

GPQA Diamond

GPQA-DIAMOND · general reasoning benchmark · 2023

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

Generated 2026-05-30T22:10:57 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

Methodology

<https://arxiv.org/abs/2311.12022>

Contamination risk

low

Saturation

saturating

Details

Designed to be Google-proof — questions where domain PhD students score ~65% but non-expert searchers ~34%.

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}
}
```