Buddy — MMLU General Knowledge — 57 subjects

AIIT-THRESHOLD · Council Hill, Oklahoma · **June 12, 2026** · stratified practice exam, all 57 subjects sampled · zero-shot

System under test: Buddy — Qwen2.5-14B fine-tune ("buddy-merged"), one RTX 3090, no cloud, nothing leaves the house, routed through his `lme_bench` clean room (no memory injection). MMLU is the most-cited LLM knowledge benchmark — 57 subjects, high-school to graduate level.

135 / 200 (67.5%)

67.5% zero-shot, single-letter, choices shuffled per question. A stratified practice exam across all 57 subjects — solid for a 14B answering cold through its own serving stack. (Stock Qwen reports ~78%, but that is 5-shot with logit scoring — a different, easier protocol.)

The shape of it (strong on words, soft on symbols)

strongest (aced)		softest (the study list)	
clinical knowledge	100%	college mathematics	25%
econometrics	100%	high-school statistics	33%
US / European history	100%	machine learning	33%
government & politics	100%	abstract algebra	50%
elementary math	100%	college computer science	50%

One clean fault line: he is strong on verbal, historical, and conceptual subjects, and soft where the **symbols and digits stack up** — the same place GSM8K caught him. Knowing the subject was rarely the problem; landing the symbol-heavy answer was.

What the practice exam diagnosed

Re-examining his quantitative misses with reasoning captured: roughly a quarter were **format-starved** — he knew the material (correct calculus on $f(x)=e^x-cx$; correct ring-characteristic definitions), and the zero-shot single-letter format simply gave him no room to work. The rest were **genuine** — fluent reasoning that lands on the wrong answer. Verified clean: his pipeline carried the math and logic symbols (√, ⊃, ≡, exponents) intact — these are real reasoning rungs, not corrupted inputs.

Grading honesty

200 of 14,042 questions, stratified so every subject is represented (~3–4 each — individual subject percentages are noisy at that depth; the **overall** and the **quantitative cluster** are the trustworthy signals). Zero-shot, single-letter, shuffled — a specific protocol, not a certified head-to-head against 5-shot leaderboard numbers. 1 of 200 unparsed. Raw rows: `~/Desktop/buddy_mmlu_chunk200.jsonl`.

Buddy — Qwen2.5-14B raised by Rhet Wike · Council Hill, Oklahoma · June 12, 2026
He knows the world in words. The symbols are the next rung.