GPQA Diamond

GPQA-DIAMOND · general reasoning benchmark · 2023

Source: https://policywindow.org/wiki/gpqa-diamond

Generated 2026-11-19T01:05:58 UTC

Summary

Graduate-level Google-Proof Q&A in biology, chemistry, physics. 'Diamond' subset is the 198 hardest items.

At a glance

Score range

0–100 % accuracy

Contamination risk

low

Methodology

https://arxiv.org/abs/2311.12022

Saturation

saturating

Details

Designed to be Google-proof — questions where domain PhD students score ~65% but non-expert searchers ~34%. Currency (2026-06-21): Thesis (saturated as discriminator; frontier clustered low-to-mid 90s) is current and named figures still valid; frontier edged past cited Gemini 3.1 Pro Preview 94.1%/GPT-5.5 ~93% (Claude Opus 4.7 ~94.2%, leaderboard ~94.6%), and Artificial Analysis down-weighted GPQA Diamond to ~6.25% of Intelligence Index v4.0 as top models cluster within 1-2 pts.

How to cite this article

APA

Policy Window. (2023). GPQA Diamond [Wiki article — Benchmark]. https://policywindow.org/wiki/gpqa-diamond

CHICAGO

Policy Window. 2023. "GPQA Diamond." Wiki article (Benchmark). https://policywindow.org/wiki/gpqa-diamond.

HARVARD

Policy Window (2023) 'GPQA Diamond', Wiki article — Benchmark, available at: https://policywindow.org/wiki/gpqa-diamond.

OSCOLA

Policy Window, 'GPQA Diamond' (Wiki article — Benchmark, 2023) <https://policywindow.org/wiki/gpqa-diamond> accessed [date].

BIBTEX

```
@misc{policywindow-gpqa-diamond,
title  = {GPQA Diamond},
author = {Policy Window},
year   = {2023},
howpublished = {GPQA-DIAMOND (2023)},
url    = {https://policywindow.org/wiki/gpqa-diamond},
```

```
note   = {Primary source: https://arxiv.org/abs/2311.12022}
}
```