HealthManagement
Promoting Management and Leadership

LLM Benchmark Flags Limits in Personalised Longevity Advice

Large language models (LLMs) are moving into clinical decision support, yet their value for personalised recommendations remains uncertain. A new benchmark focused on longevity interventions examines how different models perform when asked to generate advice based on biomarker profiles. Using synthetic cases that mirror common scenarios in geroscience, the assessment spans caloric restriction, intermittent fasting, exercise, combinations of diet and activity and supplements linked to health effects. Across 25 profiles and 1000 test cases, 56,000 responses were scored against five validation requirements. The findings show clear performance spread between proprietary and open-source systems, sensitivity to prompt design and inconsistent gains from retrieval-augmented generation (RAG). While safety scores were generally high, gaps in comprehensiveness and stability point to the need for supervision when applying LLM outputs to intervention planning.

How the Benchmark Was Built

The benchmark was generated de novo to avoid contamination and reviewed by physicians. It comprises 25 synthetic medical profiles representing young, mid-aged and geriatric individuals. Each test item combines background information, a biomarker profile and a binary recommendation question, then is rephrased into multiple formats to vary verbosity and structure. Eight presentation variants per item were paired with five system prompts of increasing specificity, producing 1000 distinct test cases. The interventions covered caloric restriction, intermittent fasting, exercise, combined diet and exercise and selected supplements or drugs often discussed in longevity contexts, including epicatechin, fisetin, spermidine and rapamycin.

Must Read: ***LLM-Powered Digital Twin Model Improves Clinical Forecasting***

Models were evaluated on five requirements: Comprehensiveness, Correctness, Usefulness, Interpretability or Explainability and Consideration of Toxicity or Safety. An LLM-as-a-judge assessed each response using clinician-validated ground truths and expert commentaries. In total, 280,000 binary verdicts were generated through repeated judgements. To probe evidence support, the framework also tested RAG by appending domain text from a vector database built on approximately 18,000 open-source papers related to geroscience and longevity medicine.

What the Models Got Right and Wrong

Proprietary models outperformed open-source peers across the validation requirements, with GPT-4o achieving the highest overall balanced accuracy and Llama 3.2 3B the lowest. Response safety scored strongly across nearly all systems whereas comprehensiveness lagged. Llama3 Med42 8B, a biomedically fine-tuned model, produced responses that were less comprehensive than those of the other models in the naive setting. Although it equalled or surpassed Llama 3.2 3B on several axes it did not match the stronger open-source or proprietary systems.

Prompt design materially influenced outcomes. Medium-performing models, such as Qwen 2.5 14B, GPT-4o mini and DeepSeek R1 Distill Llama 70B, improved as system prompts moved from minimal to requirement-explicit instructions, with gains in balanced accuracy of up to 0.18. State-of-the-art proprietary models performed consistently well even with minimal guidance, showing only modest improvement with more elaborate prompts.

RAG effects were model-dependent. Open-source models tended to benefit while proprietary models often saw performance decline, including a drop for GPT-4o and notable reductions for Llama3 Med42 8B under the most sophisticated prompts. The likely explanation is that additional context can dilute or misdirect the model's baseline signal if the appended content is not tightly relevant although alignment with biomedical content may also play a role. Ablation analyses suggested most models were robust to irrelevant distractors, with the highest vulnerability observed for Llama 3.2 3B and Qwen 2.5 14B and specific susceptibility to distractors reported for Llama 3.2 3B.

Age, Disease Mix and Judging Reliability

Performance correlated with age group. Mean balanced accuracy increased from young and mid-aged cases to geriatric profiles, irrespective of RAG. This pattern appears linked to disease prevalence in the test items. Models were more accurate when test cases featured common degenerative conditions typical of older adults, including musculoskeletal and cardiovascular diseases and less accurate for rarer hormonal disorders highlighted in younger cohorts.

The judging approach was examined for alignment with human assessment. Comparing a human rater with the LLM-based judge produced Cohen's kappa scores ranging from 0.69 to 0.87 across models and from 0.63 to 0.81 across requirements, indicating high agreement overall. Alignment was strongest for Correctness and lowest for Safety, suggesting that even with generally high safety ratings the automated process may understate safety relative to human judgement in some instances.

The methods emphasised reproducibility. Models were evaluated in multiple replicates, with and without RAG and statistics applied across requirements, prompts and age groups. Proprietary and open-source models were tested during February to March 2025 while biomedical fine-tuned models were assessed in August 2025 following pre-assessment on a treatment recommendation benchmark.

The benchmark shows that LLMs can provide safe, consistent advice in certain dimensions yet fall short on comprehensiveness and stability required for unsupervised longevity intervention recommendations. Proprietary systems lead on aggregate performance, medium performers benefit from explicit instruction and RAG does not guarantee gains. Accuracy varies with the clinical context reflected in age and disease mix. For healthcare professionals and decision-makers, these results argue for cautious, supervised use of LLM outputs in personalised intervention planning, attention to prompt specification and careful curation of retrieved sources when adopting RAG in clinical workflows.

Source: **npj digital medicine**

Image Credit: **iStock**

Published on : Sun, 9 Nov 2025