MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, Wenhu Chen

[NeurIPS 2024 Dataset Track]

Motivation

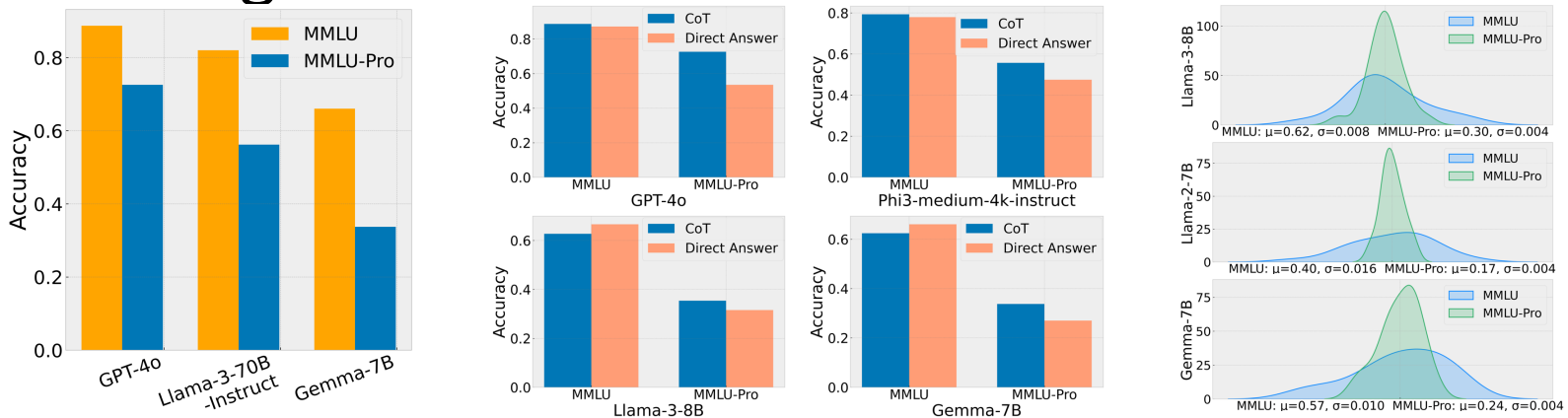
1. **Performance saturation** (90%+) on MMLU limits differentiation between advanced models

2. **Knowledge-focused questions** with 4 options enable shortcut exploitation rather than understanding

3. **Dataset noise** creates artificial performance ceiling, reducing benchmark effectiveness

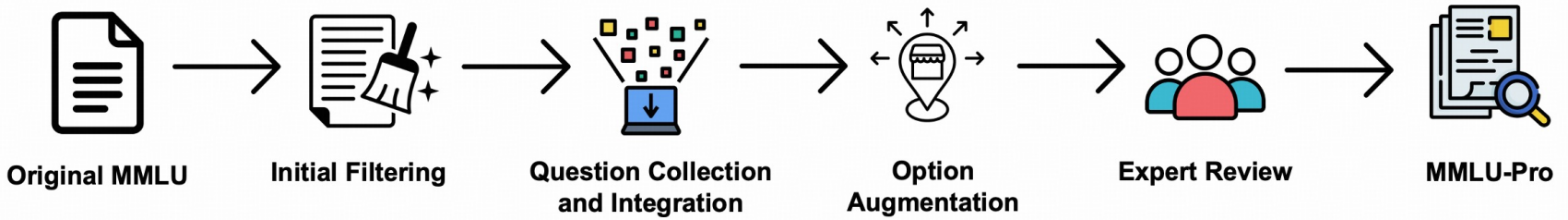
Key Differences from MMLU:

- Expanded answer choices from 4 to 10 options, reducing random guess probability from 25% to 10%
- Higher robustness to prompt variations, with sensitivity reduced from 4-5% (MMLU) to 2% (MMLU-Pro)
- Enhanced focus on reasoning over knowledge-based questions, evidenced by 20% improvement with CoT reasoning



Dataset Construction

- Initial Filtering
- Question Collection
- Option Augmentation
- Expert Review



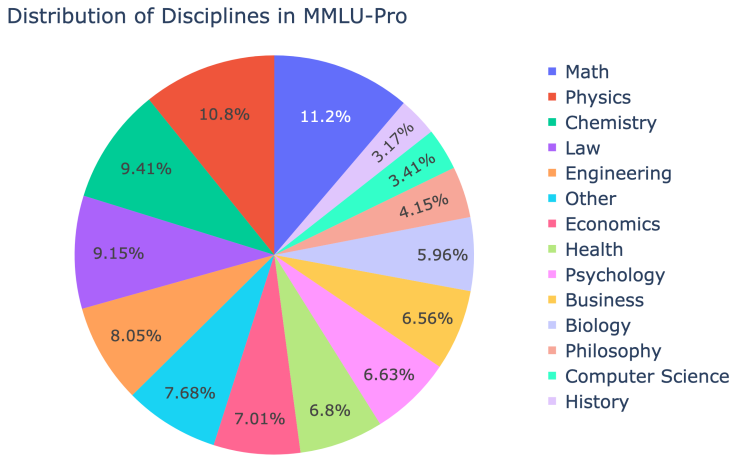
Data Source

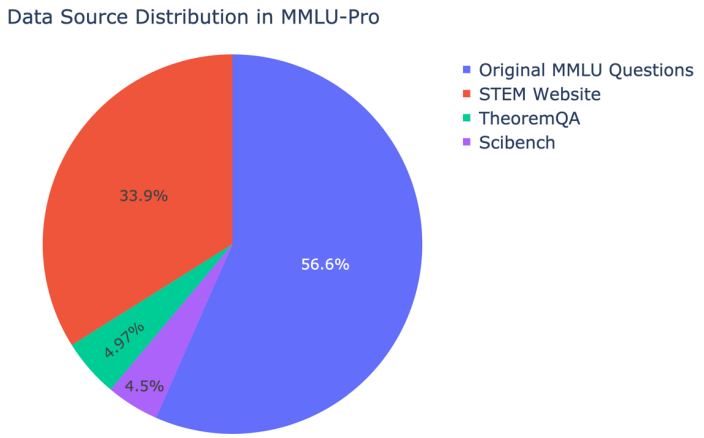
The dataset consolidates questions from several sources:

- **Original MMLU Questions:** Part of the dataset comes from the original MMLU dataset. We remove the trivial and ambiguous questions.
- **STEM Website:** Hand-picking high-quality STEM problems from the Internet.
- **TheoremQA:** A high-quality collection of human-annotated questions specifically requiring theorem application to solve.
- **SciBench:** Science questions from college exams.

Data Distribution

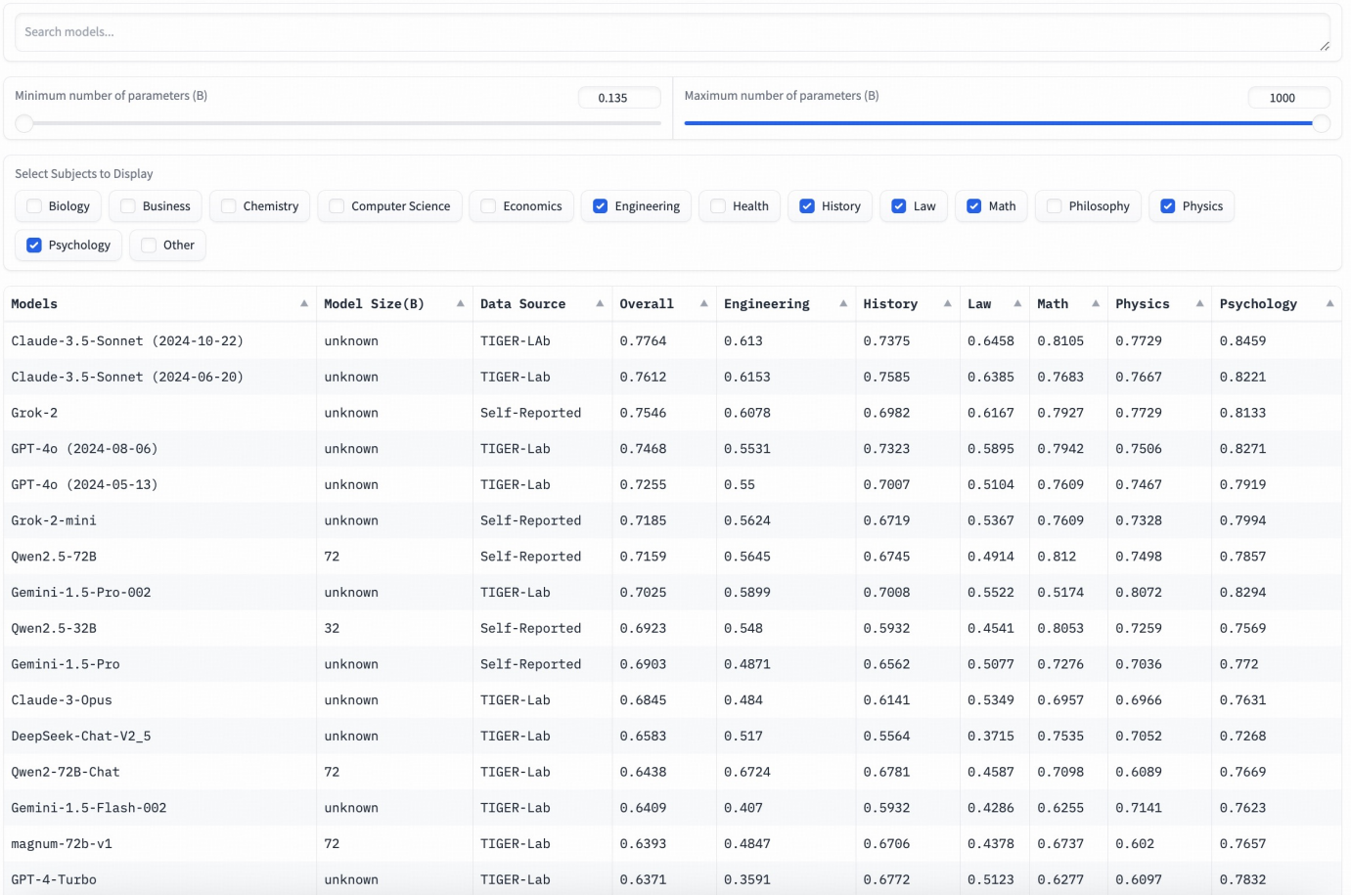
Discipline	Number of Questions	From Original MMLU	Newly Added
Math	1351	846	505
Physics	1299	411	888
Chemistry	1132	178	954
Law	1101	1101	0
Engineering	969	67	902
Other	924	924	0
Economics	844	444	400
Health	818	818	0
Psychology	798	493	305
Business	789	155	634
Biology	717	219	498
Philosophy	499	499	0
Computer Science	410	274	136
History	381	381	0
Total	**12032**	6810	5222





Leaderboard

- Platform & Scale
  - Hosted on Hugging Face Space
  - Features 100+ popular models
  - Includes both open-source and closed-source models
- Model Coverage
  - Open-source models: 135M - 399B parameters
- Features
  - Search by model name
  - Filter by parameter count and subject areas



Models	Model Size(B)	Data Source	Overall	Engineering	History	Law	Math	Physics	Psychology
Claude-3.5-Sonnet (2024-10-22)	unknown	TIGER-LAb	0.7764	0.613	0.7375	0.6458	0.8105	0.7729	0.8459
Claude-3.5-Sonnet (2024-06-20)	unknown	TIGER-Lab	0.7612	0.6153	0.7585	0.6385	0.7683	0.7667	0.8221
Grok-2	unknown	Self-Reported	0.7546	0.6078	0.6982	0.6167	0.7927	0.7729	0.8133
GPT-4o (2024-08-06)	unknown	TIGER-Lab	0.7468	0.5531	0.7323	0.5895	0.7942	0.7506	0.8271
GPT-4o (2024-05-13)	unknown	TIGER-Lab	0.7255	0.55	0.7007	0.5104	0.7609	0.7467	0.7919
Grok-2-mini	unknown	Self-Reported	0.7185	0.5624	0.6719	0.5367	0.7609	0.7328	0.7994
Qwen2.5-72B	72	Self-Reported	0.7159	0.5645	0.6745	0.4914	0.812	0.7498	0.7857
Gemini-1.5-Pro-002	unknown	TIGER-Lab	0.7025	0.5899	0.7008	0.5522	0.5174	0.8072	0.8294
Qwen2.5-32B	32	Self-Reported	0.6923	0.548	0.5932	0.4541	0.8053	0.7259	0.7569
Gemini-1.5-Pro	unknown	Self-Reported	0.6903	0.4871	0.6562	0.5077	0.7276	0.7036	0.772
Claude-3-Opus	unknown	TIGER-Lab	0.6845	0.484	0.6141	0.5349	0.6957	0.6966	0.7631
DeepSeek-Chat-V2_5	unknown	TIGER-Lab	0.6583	0.517	0.5564	0.3715	0.7535	0.7052	0.7268
Qwen2-72B-Chat	72	TIGER-Lab	0.6438	0.6724	0.6781	0.4587	0.7098	0.6089	0.7669
Gemini-1.5-Flash-002	unknown	TIGER-Lab	0.6409	0.407	0.5932	0.4286	0.6255	0.7141	0.7623
magnum-72b-v1	72	TIGER-Lab	0.6393	0.4847	0.6706	0.4378	0.6737	0.602	0.7657
GPT-4-Turbo	unknown	TIGER-Lab	0.6371	0.3591	0.6772	0.5123	0.6277	0.6097	0.7832

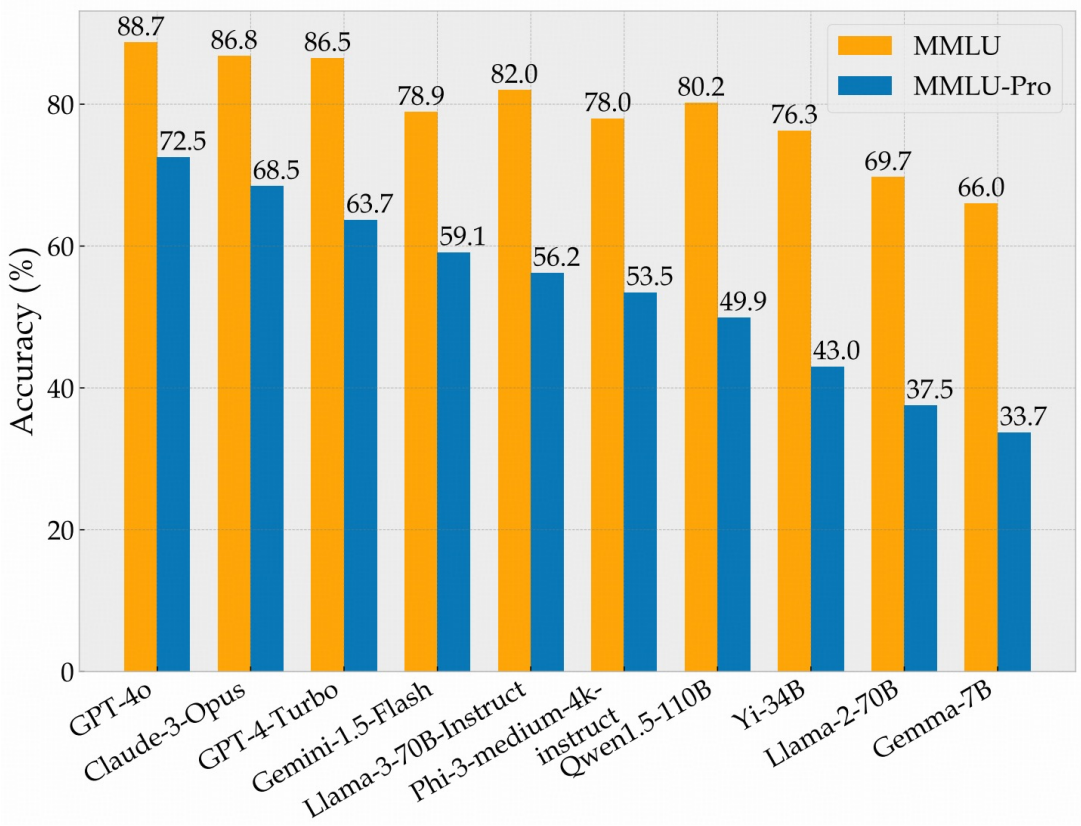
Analysis

- 1. Difficulty Level
- 2. Reasoning Level
- 3. Robustness Degree

Analysis 1: Difficulty Level

MMLU vs MMLU-Pro Model Performance Analysis

- Score Clustering in MMLU
- Better Differentiation
- Room for Improvement



Analysis 2: Reasoning Level

Model Name	MMLU			MMLU-Pro		
	CoT	Direct Answer	CoT - DA	CoT	Direct Answer	CoT - DA
GPT-4o	88.7	87.2	**1.5**	72.6	53.5	**19.1**
GPT-4-Turbo	86.5	86.7	**-0.2**	63.7	48.4	**15.3**
Phi3-medium-4k-instruct	79.4	78.0	**1.4**	55.7	47.5	**8.2**
Llama-3-8B	62.7	66.6	**-3.9**	35.4	31.5	**3.9**
Gemma-7B	62.4	66.0	**-3.6**	33.7	27.0	**6.7**

CoT vs Direct Answering: Performance Analysis

- Overall Performance Trend
- Model-Specific Improvements
- Key Implications

Analysis 3: Robustness Degree

- Tested using 24 different reasonable prompts

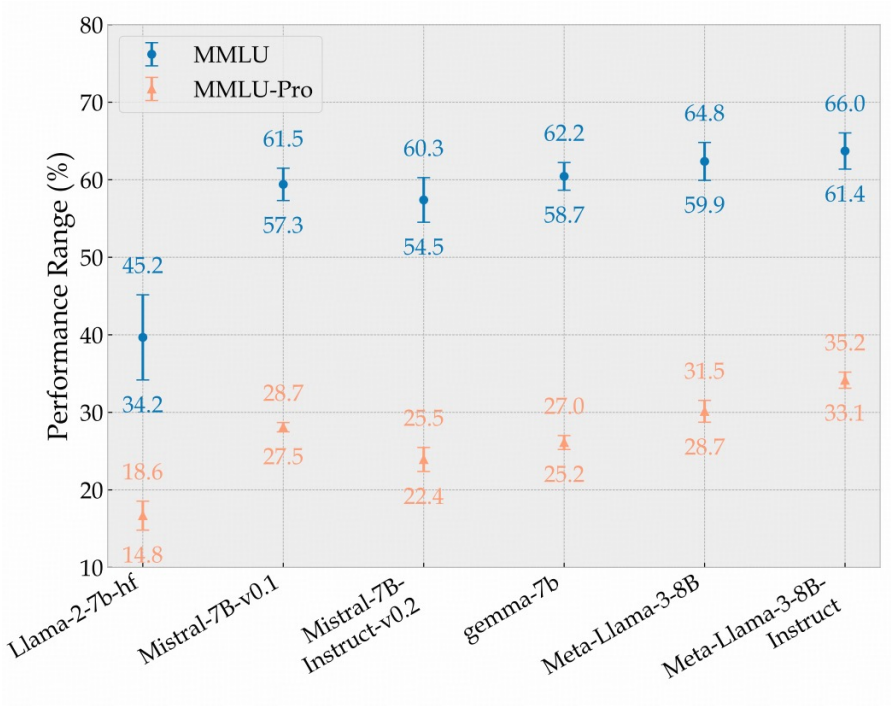
- Benchmark Comparison

  MMLU:

  - General variation: 4-5%
  - Maximum variation: 10.98%

  MMLU-Pro:

  - General variation: ~2%
  - Maximum variation: 3.74%



Performance Variability under Different Prompts on MMLU and MMLU-Pro

GPT-4o Error Analysis on MMLU-Pro

- Methodology
  - Analysis of 120 randomly selected errors
  - Evaluated by expert annotators
- Reasoning Errors: 39%
  - Logical inconsistencies
  - Pattern recognition vs true understanding
- Knowledge Gaps: 35%
  - Lack of specialized domain knowledge
  - Issues with technical applications
- Calculation Errors: 12%
  - Correct formulas but wrong computations