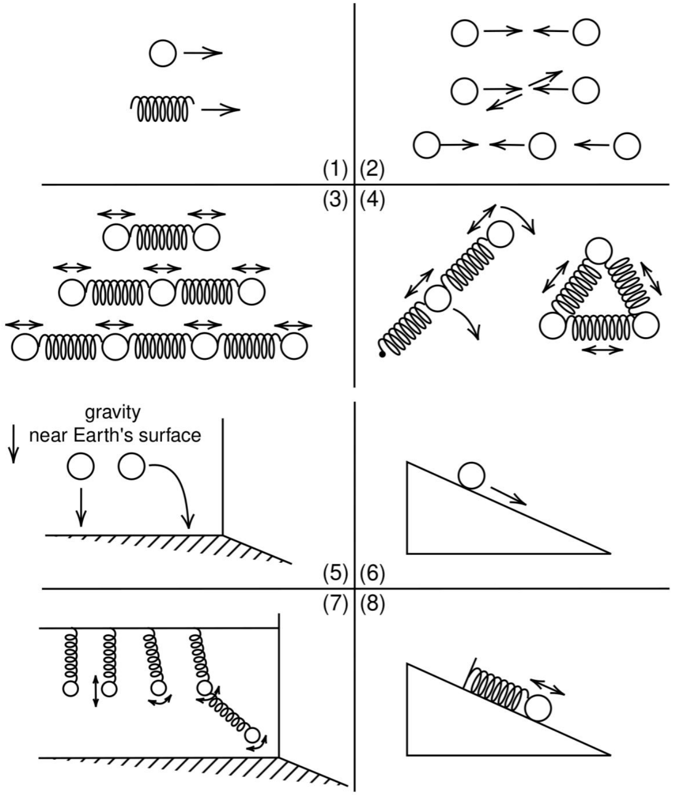
A Prototype of scientific discovery: AI-Newton

- **Based on noisy data, important natural laws are discovered!**
- **Unsupervised! Without prior physical knowledge!**

Fang, et al., 2504.01538

Physical objects

Schematic of experiments

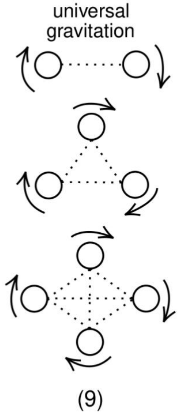


Discovered important general laws

Energy conservation

$$\sum_{\kappa\in\{x,y,z\}} T_\kappa + \sum_{\lambda\in\{k,g,G\}} \delta_\lambda V_\lambda = \text{const.},$$

where T_κ and V_λ are defined as:



$$T_\kappa = \sum_{i\in\text{Particles}} m_i v_{i,\kappa}^2,$$

$$V_k = \sum_{i\in\text{Springs}} k_i(L_i - L_{0,i})^2,$$

$$V_g = \sum_{i\in\text{Particles}} 2m_i g z_i,$$

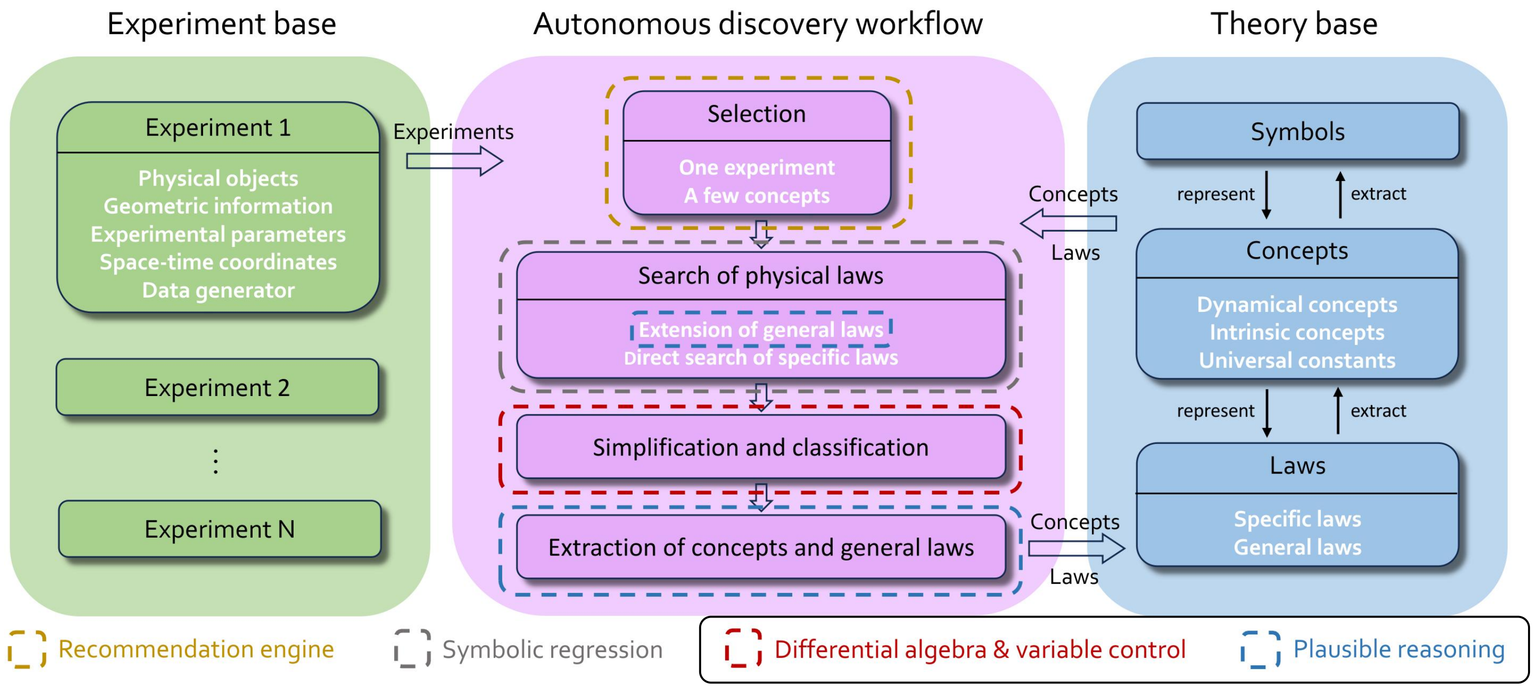
$$V_G = \sum_{i,j\in\text{Particles}} 2\left(-\frac{Gm_i m_j}{r_{ij}}\right).$$

Newton's second law

$$2a_\kappa + \sum_{\lambda\in\{k,g,G\}} \delta_\lambda \left(\frac{1}{m}\frac{\partial V_\lambda}{\partial \kappa}\right) = 0, \quad \kappa\in\{x,y,z\}.$$

($\delta_\lambda = 0$ or 1, determined spontaneously during instantiation as specific laws in experiments)

AI-Newton's bottleneck



Experiment base

Experiment 1

Physical objects
Geometric information
Experimental parameters
Space-time coordinates
Data generator

Experiment 2

⋮

Experiment N

Experiments

Autonomous discovery workflow

Selection

One experiment
A few concepts

Search of physical laws

Extension of general laws
Direct search of specific laws

Simplification and classification

Extraction of concepts and general laws

Concepts

Laws

Concepts

Laws

Theory base

Symbols

represent

extract

Concepts

Dynamical concepts
Intrinsic concepts
Universal constants

represent

extract

Laws

Specific laws
General laws

Recommendation engine

Symbolic regression

Differential algebra & variable control

Plausible reasoning

Fang, et al., 2504.01538

Replaced by
LLM agents for
real discovery

- **Mathematically simplification:**

  Rosenfield grobner algorithm in Differential algebra

- **Plausible reasoning:**

  Pre-established general rules

  1. Traversal

  2. Summary

- **The LLM Promise**

  Success for mathematical derivation:

  AlphaGeometry, AlphaProof ...

  Know all tricks people have used in history

Advancement in LLMs

- **LLMs** and their derivative products **excel in** **general domains**

- Chat client

ChatGPT

- Translation

DeepL

- Paper writng

sakana.ai

- Paper review

paperreview.ai
By Stanford ML Group

- Coding

Cline

- Research

Challenges for scientific AI

- **LLMs' reliability often drops in scientific problem-solving, which prioritize perfect performance over cost control**

- Caused from the inherent complexity of natural sciences
  - Long, multi-step and unstructured reasoning
  - Modeling of real-world scenarios
  - Understanding of fundamental laws
  - Implicit constraints
  - Deterministic & probabilistic, precise & approximate
  - …

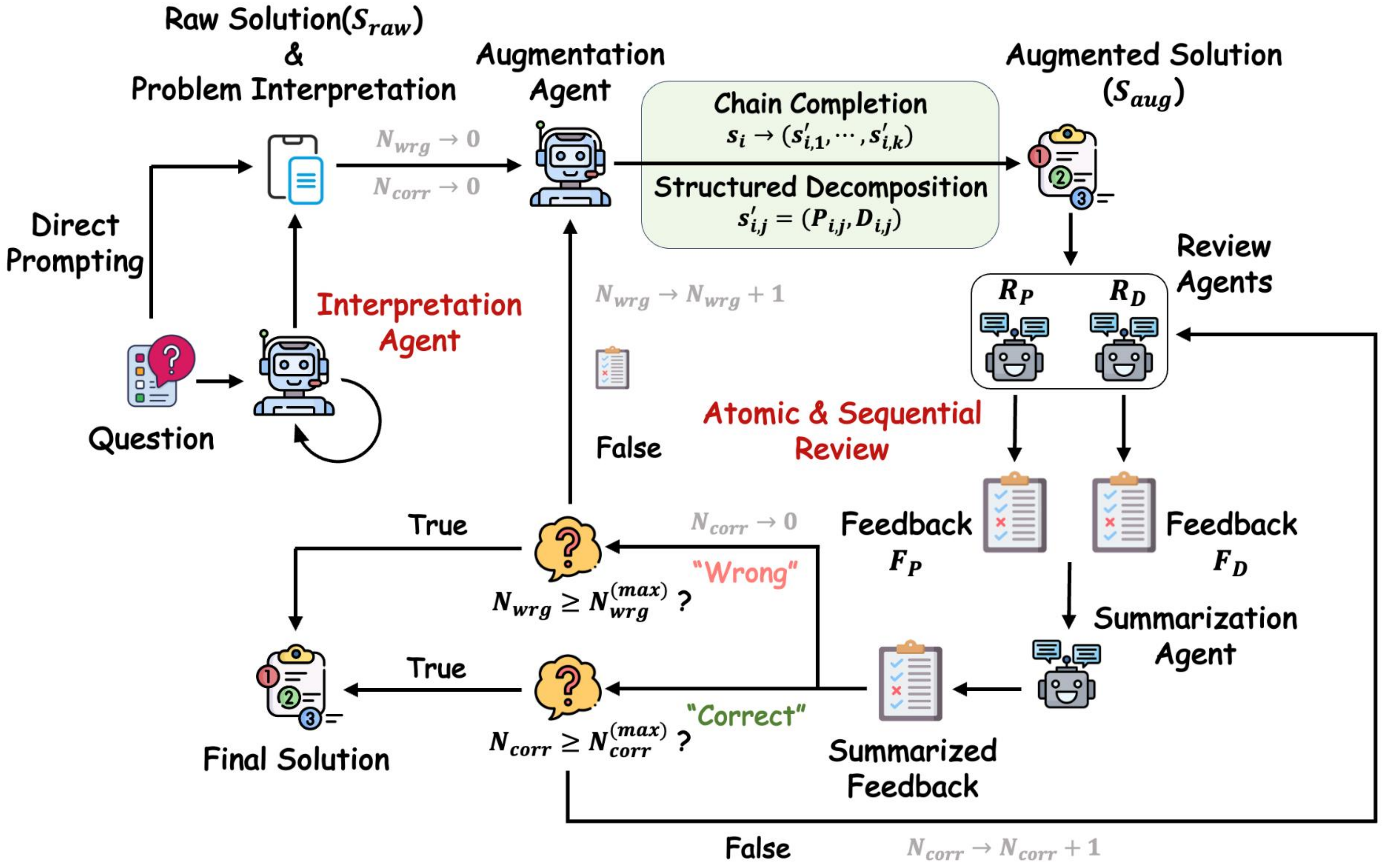
- Hard to detect due to logical leaps in the provided answers
  - Both human and AIs alike tend to omit steps they consider "obvious"

Overview of LOCA

➢ Logical Chain Augmentation (LOCA)

Fang et al., 2510.01249

Jian et al., 2511.10515



The CPhO: a challenging testbed

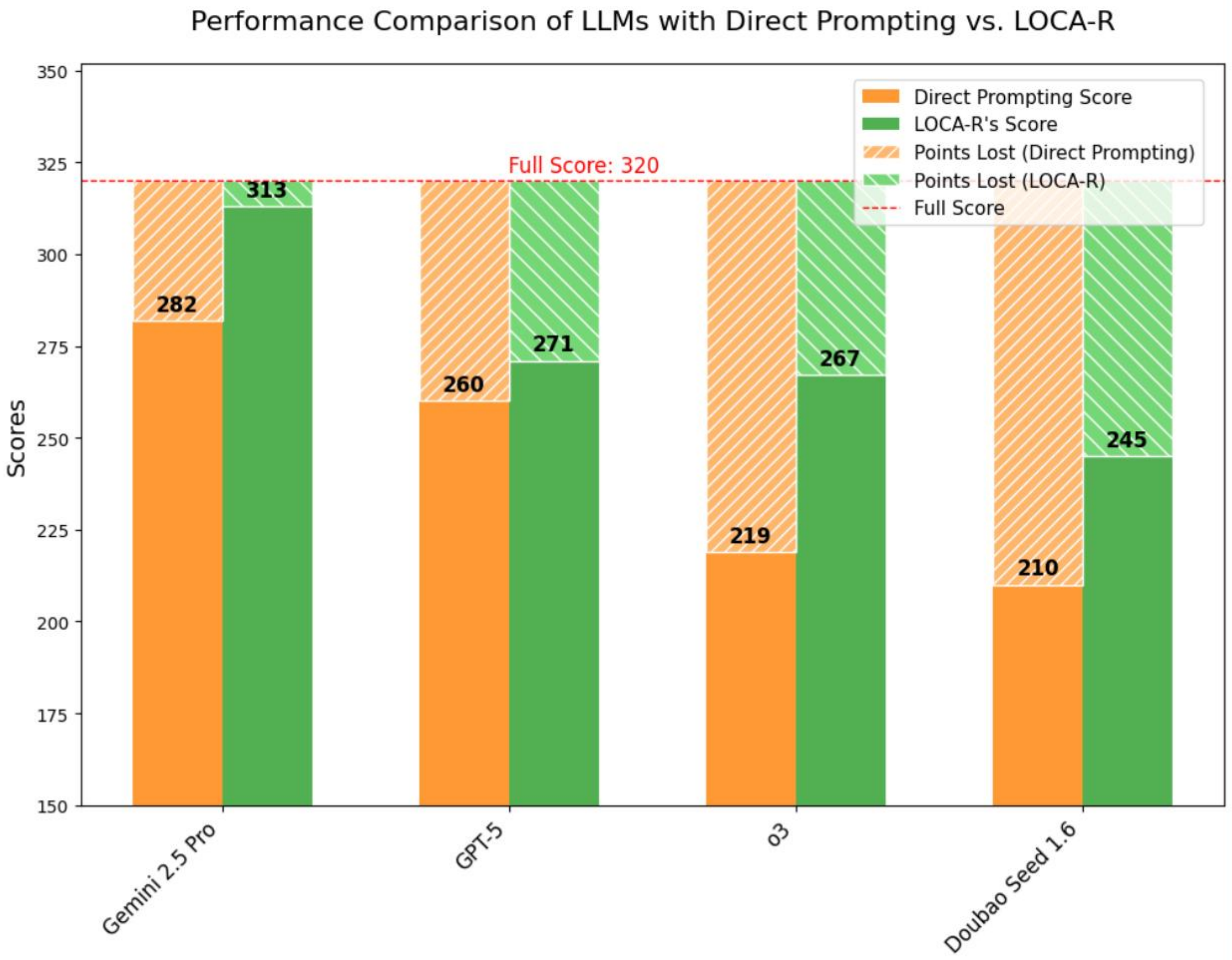
- The Chinese Physics Olympiad (CPhO): a premier national physics competition organized annually in China

  - Demands of long, multi-step reasoning
  - Multimodal problems
  - No data contamination issue

Overall performance across various base LLMs

Jian et al., 2511.10515

➢ Overall performance of LOCA-R on four mainstream LLMs



- Total scores for th. Problem: 320
- Highest score of human: 207

Comparison across more baseline methods

Jian et al., 2511.10515

➢ Comparison of LOCA-R and more baselines

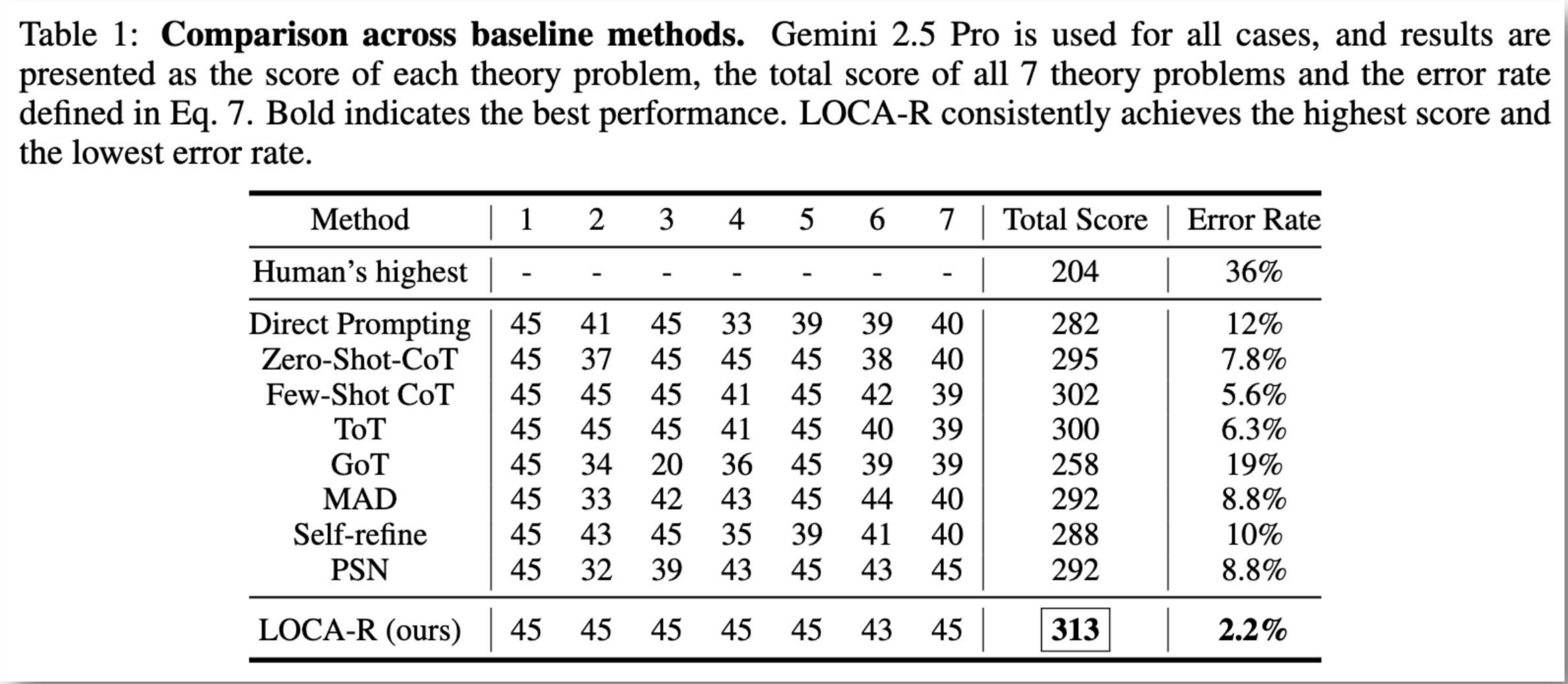


Table 1: **Comparison across baseline methods.** Gemini 2.5 Pro is used for all cases, and results are presented as the score of each theory problem, the total score of all 7 theory problems and the error rate defined in Eq. 7. Bold indicates the best performance. LOCA-R consistently achieves the highest score and the lowest error rate.

Method	1	2	3	4	5	6	7	Total Score	Error Rate
Human's highest	-	-	-	-	-	-	-	204	36%
Direct Prompting	45	41	45	33	39	39	40	282	12%
Zero-Shot-CoT	45	37	45	45	45	38	40	295	7.8%
Few-Shot CoT	45	45	45	41	45	42	39	302	5.6%
ToT	45	45	45	41	45	40	39	300	6.3%
GoT	45	34	20	36	45	39	39	258	19%
MAD	45	33	42	43	45	44	40	292	8.8%
Self-refine	45	43	45	35	39	41	40	288	10%
PSN	45	32	39	43	45	43	45	292	8.8%
LOCA-R (ours)	45	45	45	45	45	43	45	**313**	**2.2%**

Summary and outlook

- Human scientific discovery necessitates a new research paradigm, AI may help

- From model to theory:
  - Promote specific models as general theories by plausible reasoning
  - Verify general theories on specific problems by logical reasoning

- LLM for scientific problem-solving:
  - LOCA: break complex reasoning into smaller steps to enhence the reliability of review
  - Near perfect performance on CPhO

- AI for scientific discovery: remains in its infancy, but very promising

Thank you!