Comparison of the time it takes “meta-llama/Llama-3.2-1B-Instruct” to answer all the questions under MMLU subjects "astronomy" and "business_ethics" when differing device and quantization.

| model | device | quantization | wall_clock_seconds | cpu_time_seconds | gpu_time_seconds |
|---|---|---|---|---|---|
| meta-llama/Llama-3.2-1B-Instruct | cpu | 4 | 47.288663 | 42.390625 | 0.0 |
| meta-llama/Llama-3.2-1B-Instruct | cuda | 4 | 18.6522 | 17.359375 | 18.651017578125 |
| meta-llama/Llama-3.2-1B-Instruct | cuda | 8 | 31.84831 | 30.46875 | 31.84686328125 |
| meta-llama/Llama-3.2-1B-Instruct | cpu | full | 328.924944 | 2757.875 | 0.0 |
| meta-llama/Llama-3.2-1B-Instruct | cuda | full | 12.766592 | 11.625 | 12.7663037109375 |

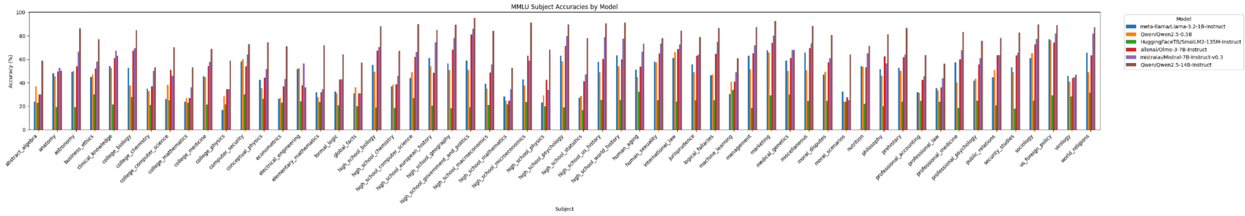
---

Comparison of the time it takes 6 LLMs to answer all the questions under the 57 MMLU.

| model | GPU | quantization | wall_clock_seconds | cpu_time_seconds | gpu_time_seconds |
|---|---|---|---|---|---|
| meta-llama/Llama-3.2-1B-Instruct | T4 | full | 607.826565 | 488.3365144730001 | 607.8305 |
| Qwen/Qwen2.5-0.5B | T4 | full | 617.224114 | 573.545525202 | 617.2289375 |

| | | | | | |
|---|---|---|---|---|---|
| HuggingFaceTB/SmolLM2-135M-Instruct | T4 | full | 723.503396 | 677.374591928 | 723.5088125 |
| allenai/Olmo-3-7B-Instruct | T4 | full | 7270.587709 | 7162.936070966 | 7270.6515 |
| mistralai/Mistral-7B-Instruct-v0.3 | A100 | full | 896.04945 | 608.617769184 | 896.045625 |
| Qwen/Qwen2.5-14B-Instruct | A100 | full | 1000.705823 | 891.8444725270001 | 1000.7006875 |

Bar Graph showing the accuracy of 6 LLMs on the 57 MMLU subjects.



Questions:

Can you see any patterns to the mistakes each model makes or do they appear random?

- The mistakes are not all random. The models all fail on the following: factual recall, numeric/units calculations and trap-style options.
- The models frequently get hyper specific facts wrong (such as dates) and numerical conversion such as the approximate conversion of 25 degrees Celsius to Kelvin.
- They also often pick answers that are plausible word choice but not correct.

Do all the models make mistakes on the same questions?

- While all the models demonstrated the same systematic weaknesses they still differed in which question they got wrong
- There were some questions that were commonly missed such as Pluto composition or Boltzmann constant. These commonly missed questions were typically obscure facts.
- Example:
    - --- Subject: astronomy ---
    - What is the correct numerical value and unit of the Boltzmann constant?
    - A. 1.38 × 10−21 m^3 · kg · s^−2 · K^−1
    - B. 1.38 × 10−22 m^2 · kg · s^−3 · K^−1
    - C. 1.38 × 10−23 m^2 · kg · s^−2 · K^−1
    - D. 1.38 × 10−24 m^2 · kg · s^−2 · K^−2

Chat history comparison:
When chat history is off the agent obviously can’t hold a real conversation. I tested by giving it a story and asking facts. It obviously doesn’t know the story. The agent that has history can recall facts. I tested an agent that summarizes old chats and it slowly forgot the story.