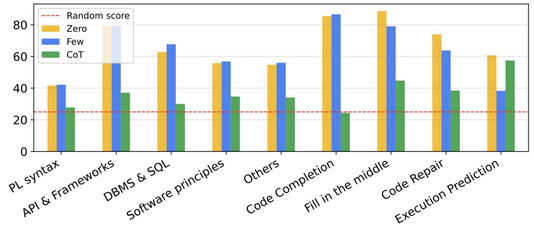
# CodeMMLU: A Multi-Task Benchmark for Assessing Code Understanding Capabilities of CodeLLMs

**Dung Nguyen Manh, Thang Phan Chau, Nam Le Hai, Thong T. Doan, Nam V. Nguyen, Quang Pham, Nghi D. Q. Bui**

Leaderboard

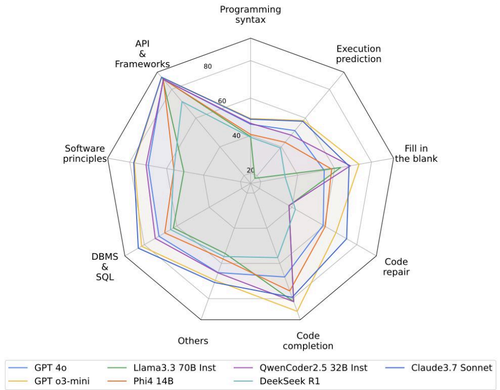
## MOTIVATION

- Existing benchmarks focus on open-ended generation, **not reasoning or understanding**.
- Practical LLM applications reveal bias, hallucinations, and misunderstanding of code semantics.
- Evaluation via **test cases limits scale**, coverage, and interpretability.

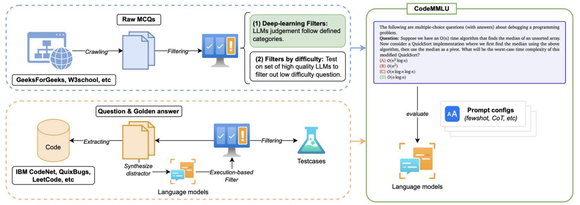



## CONTRIBUTION

- **CodeMMLU**, the first large-scale multiple-choice benchmark tailored to software and code understanding.
- Covers both *knowledge-based* and *task-based* evaluations.
- **Move beyond generation:** Evaluate models' reasoning, debugging, and software knowledge capabilities.




## BENCHMARK CONSTRUCTION




***(1) Knowledge-based test:***

- Topic: *Syntax rules, APIs/ frameworks, software principles, DBMS/SQL, etc.*
- Filtered by LLMs for clarity, difficulty, and relevance

***(2) Fundamental skills test:***

- Adapted from LeetCode, QuixBugs, CodeNet, etc
- Tasks reframed into **MCQs** with **LLM-generated distractors**
- Execution-based filtering ensures correctness and difficulty

## LEADERBOARD

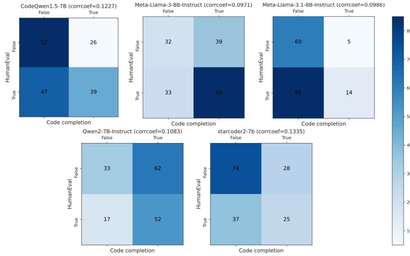
- **GPT 4o** is the top performer, but ***GPT o3-mini*** is the best compact size model.
- Instruction-tuned open-source models show strong performance.
- Fundamental tasks are more discriminative than knowledge tasks
- Scaling law holds within model families, *but fails across families* → highly the quality of pretraining data

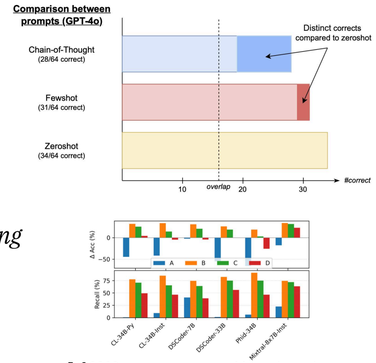
| Family | Model name | Size (B) | CodeMMLU | Rank |
|---|---|---|---|---|
| | **Closed-source models** | | | |
| Anthropic | Claude 3.7 Sonnet | - | 61.65 | *3* |
| | Claude 3.5 Sonnet | - | 59.81 | *5* |
| | Claude3 Sonnet (20240229) | - | 53.97 | *8* |
| OpenAI | GPT o3-mini | - | 62.36 | *2* |
| | GPT 4o (2024-05-13) | - | **67.00** | *1* |
| | GPT 4o-mini | - | 38.43 | *19* |
| | **Open-source models** | | | |
| Meta | Llama3.3 70B Inst | 70 | 40.66 | *17* |
| | Llama3.1 405B Inst | 405 | 58.23 | *6* |
| | Llama3.1 70B Inst | 70 | **60.00** | *4* |
| | CodeLlama34B Inst | 34 | 38.73 | *18* |
| DeepSeek | DeepSeek R1 | 671 | 43.91 | *14* |
| | DeepSeek V3 | 685 | 49.08 | *11* |
| | DeepSeekCoder 33B Inst | 33 | 36.60 | *20* |
| | DeepSeekMoE 16B Chat | 16.4 | 31.01 | *22* |
| Mistral | Mistral7B Inst (v0,3) | 7 | 43.33 | *15* |
| | Mixtral 8×7B Inst | 46.7 | 42.96 | *16* |
| | Codestral 22B | 22 | 47.60 | *12* |
| Microsoft | Phi4 | 14 | 49.19 | *10* |
| | Phi4 Mini Inst | 12 | 34.85 | *21* |
| Qwen | Qwen2.5 Max | - | 56.40 | *7* |
| | Qwen2.5 14B Inst | 14 | 51.38 | *9* |
| | QwQ 38B Preview | 57 | 46.34 | *13* |

## KEY INSIGHTS

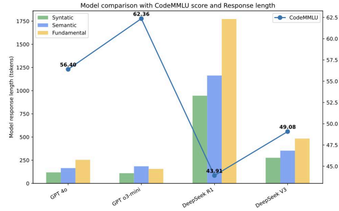
- Chain-of-Thought (CoT) often reduces performance on deterministic tasks.
- **Reasoning models** generate long, reasoning-heavy outputs but underperform. *Overreasoning leads to lower accuracy.*

| Models | A | B | C | D | STD |
|---|---|---|---|---|---|
| GPT-4o | 80.49 | 78.05 | 71.34 | 70.12 | 4.38 |
| Claude3.5 Sonnet | 90.24 | 81.1 | 85.37 | 79.27 | **4.23** |
| Claude3 Opus | 79.27 | 77.44 | 82.32 | 84.76 | **2.81** |
| Mixtral 8x7B Inst | 22.56 | 74.39 | 71.95 | 63.41 | 20.91 |
| Deepseek Coder 33B | 1.22 | 82.32 | 75.00 | 56.10 | 31.75 |
| CodeLlama 34B Py | 0.61 | 77.44 | 70.73 | 49.39 | 30.09 |
| CodeLlama 34B Inst | 9.15 | 84.76 | 65.24 | 46.34 | 27.91 |




- MCQ accuracy fluctuates with option order ($\Delta\sigma = 36.66$ in weaker models), revealing structural bias — but **top models** show much **greater robustness** to this effect.
- High HumanEval scores do not guarantee MCQ performance on the same questions ($\rho \approx 0.1$).




## TAKEAWAYS

- Knowledge ↔ Real-world Skills
- Chain-of-Thought Hurts
- (Reasoning) Longer ≠ Better