

성별 특이적 특징 기반 프롬프트 설계를 통한 LLM 기반 수면무호흡증 중증도 분류

류승연¹, 김현진², 이영한³
¹ 성신여자대학교 융합보안공학과 석사과정
² 서울아산병원 신경과 교수
³ 성신여자대학교 융합보안공학과 교수

syryu2008@gmail.com, hkimneurology@gmail.com, yhlee@sungshin.ac.kr

Adapting LLMs for OSA Severity Classification using Sex-Specific Feature-Guided Prompting

Seungyeon Ryu¹, Hyeon Jin Kim², Younghan Lee¹

¹ Dept. of Convergence Security Engineering, Sungshin Women's University

² Department of Neurology, Asan Medical Center, Seoul

Abstract

Obstructive Sleep Apnea (OSA) is a common but often underdiagnosed sleep disorder, partly due to the failure to consider differing dominant symptoms between sexes. While polysomnography (PSG) remains the diagnostic gold standard, it is resource-intensive and impractical for large-scale screening. To improve accessibility, clinicians have developed self-reported sleep questionnaires. This study investigates whether Large Language Models (LLMs) can classify OSA severity using only demographic and questionnaire features, without model training or access to PSG-derived metrics. Using data from 1,098 patients in the APPLES dataset, we evaluate two prompting strategies: a basic zero-shot role-based prompt (P1) and an enhanced three-shot prompt incorporating chain-of-thought (CoT) reasoning and sex-specific feature importance (P2). The P1 baseline underperforms compared to traditional machine learning models. However, P2 substantially improves LLM performance, achieving accuracy comparable to a Random Forest classifier. These findings highlight the critical role of prompt design in unlocking the diagnostic potential of LLMs and suggest that well-engineered prompts, even with only a few labeled examples, can provide a low-cost, interpretable, and sex-aware alternative for OSA severity classification.

1. Introduction

Obstructive Sleep Apnea (OSA) is a sleep-related breathing disorder characterized by repeated episodes of airway obstruction during sleep. If left untreated, it can lead to serious health complications including cardiovascular disease, cognitive impairment, and daytime fatigue.

The current gold standard for diagnosing OSA is polysomnography (PSG), a comprehensive overnight sleep study that monitors various physiological signals. However, PSG is resource-intensive, requiring specialized equipment, clinical staff, and overnight hospitalization, making it inaccessible for large-scale screening. In response, sleep questionnaires such as the Epworth Sleepiness Scale (ESS) and the Stanford Sleepiness Scale (SSS) have been invented as low-cost, self-reported tools to help assess OSA risk based on patient-reported symptoms and demographic factors. [1]

Recent efforts have attempted to leverage machine learning (ML) to automate OSA classification using these non-invasive features. While these ML models can offer reasonable accuracy, they require structured datasets, labeled training data, and manual preprocessing, which can limit their scalability and adaptability, especially in

medical domains where labeled data is often limited.

To address these challenges, we turn to a different paradigm: Large Language Models (LLMs). LLMs, such as GPT-based models, have demonstrated capabilities in reasoning and classification tasks, particularly when guided through in-context learning using natural language prompts. In this study, we examine whether LLMs can classify OSA severity into three levels—Normal, Moderate, and Severe—based only on questionnaire and demographic inputs, with no access to PSG data. Given that symptom presentation and risk factors for OSA can differ significantly between sexes [2], we further investigated whether prompt strategies incorporating sex-specific clinical insights can improve model performance.

We implement two different prompt strategies: the first strategy (P1) uses a zero-shot role-based prompt, providing the model with a brief patient description and a general instruction to classify OSA severity. The second strategy (P2) employs a three-shot prompt augmented with chain-of-thought (CoT) reasoning and sex-specific feature weighting, where each example highlights the most influential clinical features identified through Random Forest analysis. We compare GPT-based model that differs in its level of prompt engineering to a Random Forest classifier.

2. Background

2.1 Obstructive Sleep Apnea (OSA)

Obstructive Sleep Apnea (OSA) is a common sleep disorder marked by recurring episodes of airflow reduction or cessation due to the collapse of the upper airway during sleep. These interruptions often result in fragmented sleep, intermittent hypoxia, and frequent arousals, which can contribute to a range of long-term health issues, including cardiovascular dysfunction, metabolic disorders, impaired cognition, and daytime fatigue. [3]

The definitive method for diagnosing OSA is polysomnography (PSG), a comprehensive sleep study that records multiple physiological signals such as respiratory effort, blood oxygen levels, brain activity, and heart rate throughout the night. Although highly accurate, PSG requires clinical infrastructure, trained personnel, and overnight patient monitoring, which significantly limits its availability and scalability. [4]

To improve accessibility, clinicians have developed questionnaire-based tools that offer a more practical means of assessing OSA risk. Instruments like the Epworth Sleepiness Scale (ESS) and the Stanford Sleepiness Scale (SSS) enable patients to self-report sleep-related symptoms. When combined with demographic characteristics—such as age, sex, BMI, and neck circumference—these tools support low-cost, non-invasive screening. However, their predictive power remains limited when used in isolation, especially in the absence of physiological data. [5]

2.2 Large Language Models (LLMs)

Large Language Models (LLMs), such as OpenAI's GPT series, are neural networks trained on massive corpora of text to learn language patterns, contextual relationships, and semantic reasoning. These models can perform a wide range of natural language tasks, including question answering, summarization, translation, and classification, often matching or exceeding task-specific models in performance.

Unlike traditional machine learning models, LLMs do not require task-specific retraining or labeled datasets to be effective. These models have already been pretrained on vast amounts of diverse textual data, allowing them to acquire a broad base of knowledge. Instead of training them on new data, we simply prompt the model, using carefully crafted input instructions, to extract relevant patterns or reasoning pathways needed for a given task. This makes LLMs particularly attractive in domains like healthcare, where obtaining large, labeled datasets is often difficult due to privacy concerns, annotation costs, and variability in clinical practice. [6]

2.3 In-Context Learning

In-context learning refers to an LLM's ability to perform tasks based on examples and instructions provided within the input prompt itself, without modifying the model's internal weights. Through few-shot learning, the model is given a handful of input-output examples and then asked to make predictions on new inputs that follow the same format. This technique allows LLMs to adapt to new tasks on the fly, simply by adjusting the prompt content and structure.

This study leverages in-context learning to test whether LLMs can classify OSA severity using demographic and questionnaire data, in a setting where no model training or fine-tuning is performed. We further investigate whether carefully designed prompts, incorporating clinical insights such as sex-specific feature importance, can guide the model toward more accurate and interpretable predictions. [7]

3. Methodology

3.1 Dataset

We utilized the Apnea Positive Pressure Long-term Efficacy Study (APPLES) dataset for this study. The dataset includes comprehensive clinical and demographic information related to Obstructive Sleep Apnea (OSA), featuring variables such as age, sex,

height, weight, body mass index (BMI), and neck circumference, along with sleep questionnaire scores from the Epworth Sleepiness Scale (ESS) and the Stanford Sleepiness Scale (SSS). The dataset comprises 1098 patients, spanning a range of three OSA severity levels from normal, mild to moderate, and severe. Participants were recruited from multiple sleep centers across the United States, and all data were collected under ethical guidelines with informed consent.

To investigate the impact of sex-specific symptom patterns, we divided the dataset into male ($N = 515$) and female ($N = 286$) subsets. We then trained separate Random Forest classifiers on each group to identify the most important features contributing to OSA severity classification for each sex. The top-ranked features from these sex-specific RF models were later used to guide the prompt design in LLM-based experiments.

3.2 Preprocessing

To align with the input requirements of Large Language Models (LLMs), the structured tabular data was transformed into natural language prompts. For each patient, demographic and questionnaire features were converted into a short descriptive sentence in plain English. This preprocessing step was essential to enable effective in-context learning with LLMs, which are optimized for language-based inputs rather than structured numerical data.

3.3 Prompt Design

To assess the impact of prompt design on LLM performance, we compared two different prompting strategies, P1 and P2. P1 employed a zero-shot role-based approach, where the model was given only a natural language patient description along with role instruction. P2 adopted a three-shot role-based strategy, further enhanced by feature importance guidance and chain-of-thought (CoT) reasoning, where the model was provided with labeled examples, highlighted key features, and encouraged to generate intermediate clinical reasoning before predicting severity. Table 1 summarizes the core differences between the two prompting strategies.

Prompt template	# of Shots	Role-Based	Chain-of-Thought	Feature Importance
P1	0	Yes	No	No
P2	3	Yes	Yes	Yes

Table 1. Key differences between the two prompting strategies used in GPT-based OSA severity classification.

The detailed prompt templates for P1 and P2 are presented in Figure 1 and Figure 2, respectively. In both P1 and P2, the System Prompt refers to the role-based instruction given to the LLM (e.g., *"You are a specialized sleep disorder expert..."*), while the Human Prompt provides the patient's clinical data—either as a single natural-language description (P1) or as a structured summary of weighted features with step-by-step reasoning (P2).

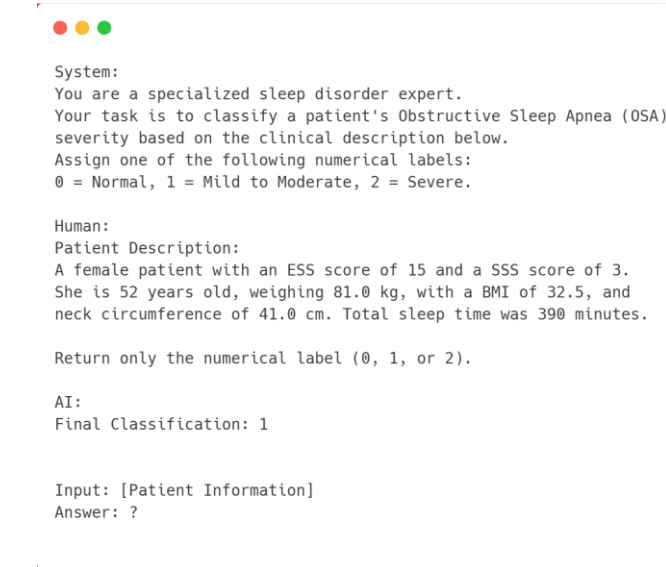
In the role-based setting, the model was explicitly instructed to act as a specialized sleep disorder expert, grounding its responses in a clinical diagnostic context.

In the zero-shot setting [7], the model received only a brief system instruction followed by a single patient description written in plain language. The prompt asked the model to assign a severity score (0 = Normal, 1 = Mild to Moderate, 2 = Severe) based on the questionnaire and demographic features, without providing any prior examples or intermediate reasoning steps.

In the three-shot setting [8], the model was provided with a structured prompt containing one labeled example for each OSA severity class (Normal, Moderate, Severe). Figure 2 shows the example corresponding to the 'Moderate' class. Each example presented the patient's clinical features alongside their corresponding feature importance scores, which were derived from sex-specific Random Forest analyses.

In the feature importance guided setting, prompts were constructed by selecting only the top five features with the highest

importance scores determined by sex-specific Random Forest models for each patient, as presented in Table 1.



System:
You are a specialized sleep disorder expert.
Your task is to classify a patient's Obstructive Sleep Apnea (OSA) severity based on the clinical description below.
Assign one of the following numerical labels:
0 = Normal, 1 = Mild to Moderate, 2 = Severe.

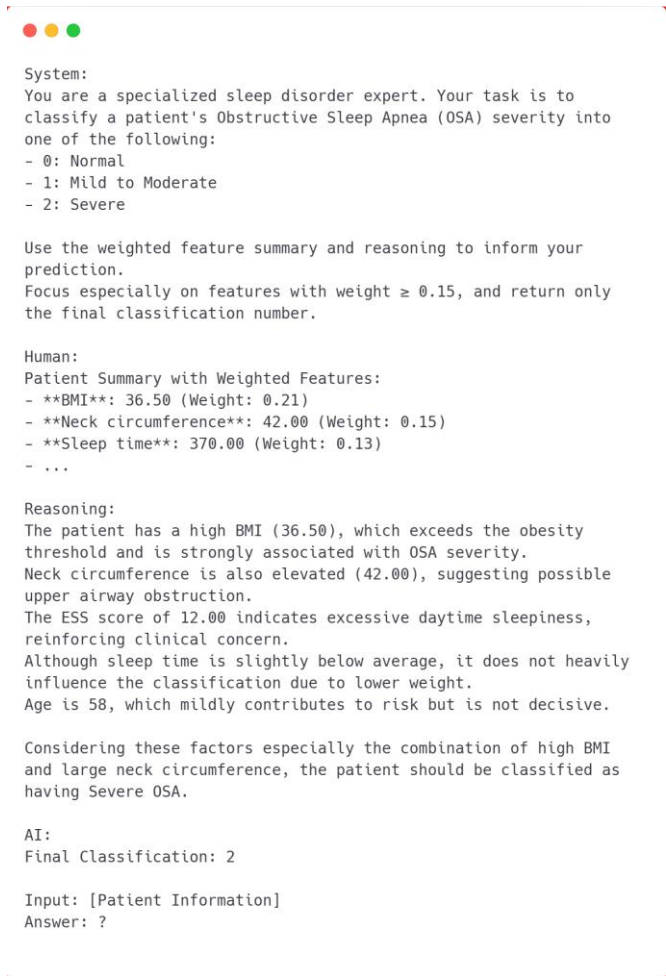
Human:
Patient Description:
A female patient with an ESS score of 15 and a SSS score of 3.
She is 52 years old, weighing 81.0 kg, with a BMI of 32.5, and neck circumference of 41.0 cm. Total sleep time was 390 minutes.

Return only the numerical label (0, 1, or 2).

AI:
Final Classification: 1

Input: [Patient Information]
Answer: ?

Figure 1. Example of the zero-shot role-based GPT prompt (P1) used in the baseline condition.



System:
You are a specialized sleep disorder expert. Your task is to classify a patient's Obstructive Sleep Apnea (OSA) severity into one of the following:
- 0: Normal
- 1: Mild to Moderate
- 2: Severe

Use the weighted feature summary and reasoning to inform your prediction.
Focus especially on features with weight ≥ 0.15 , and return only the final classification number.

Human:
Patient Summary with Weighted Features:
- **BMI**: 36.50 (Weight: 0.21)
- **Neck circumference**: 42.00 (Weight: 0.15)
- **Sleep time**: 370.00 (Weight: 0.13)
- ...

Reasoning:
The patient has a high BMI (36.50), which exceeds the obesity threshold and is strongly associated with OSA severity.
Neck circumference is also elevated (42.00), suggesting possible upper airway obstruction.
The ESS score of 12.00 indicates excessive daytime sleepiness, reinforcing clinical concern.
Although sleep time is slightly below average, it does not heavily influence the classification due to lower weight.
Age is 58, which mildly contributes to risk but is not decisive.

Considering these factors especially the combination of high BMI and large neck circumference, the patient should be classified as having Severe OSA.

AI:
Final Classification: 2

Input: [Patient Information]
Answer: ?

Figure 2. Example of the three-shot prompt with chain-of-thought reasoning and feature weighting (P2) used in the fine-tuned condition.

	Feature	Importance
All	BMI	0.17
	Weight	0.16
	Age	0.14
	Neck circumference	0.13
	Sleep time	0.13
Male	BMI	0.21
	Weight	0.15
	Sleep time	0.13
	Neck circumference	0.12
	Age	0.12
Female	BMI	0.17
	Age	0.15
	Weight	0.15
	Neck circumference	0.14
	Sleep time	0.13

Table 2. Top five most important features for OSA severity classification, as derived from the feature importance of RF models.

Each selected feature was explicitly displayed along with its corresponding value and importance score, arranged in descending order of importance. The system prompt instructed the model to prioritize features with an importance score of ≥ 0.15 to guide its diagnostic reasoning.

In the CoT (Chain-of-Thought) setting, the model was further guided by structured reasoning prompts. For each input case, a brief chain-of-thought explanation was automatically generated based on the top features and their clinical interpretations. The CoT reasoning included guidelines, such as assigning a higher risk when both BMI and neck circumference exceeded a clinically meaningful threshold. This stepwise reasoning was intended to simulate a clinician's thought process and help the model arrive at a more consistent and interpretable severity classification.

Both prompting strategies used the same GPT-4o-mini model and were evaluated on identical data splits to ensure a fair comparison. Model outputs were generated without access to ground-truth labels during inference.

4. Experimental Results

We evaluated the impact of prompt engineering on large language model (LLM) performance for OSA severity classification. Three methods were compared: a Random Forest (RF) classifier, a GPT model prompted with zero-shot role-based strategy (P1), and the same GPT model prompted with three-shot role-based strategy incorporating feature importance guidance and chain-of-thought reasoning (P2).

Table 3 summarizes the performance of these three methods—Random Forest, GPT with P1, and GPT with P2—across the entire test set as well as within sex-specific subgroups.

	Method	Accuracy	Precision	Recall	F1 Score
All	RF	0.63	0.61	0.63	0.62
	P1	0.47	0.60	0.47	0.48
	P2	0.58	0.66	0.58	0.61
Male	RF	0.55	0.54	0.55	0.54
	P1	0.37	0.51	0.37	0.37
	P2	0.53	0.58	0.53	0.55
Female	RF	0.61	0.59	0.61	0.60
	P1	0.50	0.59	0.50	0.52
	P2	0.63	0.69	0.63	0.60

Table 3. Comparison of Random Forest and GPT performance with different prompting strategies (P1, P2) on OSA severity classification.

The Random Forest classifier served as a strong baseline, achieving an overall F1 score of 0.62. Performance remained relatively consistent across male and female subgroups.

In contrast, the GPT model prompted with P1 underperformed

across all metrics. Its overall F1 score was 0.48, with especially low performance for male patients ($F1 = 0.37$), indicating difficulty in handling the sex-specific classification task without guided reasoning or prompt examples.

Notably, the GPT model prompted with P2 substantially outperformed its P1 counterpart and closely matched the Random Forest in overall performance. It achieved an F1 score of 0.61 on the full dataset, demonstrating the effectiveness of prompt engineering. The most substantial improvement was observed in the male subgroup, where the F1 score rose from 0.37 (P1) to 0.55 (P2), aligning closely with the Random Forest's male F1 score (0.54). The female group maintained strong and consistent performance across all methods.

These results collectively suggest that prompt engineering can enable LLMs to perform competitively with traditional ML models, particularly when prompts incorporate domain-specific structure and sex-specific information.

5. Discussion

We employed the lightweight GPT-4o-mini model to enable cost-efficient inference. However, evaluating larger models such as GPT-4-turbo could reveal whether increased model capacity leads to better reasoning performance, especially in complex or borderline cases. These models may also exhibit improved robustness and generalization in settings with noisy or atypical patient data.

Additionally, feature importance was derived from Random Forest classifiers trained on the structured questionnaire and demographic data. While effective, these weights are inherently limited by the assumptions and structure of tree-based models. Future work could integrate feature importance derived from domain experts, such as physicians' clinical assessments, to build more clinically aligned and interpretable prompts. Aligning LLM reasoning with established medical knowledge may improve trust and generalizability in real-world screening settings.

6. Conclusion

Our findings highlight the potential of LLMs not merely as static classifiers, but as adaptive systems whose behaviour can be shaped through prompt engineering. More importantly, it shows that prompt tuning using only a few labeled examples can close the performance gap with traditional ML models, offering a lightweight and interpretable approach to sex-sensitive, PSG-free OSA diagnostics.

Acknowledgements

This work is supported by the Ministry of Trade, Industry and Energy (MOTIE) under Training Industrial Security Specialist for High-Tech Industry (RS-2024-00415520) supervised by the Korea Institute for Advancement of Technology (KIAT), and the Ministry of Science and ICT (MSIT) under the ICAN (ICT Challenge and Advanced Network of HRD) program (No. IITP-2022-RS-2022-00156310) supervised by the Institute of Information & Communication Technology Planning & Evaluation (IITP).

This study was supported by grants from the NRF funded by MSIT (RS-2024-00359247 to HJ Kim).

References

- [1] P. J. Strollo Jr and R. M. Rogers, "Obstructive sleep apnea," *New England Journal of Medicine*, vol. 334, no. 2, pp. 99–104, 1996.
- [2] M. R. Bonsignore, A. Saareanta, and R. Riha, "Gender and sleep apnea," *European Respiratory Review*, vol. 28, no. 154, pp.

190030, 2019.

- [3] S. C. Veasey and I. M. Rosen, "Obstructive sleep apnea in adults," *New England Journal of Medicine*, vol. 380, no. 15, pp. 1442–1449, 2019.
- [4] Vishesh K Kapur, Dennis H Auckley, Susmita Chowdhuri, David C Kuhlmann, Reena Mehra, Kannan Ramar, and Christopher G Harrod. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an american academy of sleep medicine clinical practice guideline. *Journal of clinical sleep medicine*, 13(3):479–504, 2017.
- [5] Yewen Shi, Yitong Zhang, Zine Cao, Lina Ma, Yuqi Yuan, Xiaoxin Niu, Yonglong Su, Yushan Xie, Xi Chen, Liang Xing, et al. Application and interpretation of machine learning models in predicting the risk of severe obstructive sleep apnea in adults. *BMC Medical Informatics and Decision Making*, 23(1):230, 2023.
- [6] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, E. Chi, A. R. Mihaila, L. Collison, A. Sabharwal, et al., "Large language models encode clinical knowledge," *arXiv preprint arXiv:2302.10201*, 2023.
- [7] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- [8] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zeroshot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017.