CMPhysBench: A Benchmark for Evaluating Large Language Models in Condensed Matter Physics

Weida Wang[1,4*], Dongchen Huang[2,3*], Jiatong Li[6*], Tengchao Yang[1,5*], Ziyang Zheng[2,3*], Chuyi Peng[2,3], ..., Yunqi Cai[2,3†], Xi Dai[7†], Shufei Zhang[1†], Lei Bai[1†], Jinguang Cheng[2†], Zhong Fang[2†], Hongming Weng[2,3†]

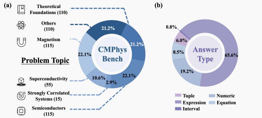
[1]Shanghai Artificial Intelligence Laboratory [2]Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences
[3]Condensed Matter Physics Data Center, Chinese Academy of Sciences [4]Fudan University [5]Tongji University [6]Hong Kong Polytechnic University
[7]Hong Kong University of Science and Technology [8]Hong Kong University of Science and Technology (Guangzhou)

Github

Motivation

- **Inadequate Benchmarks**: Current sets focus on K-12 or basic college levels, failing to test graduate-level research proficiency.
- **Complexity of Physics**: Condensed Matter Physics (CMP) requires unique mastery of conceptual principles and precise symbolic math.
- **Need for Nuance**: Brittle binary accuracy cannot capture near-miss reasoning in multi-step physical derivations.

CMPhysBench

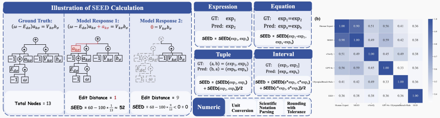


(c)

Answer Type	Example
Tuple	$(x, y, z) = (0, 1, 1)$
Expression	$\frac{1}{2}mv^2$
Interval	$[0,1]$
Numeric	$\frac{1}{16}$
Equation	$x^2 + y^2 = 1$

- ✓ **Graduate-Level Depth**: A collection of 520 meticulously curated calculation problems authored by Ph.D. students and postdocs.
- ✓ **Comprehensive Coverage**: Spans 6 core domains, including Magnetism, Superconductivity, and Strongly Correlated Systems.
- ✓ **Diverse Answer Types**: Supports open-ended solutions in five formats: Tuple, Numeric, Expression, Equation, and Interval.
- ✓ **Expert-Verified Pipeline**: Sourced from 17 classic textbooks with rigorous human review to ensure procedural correctness.

Scalable Expression Edit Distance



- Fine-grained, non-binary partial credit and interpretable error localization.
- The highest correlation with human expert judgment.

Example Problem

Consider the Anderson s-d exchange model with Hamiltonian

$$H = \sum_{k,\sigma} E_{k\sigma} n_{k\sigma} + \sum_{\sigma} E_{d\sigma} n_{d\sigma} + \frac{U}{2}\sum_{\sigma} n_{d\sigma} n_{d\bar{\sigma}} + \sum_{\sigma} n_{d\sigma} n_{d\bar{\sigma}} + \sum_{k,\sigma} V_{kd}(C^{\dagger}_{k\sigma} d_{\sigma} + d^{\dagger}_{\sigma} C_{k\sigma})$$

where

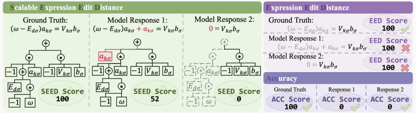
$$E_{k\sigma} = E_k + \sigma\mu_B h, \quad E_{d\sigma} = E_d + \sigma\mu_B h$$

Here, $\mu_B = \left(\left|\frac{e}{2mc}\right|\hbar\right)$ is the Bohr magneton, with a Landé factor of $g_0 = g_i = 2$ for both electrons and impurities. This is the non-degenerate orbital Anderson s-d mixing model. Derive the equation of motion for the s-d exchange model concerning the mixed Green's function $\ll C_{k\sigma}|d^{+}_{\sigma}\gg_{\omega}$.
Hint: Let $a_{k\sigma}$ symbolize $\ll C_{k\sigma}|d^{+}_{\sigma}\gg_{\omega}$, and b_{σ} symbolize $\ll d_{\sigma}|d^{+}_{\sigma}\gg_{\omega}$.

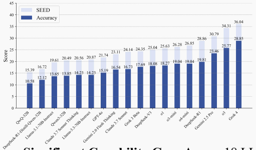
Answer Type: Equation

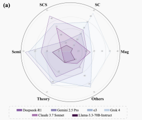
Final Answer: $(\omega - E_{d\sigma})a_{k\sigma} = V_{k\sigma}b_{\sigma}$

Topic: Strongly Correlated Systems

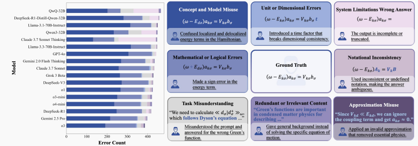


Results and Analysis





- **Significant Capability Gap**: Across 18 LLMs, the top model achieves only a 36 SEED score and 29% accuracy.
- **Asymmetric Strengths**: Models show highly uneven competence across different CMP subfields.
- **Reasoning vs. General**: Longer CoT does not guarantee better performance without solid domain knowledge.



- **Concept and Model Misuse** (~50-60%): The dominant error, exposing weak physics intuition.
- **Mathematical or Logical Errors** (20–30%): Highlights persistent bottlenecks in symbolic derivation.