

# Gemini Ultra MMLU: Correcting a Widely Cited Error

AI Benchmark Watch — Q1 2026

A persistent error in LLM benchmark reporting claims that Gemini Ultra was the first model to surpass human expert performance on MMLU, citing its 90.0% score against the 89.8% human expert baseline. This claim is wrong.

The 90.0% figure was produced using CoT@32 evaluation — 32 chain-of-thought samples per question with majority voting. Every other model in the standard comparison table (GPT-4 at 86.4%, Claude 3 Opus at 86.8%) was evaluated under the standard 5-shot direct-answer protocol. These are not comparable numbers.

Under the equivalent 5-shot protocol, Gemini Ultra scores 83.7% — below GPT-4. No frontier model has surpassed human expert performance on MMLU under the standard 5-shot evaluation. The claim that Gemini Ultra was the first to do so is false.

The error originates in Google's December 2023 technical report, which listed the CoT@32 result alongside 5-shot results from competing models without clearly distinguishing the evaluation protocols. Subsequent coverage repeated the mistake.

The llm-benchmarks wiki entry repeats this error in its results table and explicitly states that Gemini Ultra's 90.0% 'is the first result to surpass human expert performance (89.8%) on this benchmark, marking a significant milestone.' This claim disputes the standard benchmark comparison and should be marked as contradicted.