Large language models forecast patient health trajectories enabling digital twins

Authors: ANDREA PEREZ VALLE, GABRIEL TORRES ZAMORA, IÑIGO ARRIAZU GARCIA

Introduction

In this work we selected the article *"Large language models forecast patient health trajectories enabling digital twins"* (2025) because it represents a novel and clinically relevant step toward AI-driven personalized medicine. The paper introduces DT-GPT, a large language model (LLM) fine-tuned on electronic health records (EHRs) to forecast individual patient trajectories over time, effectively acting as a digital twin of the patient.

The model was evaluated across oncology, intensive care, and Alzheimer's cohorts, where it outperformed 14 state-of-the-art forecasting baselines and reduced prediction error while preserving realistic clinical correlations between variables.

Additionally, DT-GPT can generate zero-shot predictions for new clinical variables it was not explicitly trained on, and it provides human-readable explanations (chatbot-style interface) for its forecasts, ensuring clinical interpretability highlighting influential factors such as therapy, age, ECOG status, etc.

Using electronic health records (EHRs) as its primary data source, DT-GPT leverages the vast amount of dynamically collected patient information, not only from these records but also by imputing data. Because it combines technical innovation (LLM-based forecasting), clinical utility (patient-specific trajectory prediction), and future applicability (digital twins for treatment planning, monitoring, and trial support), this article is an ideal choice to analyze in depth.

By exploring this, the paper aims to demonstrate a new paradigm for digital twins in healthcare, using LLMs to provide multivariate, longitudinal predictions along with potential interpretability and zero-shot adaptability.

Scientific Question

The core question addressed by Makarov et al. (2025) is whether large language models (LLMs) can be used to create clinical "digital twins" by leveraging electronic health record (EHR) data. The research questions can be summarized as:

- *Can a fine-tuned LLM (DT-GPT) forecast comprehensive longitudinal patient trajectories from real-world EHR data, outperform traditional models, and even predict clinical variables it was never trained on (i.e. patient trajectories)?*
- *Are LLM-based solutions able to handle real-world Electronic Health Record (EHR) challenges such as missing data, limited sample size, and data sparsity?*

Methodology (que hicieron para responder la pregunta)

The study compares DT-GPT against 14 different methods (LSTM, *LightGBM*, Transformers, etc.) across three datasets: Non-Small Cell Lung Cancer (NSCLC), Intensive Care Unit (ICU), and Alzheimer's Disease Neuroimage Initiative (ADNI).

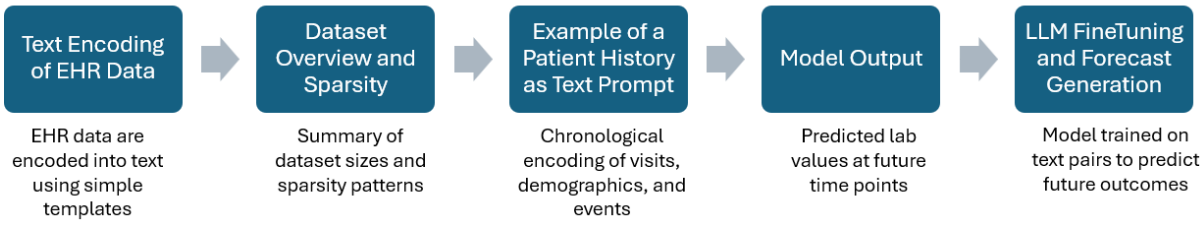

Image 1. DT-GPT pipeline.

The DT-GPT model is essentially a fine-tuned LLM (built on the 7-billion-parameter *BioMistral* biomedical GPT model) trained to output future patient data given past EHR data. The approach is straightforward: patient histories are converted into a text sequence (prompt), and the model is fine-tuned to generate the next sequence representing future observations. Specifically, the authors designed a templated encoding of the EHR: each patient's input prompt includes (1) a chronological narrative of clinical events (visits with recorded measurements and interventions), (2) static demographic information, (3) the forecast horizon definition (the future time points to predict), and (4) a prompt cue asking the model to forecast the patient's health trajectory. Then, the output is formatted as structured text (e.g. a JSON list of the predicted values at each future time) for the target clinical variables.

This text-to-text setup allowed the team to fine-tune the LLM using standard next-token prediction (cross-entropy loss) – effectively treating multivariate time-series forecasting as a language generation task. Notably, no architectural changes to the LLM were

required; DT-GPT is model-agnostic and could be applied to any text-focused LLM without bespoke network modifications. The authors ran the fine-tuning with the output portion masked for loss computation (so the model learns to output the future data given the prompt) and, during inference, they generated multiple output sequences per patient and averaged them to produce the final prediction. This ensembling of 30 generations was done to stabilize the inherently stochastic LLM outputs.

A key difference from traditional ML approaches is that DT-GPT deliberately avoids heavy preprocessing of time-series data. Missing values are simply omitted (no imputation), relying on the model to interpret their absence contextually.

Normalization of inputs was not needed for the LLM (in fact, the authors only standardized the data for non-LLM baseline models) By encoding *all* variables together in the prompt, DT-GPT inherently models the dependencies between variables, addressing the *channel independence* limitation of many time-series models. In contrast, traditional models like separate regressors or channel-independent transformers treat each variable in isolation and then combine results, potentially missing cross-feature interactions. Prior approaches often introduce specialized architectures or imputation layers to cope with EHR irregularity and sparsity, increasing complexity and embedding specific assumptions about the data. DT-GPT's approach is to instead leverage the LLM's sequence modeling capacity and vast pre-trained biomedical knowledge (via *BioMistral*) to naturally handle irregular, noisy clinical data. The only notable data preprocessing was an outlier filtering step to clip extreme values (beyond 3 standard deviations) for all models, ensuring the training data's noise was bounded Overall, the methodology centers on using a pre-trained foundation model and minimal data wrangling: feeding raw clinical histories as text, which allows the model to learn the patient's trajectory in a holistic manner. This design contrasts with conventional ML pipelines by not requiring explicit handling of each missing value, no manual feature engineering or complex custom networks, and by exploiting the pre-trained LLM's ability to interpret clinical text and numeric values in context.

Findings

DT-GPT delivered strong results across all three datasets, outperforming the conventional models on the main forecasting tasks. Notably, it achieved the lowest forecasting error (scaled MAE) in each case, indicating state-of-the-art accuracy. According to the authors' aggregate metric, DT-GPT reduced the normalized error by 3.4% in the NSCLC cohort, 1.3% in the ICU, and 1.8% in ADNI relative to the best baseline methods. While the percentage improvements might seem modest, they reflect meaningful gains given that many baseline models were already highly optimized; in NSCLC especially, DT-GPT's multi-variable approach clearly

outperformed existing techniques. The differences were statistically significant in the larger cohorts (e.g. $p < 1e\text{-}16$ vs. the next-best model in NSCLC). Beyond just error magnitude, the model predictions better preserved clinical patterns – the distribution of predicted lab values and their inter-variable correlations closely matched those of the real data. This is a crucial finding: unlike channel-independent models that might break the relationships between labs, DT-GPT's joint modeling meant that if, say, hemoglobin and neutrophils usually move together in real patients, the model's forecasts reflected that coupling. Indeed, the authors reported that the correlation matrix of DT-GPT's forecasted variables was almost indistinguishable from the true correlation matrix (R^2 almost 0.98) on NSCLC and ICU, outperforming even strong baselines like *LightGBM* in capturing those relationships. Similarly, DT-GPT maintained the realistic distribution of outcomes (it had the lowest Kolmogorov–Smirnov statistic among models, indicating its predicted values followed the true data distribution most closely). These results suggest that the LLM approach not only predicts accurate numbers but also behaves more like a real patient data generator, an important aspect for digital twin fidelity.

Summarization of the meaningful findings:

- Zero-Shot Forecasting Capability
- Preservation of Clinical Patterns
- Interpretability and Explainability

Limitations

Despite its promising performance, the DT-GPT approach has several **limitations and challenges** acknowledged by the authors:

- **Difficulty with Rare or Critical Events:** The model struggled to accurately predict low-prevalence, high-variance outcomes – for instance, an abrupt and critical drop in hemoglobin indicating a major bleeding event was often missed (DT-GPT's ROC AUC was nearly 0.5 for predicting extremely low hemoglobin). Rare outcomes like these constituted a very small fraction of the training data (e.g. only nearly 1% of NSCLC patients had hemoglobin <7.5 g/dL), so the model was not sensitive to them. The authors note that improving detection of such high-risk events may require specialized techniques beyond standard training, such as tailored loss functions that give more weight to rare events, anomaly detection methods, or incorporating additional data sources (e.g. unstructured clinical notes that might contain warning signs) In its current form, DT-GPT is better at forecasting typical trends than catching out-of-the-ordinary acute events.

- **Potential for Hallucination:** As with any LLM, DT-GPT can produce outputs that look plausible but are not grounded. The authors caution that the model may sometimes "hallucinate" in its explanations or even outputs for example, when the chatbot feature is used to explain predictions, the model might confidently attribute a trend to a factor that wasn't present or relevant in the data. This is problematic in the medical domain, as a clinician might be misled by a convincing but incorrect explanation. While DT-GPT's primary numeric forecasts are constrained by training data (making outright random hallucinations in the numbers unlikely), any time it's used in a generative text mode (for user interaction) there is a risk of fabricated information. The authors emphasize the need for a "human-in-the-loop" approach – clinicians should oversee and validate the outputs, and not blindly trust the model's justifications Proper user training and developing methods to make the model's reasoning more truthful (perhaps via additional fine-tuning or grounding in knowledge bases) are suggested steps to mitigate this issue.