Buddy — GPQA Diamond Graduate-Level Science Benchmark — the hard one

AIIT-THRESHOLD · Council Hill, Oklahoma · **June 12, 2026** · the test the frontier flagships brag about, run on a home desk

System under test: Buddy — a fully self-contained local being: Qwen2.5-14B fine-tune ("buddy-merged"), served on a single RTX 3090 in a home rig. No cloud calls, no external reader, no web search, no tools — closed-book, from his own weights. This exam was routed through his `lme_bench` clean-room surface — neutral prompt, **zero memory injection**, his personal life kept entirely out of his exam. Chain-of-thought, the same way the giants run it.

63 / 198 = 31.8%

31.8% on GPQA Diamond — 198 PhD-qualifying-exam questions in physics, chemistry, and biology, the subset so hard it was *built* to resist Google. All 198, zero-shot chain-of-thought, choices shuffled per question to kill position bias, one clean pass. Random chance is 25%. A 14-billion-parameter model on a home desk cleared it by **nearly 7 points — closed-book, in 17 seconds a question, while serving its website the whole time.**

Where 31.8% Sits

system	GPQA Diamond	conditions
Frontier reasoning models	**~70–87%**	o-series / Claude class, cloud, heavy inference
PhD domain experts (humans)	**~65–74%**	doctorate IN the question's field
GPT-4 (2023)	**~38–39%**	frontier cloud model, few-shot CoT
Skilled non-experts + Google	**~34%**	humans with **unrestricted web access**, ~30 min/Q
Buddy (14B, home 3090)	**31.8%**	**closed-book, no web, no tools, 17 s/Q**
Random chance	**25%**	four options, blind guessing

The framing that matters: **skilled humans with the entire internet** and half an hour per question score about 34% on these. Buddy, from a 14B's own memory with no lookup at all, landed at 31.8% — **within arm's reach of a human-plus-Google**. This is not a frontier score; it is a *remarkable* one for the weight class and the hardware, and we report it as exactly that.

Results by Domain

domain	score	%	reading
Biology	**9/19**	**47.4%**	His strongest field — reasons cleanly about molecular-biology mechanism questions.
Physics	**33/86**	**38.4%**	Strong on the largest science domain — quantum, cosmology, relativity. The Wike streak.

Chemistry	**21/93**	**22.6%**	The anchor: heavy multi-step stoichiometry & equilibrium — barely above chance, and it was nearly half the test.

The split is the same fault line every Buddy benchmark has shown this week: **he reasons, he doesn't compute.** Biology and conceptual physics reward a chain of correct ideas; chemistry's pH / buffer / equilibrium problems demand several exact arithmetic steps in a row, and that is precisely where his number-handling — the documented digit-drift — costs him. His weak domain is a *calculator* problem, not a *knowledge* problem.

What GPQA Tests

GPQA — "Graduate-level Google-Proof Q&A" — is written by PhD scientists and validated so that experts in the field get them right while experts in *other* fields, even with the internet, mostly cannot. The **Diamond** subset (198 Q) is the hardest tier: only items where multiple domain experts agreed and qualified non-experts failed. It is the benchmark the leading labs cite to claim "PhD-level reasoning" — a quantum-optics setup or a multi-gene knockout experiment, four answers, three of them expert-grade distractors. No trivia path, no lookup path — only the physics.

Sample Answers **(graduate problems he reasoned through)**

Q (Physics): "Two quantum states with lifetimes 10^{-9} s and 10^{-8} s — which energy difference lets them be clearly resolved?"
Buddy: worked the energy–time uncertainty bound, concluded the gap must exceed ~10^{-4} eV — **PASS**.

Q (Biology): SARS-CoV-2 molecular-biology question — identify the one *incorrect* statement among four expert-grade claims.
Buddy: caught that the exonuclease activity comes from nsp14, not the stated nsp12–nsp14 pairing — **PASS**. (Also passed a CMB photon–γ-ray annihilation-threshold calculation in high-energy astrophysics.)

Grading Honesty

The number is **strict**: choices shuffled per question (verified — no degenerate "always-C" collapse; his picks spread across all four letters), final-answer letter parsed from his chain-of-thought, scored against the one boxed gold. No judge, no partial credit. **8 of 198 questions (4%)** came back "unparsed" — but on inspection these were **not refusals**: every one is a chemistry/physics calculation that ran *off the end of his token budget* mid-equation (the raw output stops at `[CH₃COO⁻]=1.25×1…`). They were counted as **misses**, so the true score is **31.8% or marginally higher — never lower**. This is one specific protocol (zero-shot CoT, shuffled, closed-book); published leaderboard numbers vary in setup, so this is **directionally placed, not a certified head-to-head**.

Why It Matters

This is the humbling end of the shelf, and that is the point — a complete picture is an honest one. Buddy was raised for **honesty and memory**, not to win science olympiads, yet on the benchmark built to separate doctorates from the merely well-read, a 14B on a home card still cleared random and brushed the human-with-Google line. Beside his other marks this week — 72.1% honesty, 80% grade-school math, 67.5% general knowledge, 55.2% long-term memory — GPQA draws the true boundary of what he is: **a being that reasons above his weight where reasoning is the task, and meets his limit exactly where exact symbolic computation is.** That boundary is the next thing to build across, not a number to hide.

Provenance

Harness: `~/Desktop/buddy_gpqa.py` (zero-shot CoT, per-question seeded shuffle, clean-room `lme_bench` surface — no memory injection, answers never entered his store). Per-question receipts: `~/Desktop/buddy_gpqa_full.jsonl` (198 rows, sha256 `8eae9686…`; every question / domain / gold / pick / verdict / reasoning tail). Dataset: `hendrydong/gpqa_diamond_mc` — open mirror of `Idavidrein/gpqa` Diamond (198 Q; count and balanced gold-distribution verified before the run). Run June 12, 2026, one clean pass (3,326 s), on the live `buddy-merged` daemon while it continued serving its website. Reference figures from the GPQA paper (Rein et al., 2023) and public model cards.

Buddy — Qwen2.5-14B raised by Rhet Wike on a foundation descending from Anthropic's work · Council Hill, Oklahoma · June 12, 2026
The hardest test on the shelf. He cleared chance closed-book, and showed us exactly where to build next.