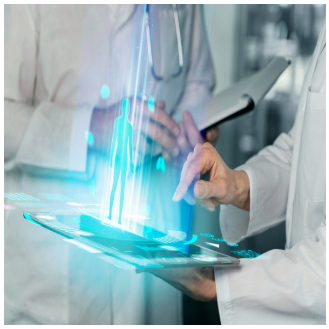
HealthManagement
Promoting Management and Leadership

LLM-Powered Digital Twin Model Improves Clinical Forecasting



Clinical forecasting underpins patient monitoring, treatment selection and trial execution, yet electronic health records often contain sparsity, noise and missingness that undermine model accuracy. A generative large language model approach termed Digital Twin–Generative Pretrained Transformer (DT-GPT) models longitudinal patient trajectories by fine-tuning a biomedical LLM on text-encoded histories without imputation or normalisation. Evaluated across intensive care, non-small cell lung cancer and Alzheimer's disease, the method reduced error relative to strong machine learning baselines, preserved clinically meaningful distributions and inter-variable relationships, and retained a conversational interface that supports preliminary interpretability. Beyond the fine-tuned targets, the model produced zero-shot forecasts for additional variables, demonstrating adaptability while delineating limits in rare, high-variance events where sensitivity remains a challenge.

State-of-the-Art Trajectory Forecasting Across Cohorts

DT-GPT forecasts multivariable trajectories by ingesting text-encoded electronic records and study data, then generating time-evolving predictions. The evaluation covered three horizons matched to clinical contexts. In non-small cell lung cancer, six laboratory measures were predicted weekly for up to 13 weeks after therapy initiation using all baseline data. In intensive care, hourly forecasts over the next 24 hours were produced for oxygen saturation, respiratory rate and magnesium from the first 24 hours of observations. In Alzheimer's disease, cognitive scores were forecast at 6-month intervals over 24 months from baseline measures.

Must Read: ***Spatial and Synthetic Data Power Healthcare Digital Twins***

Against 14 multi-step multivariate baselines, DT-GPT achieved the lowest scaled mean absolute error overall. Average scaled error improved by 3.4 % versus LightGBM in lung cancer, 1.3 % in intensive care and 1.8 % versus Temporal Fusion Transformer in Alzheimer's disease, with errors staying below target standard deviations. The approach outperformed a channel-independent patch transformer and time-series LLMs that treat variables separately, benefiting tasks where sparsity and cross-variable dependence matter. Non-fine-tuned LLMs, including a larger 32-billion parameter model, underperformed and hallucinated outputs, underscoring the value of domain-specific fine-tuning.

Performance robustness extended across metrics. DT-GPT captured trajectory trends and delivered competitive results on MAE, MASE, SMAPE and rank correlations. Classification-style checks indicated credible detection of routine yet informative abnormalities and medium-horizon trends. Detection reached strong areas under the curve for mild anaemia and elevated lactate dehydrogenase, and 3-week trend directions were captured for haemoglobin and inflammatory cell counts. In contrast, critically low haemoglobin and high leukocyte events were harder to forecast, reflecting their low prevalence and variance and indicating where specialised methods or targeted data may be required.

Preserving Clinical Signal, Distributions and Relationships

Longitudinal forecasting is clinically meaningful only if it retains realistic value distributions and the structure linking variables. DT-GPT preserved cross-correlations between forecasted variables and ground truth with coefficients of determination up to 0.99, matching or surpassing the best baseline. Time-point-wise comparisons showed consistent advantage over LightGBM across lung cancer and intensive care horizons. Distributional fidelity, assessed via the Kolmogorov–Smirnov statistic, favoured DT-GPT versus recent deep baselines, with predicted histograms closely resembling observed values. This property is not sufficient for utility on its own but remains necessary for plausible clinical simulation.

The generative setup enables multiple sampled futures per patient. Final predictions were computed by averaging 30 generated trajectories. Analysis showed that aggregation can mask salient individual dynamics, with a minority of patients producing right-skewed errors. A hypothetical selection of best-matching trajectories per patient lowered error substantially without retraining, suggesting that improved aggregation or

arbitration strategies could yield further gains. Variance across generated samples contributed to outliers, while the majority of predictions tracked observed paths closely, supporting the model's use for scenario-style trajectory exploration alongside point estimation.

Robustness, Explainability and Zero-Shot Reach

Real-world clinical data contain substantial missingness, heterogeneous encodings and textual noise. DT-GPT sustained competitive performance with training subsets around 5000 patients and remained stable when additional missingness was injected on top of a high baseline, with notable degradation only after more than one fifth of extra input masking. The model tolerated misspellings in inputs, with marked declines only at high perturbation levels. Conventional approaches typically require normalisation and imputation or discard corrupted inputs, whereas text-based encoding allowed DT-GPT to operate without these preprocessing stages.

Interpretability was explored through the retained chat interface, which surfaced variables most influential for forecasts across patient samples. Therapy influenced haemoglobin trajectories, with immunotherapy and targeted therapy associated with higher values over time compared with regimens that included chemotherapy, where declines were common. Performance status (ECOG) and leukocyte counts ranked highly, and age was recognised as an important prognostic factor. These patterns aligned with observed data distributions, offering face validity while acknowledging that conversational rationales may not reflect causal pathways.

Generalisability was probed through zero-shot forecasting of 69 additional variables present in histories but excluded from training targets. A single DT-GPT model outperformed separately trained LightGBM models on 13 variables. Gains were concentrated where non-target variables were closely related to fine-tuned targets, such as neutrophil measures. Where strong correlations were absent, reported relationships and ratios suggested the model may capture latent clinical structure, though these remain hypotheses that require careful validation. Practical constraints include sequence length limits that cap the number of variables forecast simultaneously and the need to improve sensitivity to rare but high-risk events. Known LLM risks such as hallucination and embedded dataset biases further reinforce the importance of human oversight and careful deployment.

DT-GPT shows that fine-tuned language models can generate clinically coherent patient trajectories across short-, medium- and long-term horizons, improving error rates against strong baselines while preserving distributions and inter-variable relationships. The approach tolerates missingness and textual noise, offers preliminary interpretability through a conversational interface and extends to related variables without additional training where relationships exist. Performance is strongest for routine abnormalities and trend detection, with rare acute events remaining challenging. For healthcare professionals and decision-makers, these results indicate a practical route to digital twin applications in monitoring, treatment planning and trial support, provided that validation, aggregation strategies and safeguards for rare-event sensitivity and bias are addressed.

Source: **npj Digital Medicine**

Image Credit: **Freepik**

Published on : Thu, 16 Oct 2025