

ИНФОРМАТИКА И КОМПЮТЪРНИ НАУКИ
INFORMATICS AND COMPUTER SCIENCES

**DEPLOYING AI AGENTS FOR SLEEP APNEA MANAGEMENT IN A U.S. DIGITAL
HEALTH PLATFORM: ARCHITECTURE, LLM SELECTION,
AND BUSINESS VALUE**

Krasimir Nikolov

University of Library Studies and Information Technologies

<https://doi.org/10.70300/TIVP9872>

Abstract: Sleep apnea remains a major and frequently underdiagnosed public health challenge. Epidemiological estimates indicate that obstructive sleep apnea (OSA) affects nearly 936 million adults worldwide (Benjafield et al. 2019). Home Sleep Apnea Testing (HSAT) devices have expanded the possibilities for out-of-laboratory assessment and generate structured clinical data that may support intelligent workflow automation. This paper presents a reference technical architecture for deploying AI agents within a U.S. digital health platform specializing in sleep apnea care. The proposed system is built around a self-hosted large language model, specifically Llama 3.1 8B, served via vLLM on AWS GPU infrastructure and integrated with an existing Node.js/NestJS backend through a standardized REST API. To support a HIPAA-aligned deployment, the agent is restricted to structured patient data such as screening questionnaire results, AHI values, and CPAP adherence indicators and does not access raw physiological signals. The paper examines the infrastructure configuration, LLM selection criteria, and potential business value. Under the stated AWS pricing assumptions, the core GPU and model-storage infrastructure is estimated at approximately \$295 per month; this figure is illustrative and does not represent total cost of ownership. The proposed architecture requires task-specific clinical, security, and regulatory validation before production use. **Keywords:** AI agents; sleep apnea; LLM deployment; digital health architecture; HIPAA compliance

INTRODUCTION

Sleep apnea represents a significant, yet frequently underaddressed, public health challenge in modern medicine. Epidemiological estimates indicate that obstructive sleep apnea (OSA) affects nearly 936 million adults worldwide (Benjafield et al. 2019). A substantial proportion of affected individuals remain undiagnosed, contributing to delayed treatment and an increased burden on healthcare systems. Left untreated, OSA is associated with serious comorbidities. The condition substantially elevates the risk of cardiovascular disease, type 2 diabetes, and hypertension, while simultaneously impairing cognitive performance over time (Lévy et al. 2015).

Traditionally, in-laboratory polysomnography (PSG) has served as the diagnostic reference standard for OSA. However, this procedure presents considerable practical barriers. It carries high costs, creates logistical burdens, and may contribute to delays in diagnostic assessment and treatment initiation (Kapur et al. 2017). The introduction of Home Sleep Apnea Testing (HSAT) devices – such as the WatchPAT series developed by Itamar Medical – has substantially altered this landscape. These devices can expand access to sleep-disordered breathing assessment by allowing eligible patients to complete diagnostic testing in their home environment, with results subsequently reviewed remotely by credentialed clinicians (Corral et al. 2017).

Digital health platforms that coordinate the end-to-end HSAT workflow naturally accumulate structured datasets throughout the patient journey. From initial intake and insurance verification to study fulfillment, clinical review, and CPAP treatment initiation, these platforms generate records covering screening questionnaires, diagnostic outcomes, treatment assignments, and adherence metrics. Collectively, such structured data may provide a suitable foundation for intelligent and automated clinical workflow support.

Artificial intelligence agents offer a highly relevant mechanism for augmenting these clinical operations (Wang et al. 2024). In contrast to static, rule-based automation systems, AI agents can perform generate context-dependent action sequences rather than executing fixed rules over structured inputs, interact with predefined tools, and execute complex workflows according to specified instructions and constraints. Their behavior can be conditioned on the clinical context represented in the available data. When cases require human judgment, the agent can be configured to escalate them to qualified clinical personnel, while routine monitoring tasks can be performed continuously.

This paper addresses a focused research question: How can AI agents be technically designed, deployed, and integrated within a U.S. digital health platform for sleep apnea care while maintaining strict patient data privacy and supporting compliance with applicable HIPAA requirements? Three primary contributions are offered: (1) a reference architecture for a self-hosted, LLM-based AI agent operating within a Node.js/NestJS healthcare environment; (2) a comparative evaluation of open-weight language models suited for regulated clinical deployments; and (3) a practical mapping of AI agent capabilities to real-world workflow states within the sleep apnea care pathway.

RESEARCH METHODOLOGY

The study adopts a Design Science Research (DSR) methodology – a well-established framework in information systems and applied computer science that focuses on the construction and evaluation of purposeful technical artifacts (Hevner et al. 2004). This approach was selected because the research objective extends beyond descriptive analysis: the aim is to construct a technically viable solution to a real-world operational problem and evaluate its utility against a defined set of criteria (Peppers et al. 2007).

The investigation was structured across three sequential phases. In the first phase, a targeted literature review was conducted covering AI agent architectures, large language model deployment in healthcare, HIPAA-aligned infrastructure patterns, and sleep apnea clinical workflows. Sources were drawn from IEEE Xplore, PubMed, the ACM Digital Library, and relevant vendor technical documentation. The review was intended to identify technically and clinically relevant concepts rather than to constitute a systematic review or meta-analysis. In the second phase, a comparative technical evaluation of three open-weight language models was carried out based on a defined set of clinical deployment criteria. In the third phase, a reference architecture was synthesized and mapped to an existing clinical data model representative of a typical U.S. sleep apnea digital health platform.

The LLM evaluation was structured around five criteria of particular relevance to regulated healthcare environments. These include: (1) licensing permissiveness for commercial use and applicable use restrictions; (2) data residency, meaning the ability to operate fully within a private cloud without outbound transmission of Protected Health Information (PHI) to an external model provider; (3) reasoning and tool-calling capability for reliable multi-step task execution; (4) computational footprint, specifically VRAM requirements and compatibility with cost-effective GPU instance types; and (5) governance and regulatory considerations associated with model provenance, licensing, organizational policy, and deployment jurisdiction.

Infrastructure assessment was based on technical specifications and indicative AWS pricing for GPU-enabled instance types, using the g4dn.xlarge (16 GB NVIDIA T4) as the minimum development configuration and the g5.xlarge (24 GB NVIDIA A10G) as the recommended production setup. Model inference was assessed conceptually through a deployment-oriented technical analysis using vLLM, an open-source high-throughput inference engine, deployed as a containerized ECS task within a HIPAA-eligible AWS environment. Cost projections were calculated on a 30-day continuous operation basis, incorporating both on-demand and indicative spot instance pricing.

The clinical workflow analysis was conducted against a representative PostgreSQL schema modeled with Prisma ORM, reflecting the data architecture of a typical U.S. sleep apnea digital

health platform. Key entities examined included patient screening results, HST workflow state machines, diagnostic report structures, treatment records, and compliance indicators. Through this analysis, specific agent trigger points were identified – discrete state transitions where an AI agent can meaningfully intervene without requiring any access to raw physiological signal data from HST devices.

TECHNICAL ARCHITECTURE

The proposed AI agent system is architected as a modular, HIPAA-aligned extension to an existing Node.js/NestJS digital health backend. A micro-services pattern was adopted, whereby the LLM inference engine operates as a fully independent containerized service, structurally decoupled from the core application layer. This separation has two critical benefits: it reduces disruption to established clinical workflows and allows the inference engine to be scaled or replaced without requiring redeployment of the main application. The system is organized into four functional layers: an inference layer responsible for hosting and serving the language model; an agent orchestration layer implemented as a dedicated `ai/` module within the NestJS backend; a data access layer providing the agent with structured read access to patient workflow states and clinical summaries; and a human oversight layer that routes all agent outputs to clinical staff for mandatory review prior to any patient-facing action.

The language model is served through vLLM – a high-throughput, memory-efficient inference engine optimized for transformer-based models – deployed as a Docker container within an AWS ECS task. Model weights are stored on an Amazon EFS volume mounted to the task, ensuring persistence across container restarts and eliminating redundant downloads upon reinitialization. For production environments, a `g5.xlarge` instance equipped with a 24 GB NVIDIA A10G GPU is proposed as a baseline configuration. At 4-bit quantization, the evaluated model class can typically be accommodated within this GPU memory envelope, although exact VRAM requirements depend on model architecture, quantization method, context length, KV-cache allocation, batching strategy, and inference configuration. The inference service exposes a chat completions REST endpoint on port 8000, accessible exclusively within the private VPC. No inference traffic needs to be routed to an external model provider, thereby supporting containment of PHI within the organization’s controlled AWS environment.

The agent orchestration logic resides within a dedicated `ai/` module embedded in the existing NestJS application. This module executes a continuous tool-calling loop. On each invocation, the agent receives a system prompt embedding HIPAA guardrails, a task description, and a set of available tool definitions expressed as JSON schemas. The model reasons over the supplied context, then emits either a conclusive response or a structured tool invocation request. When a tool call is detected, the NestJS module executes the corresponding internal service method – such as retrieving a patient’s current HST workflow state, fetching screening questionnaire scores, or querying CPAP compliance data – and feeds the result back as a new context message (Wang et al. 2024). The loop continues until a final output is produced or a maximum iteration safeguard is reached.

A defining architectural constraint is that the agent operates exclusively on structured summary data. It has no access to raw physiological signals from HSAT devices, which remain within the clinical device ecosystem. This restriction intentionally reduces the PHI surface area, simplifies audit-trail requirements, and supports the HIPAA principle of minimum necessary access. Agent trigger points are mapped to discrete state transitions in the patient care pathway – spanning device ordering, study completion, clinical review, and treatment initiation. In every case, all agent outputs are reviewed by clinical staff before any downstream action is executed. This human-in-the-loop design functions as a risk-control mechanism and is consistent with broader principles applied to clinical decision-support and AI-enabled medical software; however, the regulatory classification of any specific implementation must be assessed separately.

Table 1. Infrastructure configuration and indicative cost comparison

Parameter	Minimum (Dev)	Recommended (Prod)	Indicative Spot (Prod)
Instance type	g4dn.xlarge	g5.xlarge	g5.xlarge
GPU	NVIDIA T4 (16GB)	NVIDIA A10G (24GB)	NVIDIA A10G (24GB)
Hourly cost	\$0.526	\$1.006	~\$0.40
Monthly cost	~\$380	~\$725	~\$295
Model storage (EFS)	~\$5/mo	~\$5/mo	~\$5/mo
Total/month	~\$385	~\$730	~\$295

LLM SELECTION AND COMPARATIVE ANALYSIS

Selecting an appropriate language model for a HIPAA-regulated healthcare environment requires evaluation criteria that extend well beyond standard benchmark performance. In the context of a self-hosted deployment, factors such as licensing terms, model provenance, the capacity for multi-step task execution, computational efficiency, governance requirements, and compatibility with organizational security policies may carry equal or greater weight (Moor et al. 2023).

Llama 3.1 8B Instruct, developed by Meta, represents a technically attractive candidate for self-hosted U.S. healthcare deployments because of its manageable computational requirements, broad software ecosystem, and availability for local inference. Its licensing framework permits a wide range of commercial uses, subject to the terms and restrictions of the Llama 3.1 Community License (Meta 2024). Its reliability for specific clinical tool-use workflows should be established empirically using task-specific evaluation data rather than assumed from general-purpose benchmarks. Mistral 7B Instruct v0.3, developed in France by Mistral AI, operates under the Apache 2.0 license (Mistral AI 2024). Its relatively small computational footprint may offer efficiency advantages for resource-constrained deployments. Its European provenance may also be operationally relevant for organizations whose procurement, governance, or cross-border data-transfer policies favor EU-based technology providers; however, GDPR does not impose a general requirement that personal data remain physically stored within the European Union.

DeepSeek-R1-Distill-Llama-8B presents a distinct technical profile. As a distilled derivative of the DeepSeek-R1 reasoning model based on Llama 3.1 8B, it is designed to retain reasoning-oriented capabilities within an approximately 8B-parameter model class (DeepSeek-AI 2025). Its suitability for complex tool-based clinical workflows is promising but should be established through task-specific benchmarking before conclusions about superiority over Llama or Mistral models are drawn. Its development by a Chinese organization may trigger additional organizational, procurement, cybersecurity, governance, or legal review depending on the policies and regulatory obligations applicable to the deploying healthcare organization. Model provenance alone, however, should not be treated as evidence of regulatory non-compliance.

Table 2. LLM candidate evaluation matrix

Criterion	Llama 3.1 8B	Mistral 7B	DeepSeek-R1-Distill-Llama-8B
License	Llama Community	Apache 2.0	MIT; derived from Llama 3.1
Provider jurisdiction	USA	France / EU	China
VRAM (4-bit quant)	~6 GB	~5 GB	~6 GB
Single-hop tool calls	Suitable; validation required	Suitable; validation required	Suitable; validation required
Multi-step chains (3+)	Requires task-specific testing	Requires task-specific testing	Reasoning-oriented; validation required
Governance considerations	Established U.S. ecosystem; organizational review required	EU-based provider; organizational review required	Additional institutional review may be appropriate
Recommendation	Primary candidate	Alternative candidate	Benchmark candidate

Based on the above evaluation, Llama 3.1 8B Instruct is proposed as the initial production candidate, subject to task-specific validation and organizational approval. DeepSeek-R1-Distill-Llama-8B is recommended for parallel benchmarking on equivalent infrastructure. Because the architecture abstracts the inference layer from the core NestJS application, model substitution can be achieved with limited changes to the surrounding orchestration logic. Should benchmarking demonstrate superior task-level performance and the relevant governance requirements be satisfied, DeepSeek-R1-Distill-Llama-8B could represent an alternative deployment option without requiring substantial architectural modification. Mistral 7B remains a secondary option for organizations that prioritize Apache 2.0 licensing, lightweight deployment, or specific European procurement and governance considerations.

BUSINESS VALUE AND RETURN ON INVESTMENT

The deployment of AI agents in a sleep apnea digital health platform has the potential to generate value across three dimensions: operational efficiency, clinical scalability, and revenue-cycle support. Traditional care pathways depend heavily on manual touchpoints – patient intake coordination, insurance eligibility verification, HSAT device logistics, clinical review scheduling, and treatment follow-up. Each of these steps consumes staff time that tends to increase with patient volume. An AI-assisted system, by contrast, can automate or support selected repetitive workflow-monitoring tasks at comparatively low incremental compute cost once the infrastructure is operational.

The illustrative compute and storage cost of the proposed self-hosted system amounts to approximately \$295 per month under the spot-pricing scenario shown in Table 1. Because this infrastructure cost does not necessarily increase linearly with patient volume, the compute cost per monitored patient may decline as utilization increases, provided that system capacity is not exceeded. For a platform serving 500 active patients monthly, the estimated infrastructure expenditure corresponds to approximately \$0.60 per patient. This calculation represents infrastructure cost only and excludes software engineering, maintenance, security operations, monitoring, networking, backups, clinical review time, and other organizational costs.

A deployed AI agent system can process multiple workflow events concurrently, subject to the throughput, latency, context-length, batching, and GPU-capacity limits of the selected deployment configuration. Rather than replacing clinical coordinators, the agent can continuously monitor predefined workflow state transitions and prioritize cases for human attention. This capacity may be especially relevant given the documented workforce constraints in sleep medicine, which contribute to persistent bottlenecks in OSA diagnosis and treatment pathways (Kirsch 2018). The AI agent does not replace clinical judgment; it functions as a workflow-support mechanism that may allow clinical teams to allocate more attention to cases requiring professional assessment.

The proposed architecture may support documentation and monitoring activities associated with Remote Physiologic Monitoring (RPM) under the CMS reimbursement framework. Relevant CPT codes include 99453, 99454, and 99457 (CMS 2024). Whether a specific workflow qualifies for reimbursement depends on the applicable CMS coverage, billing, device, monitoring, and clinical-interaction requirements and cannot be determined solely by the presence of an AI agent. The agent can support documentation, workflow tracking, and identification of missing adherence data, while billable clinical services must continue to satisfy all applicable human-service and documentation requirements. Furthermore, by monitoring available CPAP usage data, the agent may identify patients approaching the CMS adherence threshold – PAP use of at least four hours per night on 70% of nights during a consecutive 30-day period within the first three months of initial use – and flag them for proactive clinical outreach (CMS n.d.).

From a risk management perspective, the fully self-hosted deployment model reduces dependency on external AI inference providers and may reduce the number of third parties processing PHI. This does not itself establish HIPAA compliance. Compliance depends on the broader administrative, physical, and technical safeguards implemented by the covered entity or business associate, including appropriate contractual arrangements, access controls, encryption, logging,

monitoring, incident response, and workforce procedures. As U.S. state-level health data privacy legislation continues to expand beyond federal HIPAA requirements, this architecture may provide greater organizational control over data flows and processing locations (Goldstein et al. 2023). The mandatory human-in-the-loop design reduces, but does not eliminate, risks associated with autonomous AI-generated recommendations in clinical environments.

DISCUSSION

The proposed architecture demonstrates that technically capable, privacy-preserving AI agents for sleep apnea care management can be designed using open-weight models and self-hosted infrastructure in a manner that avoids routine transmission of PHI to third-party model API providers.

Several architectural decisions are particularly important. The restriction of the agent to structured summary data – excluding raw physiological signals from HSAT devices – represents both a deliberate constraint and a privacy advantage. It significantly reduces the PHI surface area, simplifies audit-trail requirements, and supports a privacy-by-design and compliance-oriented architecture. The tool-calling loop implemented within the NestJS ai/ module provides a flexible extensibility pattern. Because each available tool is defined as a discrete JSON schema, new capabilities can be added without modifying the core orchestration logic – a pattern that aligns with established principles of maintainable healthcare software architecture (Blobel 2018).

The LLM evaluation reveals that no single model can be considered universally optimal across all relevant criteria. Llama 3.1 8B offers a potentially strong combination of ecosystem maturity, manageable computational requirements, and suitability for structured tool execution in U.S.-based deployments. DeepSeek-R1-Distill-Llama-8B presents a potentially valuable reasoning-oriented alternative for multi-step workflows, but its comparative advantage requires direct empirical validation using the target clinical tasks. This finding highlights a recurring tension in healthcare AI adoption: strong general-purpose reasoning performance does not necessarily provide the optimal balance of accuracy, governance, licensing, operational maturity, computational cost, and regulatory suitability for a particular clinical environment. Technical and governance criteria should therefore be evaluated jointly throughout the model-selection process.

Several limitations of the proposed system require acknowledgment. First, the architecture was designed and evaluated at the level of technical specification; empirical validation through production deployment and clinical outcome measurement remains future work. Second, the agent’s effectiveness is bounded by the completeness, accuracy, and timeliness of data entered into the underlying clinical information systems – a limitation common to EHR-integrated decision-support systems (Sittig and Singh 2012). Third, the explainability of LLM-generated outputs remains an open challenge. While the human-in-the-loop design mitigates some immediate operational and clinical risks, full transparency into the basis of individual model outputs is not guaranteed by the current architecture. Explainability, traceability, provenance of retrieved information, and reproducibility of agent actions should therefore be treated as explicit requirements in future evaluation (Tonekaboni et al. 2019).

Three directions for future development are identified. First, the integration of multi-agent coordination – where specialized sub-agents handle distinct aspects of the care pathway through a shared orchestration layer – could increase task specialization across individual workflow stages. Second, the application of retrieval-augmented generation (RAG) over clinical guidelines and payer policy documents could allow the agent to ground its responses in current clinical and reimbursement information without requiring frequent model retraining. Such a system would require version control, source provenance, document-governance procedures, and mechanisms to ensure that outdated policies are not retrieved as current guidance. Third, fine-tuning the base models on de-identified sleep apnea clinical notes could be investigated as a method for improving task-specific performance, provided that de-identification, data governance, training provenance, and resulting model-privacy risks are appropriately evaluated.

CONCLUSION

This paper presents a reference technical architecture for deploying AI agents within a U.S. digital health platform specializing in sleep apnea care. The structured clinical data generated by modern HSAT workflows offers a potentially powerful foundation for intelligent automation. The proposed system demonstrates a technical pathway for using such data without requiring the AI agent to access raw physiological signals or routinely transmit patient information to external model API providers.

The architecture rests on three core pillars: a self-hosted open-weight language model served within a private AWS cloud environment; a modular NestJS tool-calling orchestration layer; and a tightly controlled data-access model restricted to structured clinical summaries. This design can support organizational control over PHI processing and facilitate implementation of minimum-necessary access principles, while enabling the agent to surface actionable information for clinical staff review. HIPAA compliance, however, remains dependent on the complete set of administrative, physical, technical, contractual, and organizational safeguards surrounding the deployment.

Llama 3.1 8B Instruct is identified as the preferred initial candidate for production evaluation, offering a potentially favorable balance of ecosystem maturity, deployment flexibility, licensing characteristics, and computational efficiency for U.S. healthcare environments. Under the illustrative spot-instance pricing assumptions presented in this study, the estimated GPU and model-storage infrastructure cost remains near \$295 per month. This estimate should not be interpreted as the total cost of ownership of the AI system. The mandatory human-in-the-loop review principle is designed to augment, rather than autonomously replace, clinical judgment and provides an important risk-control mechanism; regulatory classification and compliance obligations must nevertheless be determined for each specific implementation and intended use.

Future work should prioritize empirical validation through live production deployment, task-specific benchmarking of candidate models, quantitative assessment of latency and throughput, exploration of multi-agent coordination for more granular clinical pathway coverage, and the application of retrieval-augmented generation to improve response grounding. The architecture presented in this study offers a replicable technical foundation for further evaluation by healthcare organizations seeking to integrate AI-agent capabilities into existing digital health infrastructure while maintaining strong privacy, governance, and human-oversight controls.

REFERENCES

- BENJAFIELD, A. V., AYAS, N. T., EASTWOOD, P. R., HEINZER, R., LIM, D. C., PARTHASARATHY, S., QUAN, S. F., TAPIA, I. E., WOEHRLE, H. and MALHOTRA, A., 2019. Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis. *The Lancet Respiratory Medicine*, vol. 7, issue 8, pp. 687–698. DOI: [https://doi.org/10.1016/S2213-2600\(19\)30198-5](https://doi.org/10.1016/S2213-2600(19)30198-5)
- BLOBEL, B., 2018. Architectural approach to eHealth for enabling paradigm changes in health. *Studies in Health Technology and Informatics*, vol. 248, pp. 10–24. DOI: <https://doi.org/10.3233/978-1-61499-951-5-10>
- CENTERS FOR MEDICARE & MEDICAID SERVICES (CMS), 2024. *Remote Physiologic Monitoring: CPT Codes 99453, 99454, 99457*. [viewed 10 April 2026]. Available from: <https://www.cms.gov/medicare/payment/fee-schedules/physician>
- CENTERS FOR MEDICARE & MEDICAID SERVICES (CMS). *Local Coverage Determination (LCD): Positive Airway Pressure (PAP) Devices for the Treatment of Obstructive Sleep Apnea (L33718)*. [viewed 28 June 2026]. Available from: <https://www.cms.gov/medicare-coverage-database/view/lcd.aspx?LCDId=33718>
- CORRAL, J., MASA, J. F., DURAN-CANTOLLA, J., TERAN, J., CABELLO, M., HERNANDEZ-VOTH, A. and MONTSERRAT, J. M., 2017. Alternatives to polysomnography for the diagnosis of sleep apnea syndromes. *Archivos de Bronconeumología*, vol. 53, issue 8, pp. 432–438. DOI: <https://doi.org/10.1016/j.arbres.2017.02.005>
- DEEPSEEK-AI, 2025. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv preprint, arXiv:2501.12948. [viewed 10 April 2026]. Available from: <https://arxiv.org/abs/2501.12948>
- U.S. FOOD AND DRUG ADMINISTRATION (FDA), 2019. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning-Based Software as a Medical Device*. [viewed 10 April 2026]. Available from: <https://www.fda.gov/media/122535/download>

- GOLDSTEIN, M. M., PEWEN, W. F. and BLUMENTHAL, D., 2023. State health data privacy laws and the limits of HIPAA preemption. *Health Affairs*, vol. 42, issue 1, pp. 93–101. DOI: <https://doi.org/10.1377/hlthaff.2022.00734>
- HEVNER, A. R., MARCH, S. T., PARK, J. and RAM, S., 2004. Design science in information systems research. *MIS Quarterly*, vol. 28, issue 1, pp. 75–105. DOI: <https://doi.org/10.2307/25148625>
- KAPUR, V. K., AUCKLEY, D. H., CHOWDHURI, S., KUHLMANN, D. C., MEHRA, R., RAMAR, K. and HARROD, C. G., 2017. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, vol. 13, issue 3, pp. 479–504. DOI: <https://doi.org/10.5664/jcsm.6506>
- KIRSCH, D. B., 2018. To sleep, perchance to dream of a sleep medicine workforce. *Journal of Clinical Sleep Medicine*, vol. 14, issue 4, pp. 693–695. DOI: <https://doi.org/10.5664/jcsm.7070>
- LÉVY, P., KOHLER, M., McNICHOLAS, W. T., BARBÉ, F., McEVOY, R. D., SOMERS, V. K., LAVIE, L. and PÉPIN, J. L., 2015. Obstructive sleep apnoea syndrome. *Nature Reviews Disease Primers*, vol. 1, Article 15015. DOI: <https://doi.org/10.1038/nrdp.2015.15>
- META, 2024. *Llama 3.1: Model Information and Community License Agreement*. [viewed 28 June 2026]. Available from: <https://github.com/meta-llama/llama-models>
- MISTRAL AI, 2024. *Mistral 7B (v0.3)*. [viewed 28 June 2026]. Available from: <https://docs.mistral.ai/models/mistral-7b-0-3>
- MOOR, M., BANERJEE, O., ABAD, Z. S. H., KRUMHOLZ, H. M., LESKOVEC, J., TOPOL, E. J. and RAJPURKAR, P., 2023. Foundation models for generalist medical artificial intelligence. *Nature*, vol. 616, pp. 259–265. DOI: <https://doi.org/10.1038/s41586-023-05881-4>
- PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M. A. and CHATTERJEE, S., 2007. A design science research methodology for information systems research. *Journal of Management Information Systems*, vol. 24, issue 3, pp. 45–77. DOI: <https://doi.org/10.2753/MIS0742-1222240302>
- SITTIG, D. F. and SINGH, H., 2012. Electronic health records and national patient-safety goals. *New England Journal of Medicine*, vol. 367, issue 19, pp. 1854–1860. DOI: <https://doi.org/10.1056/NEJMs1205420>
- TONEKABONI, S., JOSHI, S., McCRAIDEN, M. D. and GOLDENBERG, A., 2019. *What Clinicians Want from Explainable Artificial Intelligence for Diagnostic Decision Support*. arXiv preprint, arXiv:1905.05134. [viewed 10 April 2026]. Available from: <https://arxiv.org/abs/1905.05134>
- WANG, L., MA, C., FENG, X., ZHANG, Z., YANG, H., ZHANG, J., CHEN, Z., TANG, J., CHEN, X., LIN, Y. and ZHAO, W., 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, vol. 18, issue 6, Article 186345. DOI: <https://doi.org/10.1007/s11704-024-40231-1>
- WATSON, N. F., 2016. Health care savings: The economic value of diagnostic and therapeutic care for obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, vol. 12, issue 8, pp. 1075–1077. DOI: <https://doi.org/10.5664/jcsm.5594>

ВНЕДРЯВАНЕ НА АГЕНТИ НА ИИ (AI) ЗА УПРАВЛЕНИЕ НА СЪННА АПНЕЯ В АМЕРИКАНСКА ЦИФРОВА ЗДРАВНА ПЛАТФОРМА: АРХИТЕКТУРА, ИЗБОР НА ГЕМ (LLM) И БИЗНЕС СТОЙНОСТ

Резюме: Сънната апнея остава значим и често недиагностициран проблем на съвременното здравеопазване. Епидемиологичните оценки показват, че обструктивната сънна апнея (OSA) засяга близо 936 милиона възрастни в световен мащаб (Benjafield et al. 2019). Навлизането на устройствата за домашно изследване за сънна апнея (HSAT) разширява възможностите за извънболнична диагностика и генерира структурирани клинични данни, които могат да подпомогнат интелигентната автоматизация на работните процеси. В настоящата статия се представя референтна техническа архитектура за внедряване на AI агенти в дигитална здравна платформа, специализирана в грижата за пациенти със сънна апнея. Предложената система се основава на локално хостван голям езиков модел, Llama 3.1 8B, обслужван чрез vLLM върху AWS GPU инфраструктура и интегриран с Node.js/NestJS бекенд посредством стандартизиран REST API. За подпомагане на внедряване, съобразено с изискванията на HIPAA, агентът е ограничен до структурирани пациентски данни – резултати от скринингови въпросници, АНІ стойности и показатели за придържане към CPAP терапията – и няма достъп до сурови физиологични сигнали. Разгледани са инфраструктурната конфигурация, критериите за избор на LLM и потенциалната бизнес стойност. При посочените

допускания за AWS ценообразуване основният разход за GPU и съхранение на модела е оценен приблизително на \$295 месечно; тази стойност е илюстративна и не представлява общата цена на притежание. Предложената архитектура изисква специфична клинична, техническа, регулаторна и сигурностна валидация преди внедряване в реална среда.

Ключови думи: AI агенти; сънна апнея; внедряване на LLM; цифрово здраве; HIPAA съответствие

Красимир Николов, студент
Университет по библиотекознание и информационни технологии
София, България
E-mail: krnikolov@gmail.com