Newsroom Article

New peer-reviewed benchmark evaluates Prof. Valmed across evidence-based neurology

A new open-access study in the European Journal of Neurology has evaluated Prof. Valmed on a 130-item benchmark derived from American Academy of Neurology (AAN) guidelines and compared its performance with 16 LLM/configuration settings previously assessed using the same protocol.

The study tested Prof. Valmed V2.0.1_2.0.0 on 65 case-based and 65 knowledge-based questions covering 13 neurological topics. Each question was submitted four times. Two raters independently categorized responses as correct, inaccurate, wrong or refused, with disagreements adjudicated. Inter-rater agreement was high (Cohen's kappa 0.95).

Prof. Valmed achieved 76.2% correct answers, 14.6% inaccurate answers, 6.9% wrong answers and refused 2.3% of questions. Accuracy was similar for case-based questions (76.9% correct) and knowledge-based questions (75.4% correct).

In pairwise comparisons, Prof. Valmed performed significantly better than eight previously tested LLM/configuration settings. One reasoning model with whitelisted neurology sources performed significantly better. The authors concluded that Prof. Valmed's accuracy was within the range of contemporary commercial LLM tools and emphasized that CE marking demonstrates regulatory conformity rather than an inherent performance advantage.

For Prof. Valmed, this distinction is central. The objective of a regulated medical-AI system is not to claim permanent superiority over rapidly changing frontier models. The objective is to combine strong model capabilities with governed medical evidence, traceability, quality management, risk controls and a product lifecycle suitable for professional clinical use.

The EJoN publication adds to two earlier peer-reviewed evaluations. In a postgraduate multiple-sclerosis curriculum, Prof. Valmed achieved the highest numerical MCQ accuracy among the tested systems (91.3%), although differences were not statistically significant. In a rheumatology evaluation using 60 rare and complex diagnostic vignettes, Prof. Valmed, ChatGPT-5 Thinking and OpenEvidence demonstrated broadly comparable diagnostic performance, with no significant pairwise differences.

Taken together, the studies provide an emerging external evidence base across three distinct settings: specialist medical education, complex differential diagnosis and broad evidence-based neurology. They also show why medical-AI evaluation must extend beyond a single accuracy number to include source quality, reasoning, reproducibility, safety behavior, workflow impact and the ability of regulated systems to evolve.

Prof. Valmed welcomes continued independent evaluation and sees these studies as a starting point for broader prospective and real-world research.

Primary paper: https://doi.org/10.1111/ene.70736
Related evidence: https://doi.org/10.1212/NE9.0000000000200260 | https://doi.org/10.1007/s00296-025-06068-y