NORTHERN JOURNAL of HEALTH SCIENCES

Research Article

Comparison of the Performance of Large Language Models (LLMs) in Predicting International Classification of Diseases Codes (ICD-10) Using Turkish Neurology Doctor Reports*

Murat Koçak[1], Seda Kibaroğlu[2], Hüseyin Ademoğulları[3], Mehmet Çağlar Akpınar[3], Meryem Koruk[3], Mehmet İbrahim Öksüz[3], Yasmin Ayşe Öztoklu[3] , Sevin Suyla Turhan[3]

[1]Baskent University, Faculty of Medicine, Department of Medical Informatics, Ankara, Türkiye.
[2]Baskent University, Faculty of Medicine, Department of Neurology, Ankara, Türkiye.
[3]Baskent University, Faculty of Medicine, Ankara, Türkiye.

Cite this article as: Koçak, M., Kibaroğlu, S., Ademoğulları, H., Akpınar, M.Ç., Koruk, M., Öksüz, M.İ., Öztoklu, Y.A., Turhan, S.S. (2025). Comparison of the performance of Large Language Models (LLMs) in predicting international classification of diseases codes (ICD-10) using Turkish neurology doctor reports. *Northern Journal of Health Sciences,* 1(3),190-201.

Corresponding Author:
Murat Koçak
E-mail:
muratkocak25@gmail.com

ORCID IDs of the authors:
M.K. 0000-0001-6510-3666
S.K. 0000-0002-3964-268X

Received: June 14, 2025
Revision Requested: July 28, 2025
Accepted: August 19, 2025
Publication Date: September 26, 2025

*This study was presented as an oral presentation at the International Congress on Digitalization and Artificial Intelligence in Health, 21-23 May 2025, Sinop/Türkiye.

Abstract

Objective: Accurate and efficient use of International Classification of Diseases Clinical Modification Codes (ICD-10) in neurology is vital for healthcare reimbursement, research, and patient health surveillance. However, manually extracting these codes from physician reports is both time-consuming and prone to errors. This study evaluates the performance of several large language models (LLMs) in automatically predicting ICD-10 diagnosis codes specifically from Turkish neurology physician reports.

Method: The study evaluates the performance of ten LLMs (ChatGPT, Cohere Coral, Claude, DeepSeek, Qwen, Groq, Gemini, Meta Llama, Mistral, and Perplexity) on a dataset of 51 de-identified neurology doctor reports. A standardized prompt was used to instruct each LLM to extract ICD-10 codes relevant to the diagnoses documented in the reports. The LLM-generated codes were then compared to a gold standard set of codes assigned by certified neurology coding specialists. Performance metrics such as accuracy, precision, recall and F1-score, were used to assess the models' effectiveness.

Results: Among the LLMs, ChatGPT emerged as the top performer with an accuracy of 68.6% and an F1-score of 0.812, demonstrating strong precision (0.686) and perfect recall (1.0). It excelled in identifying common neurological conditions such as migraines (G45.9), transient ischemic attacks (TIA), and motor neuron disorders. Gemini followed closely with 58.8% accuracy (F1-score: 0.750), while Qwen and Claude showed moderate performance (54.9% and 49.0% accuracy, respectively). Conversely, Groq and Meta AI exhibited significant limitations, with accuracies of 25.5% and 27.5%, respectively.

Conclusion: While LLMs show promise for automating ICD-10 coding from neurology reports, there is considerable variability in their performance. High-performing models like ChatGPT demonstrate strong potential, but further refinement is needed to improve the accuracy and reliability of lower-performing systems. Future research should focus on enhancing training datasets, incorporating rule-based algorithms, and integrating human oversight to address discrepancies, particularly in complex or rare neurological cases.

Keywords: Artificial Intelligence, Neurology Reports, Large Language Models (LLM), Natural Language Processing (NLP), ICD-10 Coding, Medical Informatics

INTRODUCTION

The International Classification of Diseases (ICD), published by the World Health Organization (WHO), serves as a globally recognized medical classification system that categorizes diseases, disorders, injuries, and other health-related conditions (WHO, 2015). Since its inception in 1893, the ICD has evolved into an indispensable tool for healthcare management, insurance processes, medical research, and public health surveillance (Barrit et al, 2025). In neurology, accurate assignment of ICD-10 codes is particularly critical for reimbursement, clinical research, and patient health monitoring. However, the current manual process of assigning these codes is labor-intensive, time-consuming, and prone to human error (Puts et al., 2025).

In most healthcare settings, certified professional coders dedicate significant time to reviewing extensive medical records on a case-by-case basis to assign ICD-10 codes manually (Soroush et al., 2024). While this approach ensures high accuracy when performed by experienced coders, it is inherently inefficient for large-scale applications. Additionally, in outpatient scenarios, physicians often perform the coding themselves, which may further increase the likelihood of inaccuracies due to limited familiarity with coding guidelines (Kocaman, 2024).

To address these challenges, there has been growing interest in leveraging Natural Language Processing (NLP) and Large Language Models (LLMs) to automate the extraction of ICD-10 codes from clinical documentation (Lee & Lindsey, 2024). Automated medical coding holds the potential to significantly improve the efficiency and reliability of healthcare administration processes. However, achieving this goal is not without its complexities.

Previous studies have shown that LLMs face significant challenges when tasked with generating accurate ICD codes from clinical notes or descriptions (Stanfill et al., 2010). For instance, Dai et al. (2024) demonstrated that even state-of-the-art LLMs struggle with understanding the nuances of medical terminology and adhering to strict coding guidelines (Dai et al., 2024). Similarly, Dong (2022) highlighted that while LLMs can extract relevant information from clinical notes, their performance degrades when dealing with ambiguous or complex cases (Dong et al., 2022).

Furthermore, the heterogeneity and ambiguity of medical language, particularly within specialized fields like neurology, present additional hurdles for automated coding systems (Reshma et al., 2025). Neurological diagnoses frequently involve intricate descriptions of symptoms, signs, and anatomical locations, demanding a high level of linguistic and medical domain expertise for accurate code assignment (Kalani & Anjankar, 2024). Puts et al. (2024) emphasized that LLMs often fail to capture the subtleties of clinical context, leading to misclassification of conditions such as migraines, transient ischemic attacks (TIAs), and motor neuron disorders.

Recent evaluations of LLM performance in medical coding have revealed varying degrees of success. For example, Schumacher et al. (2025) found that models like ChatGPT 4 and Gemini exhibit strong performance in identifying common neurological conditions but falter when faced with rare or less frequent diagnoses (Schumacher et al., 2025). Conversely, lower-performing models, such as Groq and Meta AI, consistently underperform in both common and rare cases, highlighting the need for more robust training methodologies (Albassam et al., 2025).

Building upon these findings, this study aims to benchmark the performance of seven leading LLMs in automatically predicting ICD-10 codes for neurology diagnosis from de-identified neurology physician reports (Simmons et al., 2024). By comparing their outputs against a gold standard set of codes assigned by certified neurology coding specialists, we aim to provide a comprehensive assessment of the current capabilities and limitations of LLMs in this specialized coding task. Our findings will inform future research on refining LLM-based coding systems and contribute to the ongoing development of more accurate and efficient automated medical coding solutions in neurology.

METHOD

This study utilized a dataset consisting of 51 neurology physician patient reports. Each report included detailed descriptions of patient symptoms, physical examination findings, laboratory results, and diagnoses. These reports covered a wide range of neurological conditions, making them suitable for evaluating the accuracy of ICD-10 code predictions.

The gold standard for comparison was established by certified neurology coding specialists who manually assigned ICD-10 codes to each report. This set of expert-assigned codes served as the benchmark for assessing the performance of the LLMs.

Ten LLMs were evaluated in this study:

- ChatGPT 4o
- Claude 3.7 Sonnet
- Cohere Coral
- Groq
- DeepSeek R1
- Gemini 1.5 Pro
- Qwen3
- Mistral
- Meta Llama 3 70B
- Perplexity

Each model was instructed using a standardized prompt designed to extract ICD-10 codes relevant to the diagnoses documented in the neurology reports. The prompt was structured as follows:

"From the following neurology physician report, extract the ICD code(s) corresponding to the documented diagnosis: [You can review the supplementary file for detailed neurology physician reports.]"

To ensure reproducibility, all LLMs were run with deterministic parameters (temperature = 0, top-p = 1). A single standardized prompt was deliberately used to maintain comparability across models; however, future work will systematically explore multiple prompt variations to assess prompt sensitivity.

The performance of the LLMs was assessed using the following metrics:

- ***Accuracy:*** The proportion of correctly predicted codes out of all predictions.
- ***Precision:*** The ratio of true positives (correctly predicted codes) to the total number of positive predictions (true positives + false positives).
- ***Recall:*** The ratio of true positives to the total number of actual positives. In this study, recall was consistently 1.0 because FN (false negatives) was zero—each case had only one predefined correct code.
- ***F1-Score:*** The harmonic mean of precision and recall, providing a balanced measure of a model's performance.

RESULTS

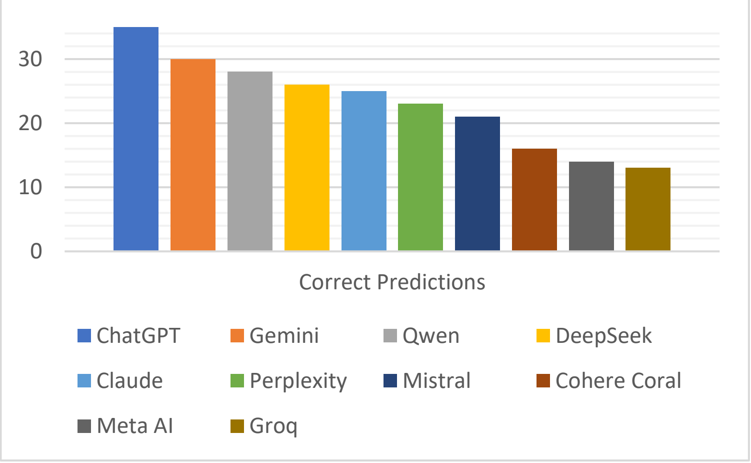
The evaluation of large language models (LLMs) in predicting International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes from Turkish neurology reports reveals significant insights into their performance and limitations. This section presents a detailed analysis of the accuracy, precision, recall, and F1-scores achieved by the evaluated systems, highlighting their ability to handle both common and complex neurological diagnoses. By comparing the LLM-generated codes with the gold standard set by certified neurology coding specialists, we aim to identify the strengths and weaknesses of each model, providing a foundation for understanding their potential applications and areas requiring improvement in automated medical coding. The goal is to evaluate the performance of different AI LLMs systems in predicting ICD-10 codes for 51 patient cases.

Overview

Table 1 presents a comparison of various AI systems based on the number of correct predictions each has made. ChatGPT leads the pack with the highest number of correct predictions (35), showcasing its superiority in delivering accurate results compared to other systems. Following closely are Gemini and Qwen , with 30 and 28 correct predictions, respectively, indicating their strong performance as well. On the other hand, systems like Groq and Meta AI lag significantly behind, with only 13 and 14 correct predictions, respectively, highlighting their lower effectiveness in this context. The remaining AI systems (DeepSeek , Claude , Perplexity , Mistral , and Cohere Coral) fall somewhere in between, demonstrating moderate performance levels. Overall, this data clearly illustrates the varying degrees of accuracy among the different AI systems, with ChatGPT emerging as the most reliable model based on the number of correct predictions.

Table 1. The correct prediction counts for each LLMs

LLMs	Correct Predictions
ChatGPT	35
Gemini	30
Qwen	28
DeepSeek	26
Claude	25
Perplexity	23
Mistral	21
Cohere Coral	16
Meta AI	14
Groq	13



Performance Metrics

Accuracy

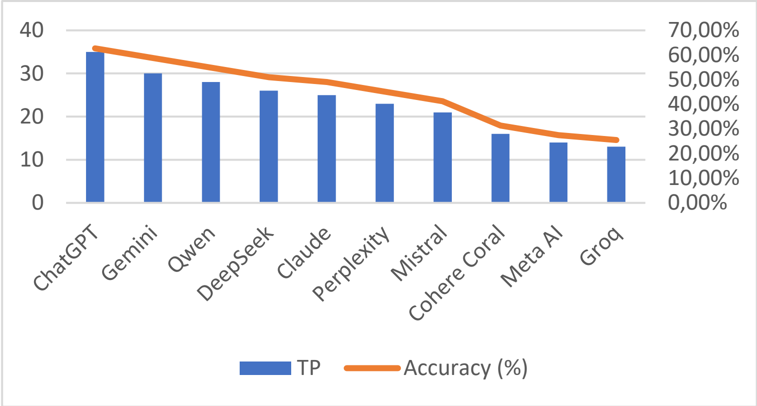
Table 2 compares the performance of various AI systems based on two key metrics: true positives (TP) and accuracy (%). ChatGPT leads the pack with the highest number of true positives (35) and the best accuracy (68.6%), making it the top-performing system in this evaluation. Following closely are Gemini and Qwen , which also demonstrate strong performance with accuracies of 58.80% and 54.90%, respectively. On the other hand, systems like Cohere Coral , Meta AI , and Groq show significantly lower accuracies, ranging from 25.50% to 31.40%, indicating weaker overall performance. While all systems correctly identified a certain number of true positives, their accuracy percentages reveal a clear divide between high-performing models such as ChatGPT and underperforming ones like Groq. This highlights the importance of selecting an AI system that not only identifies true positives effectively but also maintains a high overall accuracy.

Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Cases}}$$

Table 2. LLMs' performance: true positives and accuracy rates

LLMs	TP	Accuracy (%)
ChatGPT	35	68.6%
Gemini	30	58.80%
Qwen	28	54.90%
DeepSeek	26	51.00%
Claude	25	49.00%
Perplexity	23	45.10%
Mistral	21	41.20%
Cohere Coral	16	31.40%
Meta AI	14	27.50%
Groq	13	25.50%



Precision

Table 3 provides a performance analysis of various AI systems based on three key metrics: true positives (TP), false positives (FP), and precision. Precision is calculated as the ratio of true positives to the total number of predicted positives (TP + FP), indicating how reliable the system is when it makes a positive prediction. ChatGPT leads with the highest precision score of 0.686, achieved by correctly identifying 35 true positives while generating only 16 false positives, demonstrating its effectiveness in minimizing errors. Following ChatGPT, Gemini and Qwen also show relatively strong performance with precision scores of 0.588 and 0.55, respectively. On the other hand, systems like Groq , Meta AI , and Cohere Coral struggle with significantly lower precision scores (ranging from 0.255 to 0.314), which is largely due to their higher false positive rates compared to true positives. This suggests that these models are more prone to incorrect positive predictions, making them less reliable in scenarios where high precision is critical. Overall, the data highlights a clear performance gap between top-tier systems such as ChatGPT and lower-performing ones like Groq and Meta AI.

Precision measures the proportion of true positives among all positive predictions:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Table 3. Evaluation of LLMs models based on precision, true positives, and false positives

LLMs	TP	FP	Precision
ChatGPT	35	16	0.686
Gemini	30	21	0.588
Qwen	28	23	0.55
DeepSeek	26	25	0.511
Claude	25	26	0.49
Perplexity	23	28	0.451
Mistral	21	30	0.412
Cohere Coral	16	35	0.314
Meta AI	14	37	0.275
Groq	13	38	0.255

Recall

Recall measures the proportion of true positives among all actual positives. Since FN = 0 for all systems:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Each patient has only one correct code and the system is supposed to predict that one code, then FN is zero because the system either predicted it (TP) or another code (FP). So in this setup, FN is always zero, leading to recall of 1. That's a possible explanation.

Recall = 1.0 indicates that all correct ICD codes were identified by at least one system . Since each patient report in the dataset was associated with only one predefined correct ICD-10 code, the models either predicted the correct code (true positive) or produced a different code (false positive). Consequently, false negatives (FN) were always zero, leading to a recall value of 1.0 for all systems. This methodological feature explains why recall remained constant across all models.

F1-Score

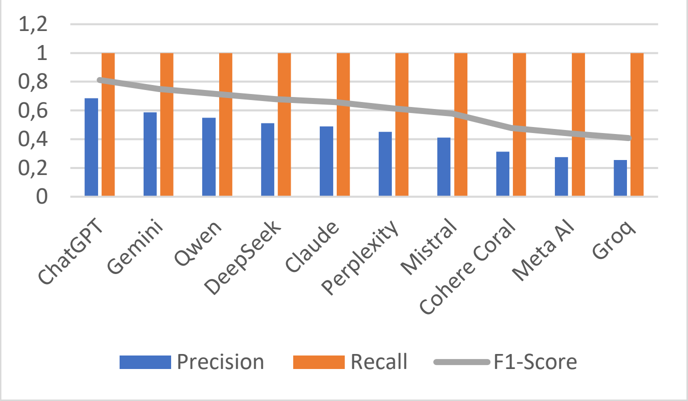
F1-Score combines Precision and Recall into a single metric:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 4 presents a performance comparison of various AI systems based on three key metrics: precision, recall, and F1-score. Notably, all systems achieve a perfect recall score of 1, indicating that they successfully identify all positive cases without missing any. However, their precision scores vary significantly, which directly impacts their F1-scores, a metric that balances precision and recall. ChatGPT leads the pack with the highest precision (0.686) and F1-score (0.812), making it the most effective system in this evaluation. Following ChatGPT, Gemini and Qwen also demonstrate strong performance with F1-scores of 0.75 and 0.714, respectively. On the other hand, systems like Cohere Coral , Meta AI , and Groq struggle with lower precision scores (ranging from 0.255 to 0.314), resulting in significantly lower F1-scores (between 0.408 and 0.478). This highlights a clear divide between high-performing models such as ChatGPT and underperforming ones like Groq, emphasizing the importance of selecting an appropriate AI system based on its precision and overall effectiveness.

Table 4. Comparison of LLMs systems based on precision, recall, and F1-score

LLMs	Precision	Recall	F1-Score
ChatGPT	0.686	1	0.812
Gemini	0.588	1	0.75
Qwen	0.55	1	0.714
DeepSeek	0.511	1	0.678
Claude	0.49	1	0.658
Perplexity	0.451	1	0.615
Mistral	0.412	1	0.578
Cohere Coral	0.314	1	0.478
Meta AI	0.275	1	0.441
Groq	0.255	1	0.408



Summary Tables

Table 5 provides a comprehensive comparison of various AI systems based on critical performance metrics such as correct predictions, accuracy, precision, recall, and F1-score. Among these systems, ChatGPT stands out as the top performer, achieving the highest accuracy (68.60%) and F1-score (0.812). Its strong balance between precision (0.686) and recall (1) makes it particularly effective in delivering accurate and reliable results. Following ChatGPT, Qwen and Gemini also demonstrate notable performance, with accuracies of 54.90% and 58.80%, respectively. Both systems maintain high recall values (1), ensuring that they do not miss any positive cases, while their F1-scores (0.714 for Qwen and 0.75 for Gemini) indicate a reasonable trade-off between precision and recall. On the other hand, systems like DeepSeek , Claude , and Perplexity show moderate performance, with accuracies ranging from 45.10% to 51.00%. These models still achieve perfect recall but struggle with lower precision, which affects their overall effectiveness.

Table 5. Performance comparison of LLMs

LLMs	Correct Predictions	Accuracy (%)	Precision	Recall	F1-Score
ChatGPT	35	68.60%	0.686	1	0.812
Gemini	30	58.80%	0.588	1	0.75
Qwen	28	54.90%	0.55	1	0.714
DeepSeek	26	51.00%	0.511	1	0.678
Claude	25	49.00%	0.49	1	0.658
Perplexity	23	45.10%	0.451	1	0.615
Mistral	21	41.20%	0.412	1	0.578
Cohere Coral	16	31.40%	0.314	1	0.478
Meta AI	14	27.50%	0.275	1	0.441
Groq	13	25.50%	0.255	1	0.408

In contrast, several systems exhibit significantly weaker performance. For instance, Groq , Cohere Coral , and Meta AI have accuracies below 32%, making them less reliable in practical applications. Groq, with an accuracy of just 25.50% and an F1-score of 0.408, ranks as one of the least effective systems in this evaluation. Similarly, Cohere Coral and Meta AI, with accuracies of 31.40% and 27.50%, respectively, also struggle to deliver consistent results. Despite having perfect recall, these systems suffer from low precision, indicating a higher likelihood of false positives. This imbalance between precision and recall leads to lower F1-scores, which is a critical metric for evaluating the overall effectiveness of a model. Overall, the data underscores the significant performance gap between leading systems like ChatGPT and underperforming ones like Groq, highlighting the importance of selecting the right model for specific tasks (Table 5).

Table 6 divides the AI systems into three distinct categories. 1.Top Performers, 2. Mid-Range Performers, and 3. Low Performers based on their accuracy and F1-score. ChatGPT emerges as the leading system with the highest accuracy (68.6%) and F1-score (0.812), demonstrating its superior balance between precision and recall. Following closely are Gemini and Qwen , which occupy the second and third positions, respectively, with accuracies of 58.8% and 54.9%, and F1-scores of 0.750 and 0.714. These top performers excel in delivering reliable and accurate results.

In the Mid-Range Performers category, systems like Claude, Perplexity, and Mistral show moderate performance, with accuracies ranging from 41.2% to 49.0% and F1-scores between 0.578 and 0.658. While these systems are not as strong as the top performers, they still provide reasonable results for less demanding tasks.

Finally, the Low Performers category includes Meta AI , Groq , and Cohere Coral , which exhibit the weakest performance among the evaluated systems. Meta AI and Groq, in particular, have the lowest accuracy and F1-score values (27.5% and 25.5% accuracy, respectively), indicating significant room for improvement in their precision and recall capabilities. Cohere Coral performs slightly better than these two but still lags behind the mid-range and top performers (Table 6).

Table 6. Evaluation of LLMs models by performance categories

Category	LLMs	Accuracy (%)	F1-Score
Top Performers	ChatGPT	68.6	0.812
	Gemini	58.8	0.75
	Qwen	54.9	0.714
Mid-Range Performers	Claude	49	0.658
	Perplexity	45.1	0.615
	Mistral	41.2	0.578
Low Performers	Meta AI	27.5	0.441
	Groq	25.5	0.408
	Cohere Coral	31.4	0.478

Distribution of Correct ICD Codes

Table 7 represents a classification of various health conditions and diseases using an international system. These codes are used to describe patients' health status, diseases, and symptoms.

From a class-level perspective, common neurological conditions such as G43.9 (migraine, unspecified), G20 (Parkinson’s disease), and G40.9 (epilepsy) were consistently predicted across high-performing models. In contrast, rare or less frequently observed diagnoses, including G12.20 (amyotrophic lateral sclerosis, unspecified) and G04.8 (encephalitis), showed lower prediction consistency. This highlights that while LLMs are effective for frequent conditions, their accuracy declines for less common diagnoses.

It becomes apparent that a significant portion of the codes relate to neurological conditions and symptoms. For instance, codes such as G20 (Parkinson's disease), G35 (multiple sclerosis), G40.9 (epilepsy), and G44.2 (tension-type headache) fall under this category. Additionally, codes like I63.5 and I63.9 describe cerebral infarction (brain hemorrhage) and transient cerebral ischemia (brain vessel occlusion), respectively.

Table 7 also includes codes related to symptoms. For example, R26.0 (ataxic gait), R41.0 (disturbances of skin sensation), R41.3 (other and unspecified disturbances of consciousness), R42 (dizziness and giddiness), and R47.0 (aphasia) describe various symptoms. Furthermore, the list contains codes for infectious diseases, such as B02.9 (herpes simplex infection). Analyzing ICD codes is crucial for planning healthcare services, managing diseases, and effectively managing symptoms.

Table 7 provides an overview of the distribution of ICD codes across various categories. The most common category is neurological diseases and symptoms. Other categories, such as circulatory system, eye and adnexa, digestive system, and more, also contain codes. This table serves as a useful tool for understanding the classification and meaning of ICD codes. This information is essential for planning healthcare services, managing diseases, and effectively managing symptoms. Even high-performance models yielding inconsistent results in rare diagnoses (e.g., G12.20 - Amyotrophic lateral sclerosis) underscores the critical need for expert oversight in such complex cases. For example, a mix of transient ischemic attack (G45.9) and migraine (G43.9) carries the potential to cause significant deviations in patient management and treatment planning due to miscoding.

Table 7. Distribution of correct ICD code categories and descriptions

1. **Certain infectious and parasitic diseases**: None in your list.	
2. **Neoplasms:** None in your list.	
3. **Blood and blood-forming organs:** None in your list.	
4. **Endocrine, nutritional, and metabolic diseases**: None in your list.	
5. **Mental and behavioural disorders:**	
	▪ F02.80: Mental disorder due to known physiological condition, unspecified, with other specified behavioral disturbance
	▪ F05: Delirium not induced by psychoactive substances
	▪ F41.0: Tension-type headache
6. **Nervous system:**	
	▪ B02.9: Herpes simplex without complication (infections)
	▪ G04.8: Other specified encephalitis
	▪ G12.2: Amyotrophic lateral sclerosis
	▪ G12.20: Amyotrophic lateral sclerosis, unspecified
	▪ G20: Parkinson's disease
	▪ G31.83: Other specified degenerative diseases of the nervous system
	▪ G35: Multiple sclerosis
	▪ G40.9: Epilepsy, unspecified
	▪ G43.0: Migraine without aura
	▪ G43.9: Migraine, unspecified
	▪ G44.2: Tension-type headache
	▪ G44.209: Other specified headache syndromes, not elsewhere classified
	▪ G44.4: Cluster headache and related syndromes
	▪ G44.81: Other headache syndromes
	▪ G44.89: Other specified headache disorders
	▪ G45.9: Transient cerebral ischemic attack, unspecified
	▪ G50.0: Trigeminal neuralgia
	▪ G51.0: Bell's palsy
	▪ G56.0: Mononeuropathy of upper limb
	▪ G61.0: Idiopathic neuropathy
7. **Eye and adnexa:**	
	▪ H46.9: Optic neuritis, unspecified
	▪ H49.1: Paresis of eyelid
	▪ H49.2: Blepharoptosis
	▪ H53.40: Unspecified visual field defect
8. **Circulatory system:**	
	▪ I61.9: Nontraumatic intracerebral hemorrhage, unspecified
	▪ I63.5: Cerebral infarction due to unspecified occlusion or stenosis of cerebral arteries

- I63.9: Cerebral infarction, unspecified

9. **Respiratory system:** None in your list.

10. **Digestive system**:
    - J34.89: Other specified diseases of nose and sinuses

11. **Symptoms, signs, and abnormal findings**:
    - R26.0: Ataxic gait
    - R41.0: Disturbances of skin sensation
    - R41.3: Other and unspecified disturbances of consciousness
    - R42: Dizziness and giddiness
    - R43.1: Disturbances of smell
    - R47.0: Aphasia

DISCUSSION

This study aimed to evaluate the performance of large language models (LLMs) in assigning accurate and consistent ICD-10 diagnostic codes based on clinical narratives derived from real patient cases presenting with diverse neurological symptoms. The dataset included 51 patient records encompassing detailed medical histories, physical examination findings, and neurologic assessments. Ten different LLMs were tested for their ability to extract relevant diagnostic information and map it correctly to standardized ICD-10 codes.

The results demonstrated varying levels of accuracy among the models. ChatGPT achieved the highest overall accuracy rate at 68.6%, followed by Gemini and Qwen, which also performed significantly well. These models showed a relatively strong capability in interpreting complex clinical narratives, identifying key symptom patterns, and aligning them with appropriate diagnostic categories. In contrast, Groq, Cohere Coral, and Meta AI exhibited lower accuracy, highlighting limitations in understanding nuanced medical terminology or contextual clues within Turkish clinical texts.

Notably, discrepancies between the correct ICD-10 codes and model-generated outputs were often due to overlapping symptomatology across multiple conditions, such as stroke, migraine, Parkinson’s disease, multiple sclerosis, and peripheral nerve disorders. For instance, some models misclassified transient ischemic attack (TIA)-like presentations as migraines or vice versa, indicating challenges in distinguishing between similar clinical syndromes without explicit temporal or imaging markers.

A key limitation of this study is the relatively small dataset of 51 reports, all drawn from Turkish neurology practice. This restricts the generalizability of the findings. Future studies should include larger and more diverse datasets from multiple institutions and specialties to validate the outcomes. Another limitation is the reliance on a single standardized prompt. While this ensured comparability among models, prompt engineering is known to significantly influence LLM performance. Therefore, future work will systematically test multiple prompt variations. Finally, although deterministic parameters (temperature = 0, top-p = 1) were applied, LLMs may still exhibit minor variability in outputs, and this inherent non-determinism should be considered when deploying such systems in clinical workflows.

Furthermore, incorrect coding can not only create statistical performance differences, but can also directly affect patient care. For example, confusing transient ischemic attack (G45.9) with migraine (G43.9) can lead to significant differences in patient diagnosis and treatment. Similarly, in rare but critically important conditions such as amyotrophic lateral sclerosis (G12.20), incorrect coding can lead to delayed early intervention and appropriate care. These findings demonstrate the vital importance of accurate coding in a clinical context as well as high accuracy rates.

Future research should focus on expanding the dataset with a broader range of clinical cases, incorporating multimodal inputs (e.g., MRI reports, lab results), and fine-tuning models on Turkish medical corpora to enhance linguistic and semantic comprehension. Additionally, integrating LLMs into clinical decision support systems could provide valuable assistance in streamlining diagnostic coding, reducing clerical burden, and improving documentation consistency provided they are used under physician supervision. Moreover, LLM-based automatic coding contributes to clinical processes not only in neurology but also in other disciplines such as cardiology, psychiatry, and internal medicine.

CONCLUSION

In conclusion, while LLMs like ChatGPT demonstrate considerable potential in supporting ICD-10 coding tasks in neurology, they cannot replace expert clinical judgment. Their role should be seen as complementary—offering efficient, scalable, and intelligent assistance rather than autonomous decision-making. With further refinement and validation, these models can become integral components of modern healthcare informatics, contributing to more accurate diagnoses, improved patient care, and enhanced health data management.

Limitations of the Study

Despite these promising results, several limitations must be acknowledged. First, the sample size was relatively small and limited to Turkish clinical narratives, which may affect the generalizability of the findings. Second, the presence of ambiguous or incomplete clinical descriptions posed interpretational challenges even for human experts, further complicating model evaluation. Third, the current versions of LLMs do not have access to real-time diagnostic reasoning tools such as laboratory data or radiological images, which are crucial for definitive diagnosis in neurology. Furthermore the dataset size (51 reports) is relatively small, which may limit the generalizability of the findings, the reports are in Turkish.

Additionally, although deterministic parameters were applied (temperature = 0, top-p = 1), LLMs may still produce slightly different outputs across repeated runs. The use of a single standardized prompt, while ensuring fairness across models, limited exploration of prompt sensitivity. Future research should therefore address both reproducibility and prompt variability in greater depth.

Declaration of Interests: The authors declare no competing interests

Financial Support: Not applicable

Ethical Considerations and Data Security

This study was conducted in accordance with ethical principles for medical research. To ensure patient privacy and data security, all physician reports were thoroughly de-identified prior to analysis. All personally identifiable information (PII) was removed. Access to the dataset was restricted to the research team through encrypted and secure environments. The study protocol was approved by the Baskent University Institutional Review Board (Project No: KA25/180).

Appendix 1. Spreadsheet of all questions, annotations, and LLMs responses

REFERENCES

Albassam, D., Cross, A., & Zhai, C. (2025). Leveraging LLMs for Predicting Unknown Diagnoses from Clinical Notes. *arXiv preprint arXiv:*2503.22092.

Barrit, S., Torcida, N., Mazeraud, A., Boulogne, S., Benoit, J., Carette, T., Carron, T., Delsaut, B., Diab, E., Kermorvant, H., Maarouf, A., Maldonado Slootjes, S., Redon, S., Robin, A., Hadidane, S., Harlay, V., Tota, V., Madec, T., Niset, A., ... Carron, R. (2025). Specialized Large Language Model Outperforms Neurologists at Complex Diagnosis in Blinded Case-Based Evaluation. *Brain Sciences,* 15(4), 347.

Dai H, Wang C, Chen C, Liou C, Lu A, Lai C, Shain B, Ke C, Wang W, Mir T, Simanjuntak M, Kao H, Tsai M, Tseng V. (2024). Evaluating a Natural Language Processing–Driven, AI-Assisted International Classification of Diseases, 10th Revision, Clinical Modification, Coding System for Diagnosis Related Groups in a Real Hospital Environment: Algorithm Development and Validation Study*. J Med Internet Res;*26:e58278,

Dong, H., Falis, M., Whiteley, W., Alex, B., Matterson, J., Ji, S., Chen, J., & Wu, H. (2022). Automated clinical coding: What, why, and where we are? *Npj Digital Medicine*, 5(1), 1-8.

Kalani, M., & Anjankar, A. (2024). Revolutionizing Neurology: The Role of Artificial Intelligence in Advancing Diagnosis and Treatment. *Cureus,* 16(6), e61706.

Kocaman, V. (2024, April 20). Comparing Spark NLP for healthcare and ChatGPT in extracting ICD10-CM codes from clinical notes . John Snow Labs. Retrieved [Insert Date of Retrieval] from https://www.johnsnowlabs.com/comparing-spark-nlp-for-healthcare-and-chatgpt-in-extracting-icd10-cm-codes-from-clinical-notes/

Lee, S. A., & Lindsey, T. (2024). Can Large Language Models abstract Medical Coded Language?. arXiv preprint arXiv:2403.10822.

Puts, S., Zegers, C. M. L., Dekker, A., & Bermejo, I. (2025). Developing an ICD-10 Coding Assistant: Pilot Study Using RoBERTa and GPT-4 for Term Extraction and Description-Based Code Selection. *JMIR Formative Research,* 9, e60095.

Reshma, O. K., Saleena, N., & Nazeer, K. A. (2025). Context-aware automated ICD coding: A semantic-driven approach. *Information Systems,* 132, 102539.

Schumacher, E., Naik, D., & Kannan, A. (2025). Rare Disease Differential Diagnosis with Large Language Models at Scale: From Abdominal Actinomycosis to Wilson's Disease. arXiv preprint arXiv:2502.15069.

Simmons, A., Takkavatakarn, K., McDougal, M., Dilcher, B., Pincavitch, J., Meadows, L., ... & Sakhuja, A. (2024). Benchmarking large language models for extraction of international classification of diseases codes from clinical documentation. medRxiv, 2024-04.

Soroush, A., Glicksberg, B. S., Zimlichman, E., Barash, Y., Freeman, R., Charney, A. W., ... & Klang, E. (2024). Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI,* 1(5),

Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A., & Hersh, W. R. (2010). A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association : JAMIA, 17(6), 646–651.

World Health Organization. (2015). The International Classification of Diseases, 10th Revision. https://icd.who.int/browse10/2015/en