EMGLLM: Data-to-Text Alignment for Electromyogram Diagnosis Generation with Medical Numerical Data Encoding

Zefei Long[1], **Zhenbiao Cao**[2], **Wei Chen**[2*], **Zhongyu Wei**[1,3,4*]
[1]School of Data Science, Fudan University, China
[2]School of Software Engineering, Huazhong University of Science and Technology, China
[3]Shanghai Innovation Institute, China
[4]Research Institute of Intelligent Complex Systems, Fudan University, China
zflong23@m.fudan.edu.cn, {m202477011, lemuria_chen}@hust.edu.cn, zywei@fudan.edu.cn

Abstract

Electromyography (EMG) tables are crucial for diagnosing muscle and nerve disorders, and advancing the automation of EMG diagnostics is significant for improving medical efficiency. EMG tables contain extensive continuous numerical data, which current Large Language Models (LLMs) often struggle to interpret effectively. To address this issue, we propose EMGLLM, a data-to-text model specifically designed for medical examination tables. EMGLLM employs the EMG Alignment Encoder to simulate the process that doctors compare test values with reference values, aligning the data into word embeddings that reflect health degree. Additionally, we construct ETM, a dataset comprising 17,250 real cases and their corresponding diagnostic results, to support medical data-to-text tasks. Experimental results on ETM demonstrate that EMGLLM outperforms various baseline models in understanding EMG tables and generating high-quality diagnoses, which represents an effective paradigm for automatic diagnosis generation from medical examination table. [1]

1 Introduction

Electromyography (EMG) refers to the pattern of electrophysiological signal concomitant with musculations recorded with an electromyograph (Ni et al., 2020), which plays a significant role in evaluating human activities (Cooray et al., 2022; Smedemark-Margulies et al., 2023; Rakhmatulin, 2024). In medicine, the EMG is one of the major diagnostic tools for identifying and characterizing motor unit disorders (Daube, 2002), which is commonly used to examine nerve and muscle excitability and conduction functions, thereby determining the functional status of peripheral nerves, neurons, neuromuscular junctions, and the muscles themselves. After the EMG examination, the physicians perform a two-step analysis based on the records of the electrical signals. They first analyze the waveforms, converting the complex electrical signals into easily interpretable data tables, which contain essential information for medical diagnosis, such as amplitude, conduction velocity, and latency. Subsequently, by completing quantitative analysis, the doctors interpret the converted table data to render their final diagnosis and form a diagnostic report (Boon et al., 2008). In this paper, we focus on the data-to-text task of automatic diagnosis generation from EMG tabular data.

Figure 1 shows an anonymized EMG diagnosis including two parts, *Findings* and *Impression*. In the context of an EMG examination, *Findings* refer to observations of tables, aiming to objectively describe the phenomena reflected by the data, thus facilitating further analysis by the physician. To accurately identify various neuromuscular disorders within tabular data and translate them into *Findings*, physicians must possess a deep understanding of the distinct patterns associated with neuromuscular junction disorders, radiculopathies, upper motor neuron lesions, and so on. In terms of *Impression*, it consists of two aspects: a summary and interpretation of the test results, as well as an analysis of the clinical significance of the *Findings*, which may include diagnostic suggestions or potential issues. Therefore, *Impression* requires a certain level of clinical experience from doctors. (Katirji, 2002) Basically, EMG diagnosis writing can be error-prone and tedious for under-experienced physicians, and onerous and time-consuming for experienced physicians. Therefore, considering the powerful reasoning and text generation capabilities of Large Language Models (LLMs) in the medical field (Fan et al., 2024), we are motivated to explore methods for using LLMs to process examination tables and automatically

* Corresponding Author.
[1] Further resources can be found at https://github.com/zefeilong/EMGLLM.

Findings of the Association for Computational Linguistics: ACL 2025, pages 20470–20480
July 27 - August 1, 2025

generate medical EMG diagnoses.

The automatic generation of EMG diagnoses involves two major challenges:

- **Reference value comparison:** This task requires analyzing from the relative size of EMG test values compared to their corresponding reference values to assess the degree of health or abnormality of the test items. Moreover, differences in equipment, environment, and other factors across hospitals may result in varying optimal reference values, which increases the complexity of analysis.

- **Intensive numerical data input:** It is necessary for EMG diagnosis to quantitatively understand of medical examination tables containing large amounts of continuous numerical data to generate diagnostic results. For LLMs, directly handling numerical data may present certain difficulties (Golkar et al., 2023) since LLMs are not well-adept at comparing numerical values and quantitatively diagnosing the normality of these values, which may lead to errors.

To address this, we proposes EMGLLM, a novel data-to-text framework for automatic diagnosis from medical examination tables, which introduces EMG Alignment Encoder specialized in encoding continuous numerical data in EMG examination tables. The EMG Alignment Encoder can compare the test values with reference values, encoding them into virtual tokens that represent the degree of abnormality, and aligning numerical data to diagnostic text. This allows the LLM to better understand EMG tables, thereby generating more accurate and comprehensive diagnoses. Our main contributions include:

- For automatic diagnosis generation from medical examination tables, we propose a data-to-text model, EMGLLM, which includes an EMG Alignment Encoder designed to encode continuous numerical values and enhance data understanding.

- We construct a dataset ETM comprising about 17,000 real EMG tables with their diagnoses annotated by authoritative physicians, which can provide support for researches on automatic diagnosis generation.

Compared to all baseline methods, EMG diagnoses generated by EMGLLM demonstrates higher quality in all evaluation metrics, fully proving the effectiveness and robustness of EMGLLM. This method can also be applied to other medical examination tables.

2 Related Work

2.1 Data-to-text Generation

Data-to-text is a significant branch of natural language processing (Sharma et al., 2024). Its goal is to transform complex numerical data and tables into textual descriptions, assisting users in understanding and analyzing data, thereby improving the efficiency of data analysis. Data has the characteristics of complex structure and information density, and many studies have proposed methods to address this challenge. For example, splicing nearest neighbors (Wiseman et al., 2021) is an effective data-to-text policy by inserting or replacing text segments directly from neighbor source-target pairs to construct generations. Search-and-learning method (Jolly et al., 2021) is aimed at enhancing semantic coverage in few-shot data-to-text generation. Recently, some research applied LLMs to complete data-to-text. MURMUR (Saha et al., 2022) and TAT-LLM (Zhu et al., 2024) respectively enhanced data-to-text generation capabilities through multi-step and discrete reasoning frameworks. TableLLaMA (Zhang et al., 2024a) and TableLLM (Zhang et al., 2024b) were implemented supervised fine-tuning on table datasets for proficiently handling tabular data.

In the medical fields, data-to-text generation also holds vast application prospects. For instance, language models can complete automatic drug description generation from medical information tables (Yermakov et al., 2021) and diagnosis from examination tables (Gu et al., 2020; Guo et al., 2024).

2.2 EMG Diagnosis

EMG has a wide range of applications in medical diagnosis (Gaso et al., 2021; Nguyen et al., 2023; Li et al., 2023). EMG signals can be used to construct an end-to-end sleep stage neural classification model for diagnosing sleep disorders (Chambon et al., 2017). They can also be classified by Markov model (Bureau et al., 2021) for diagnosing potential neuropathies. Specifically, a dataset MIME (Gu et al., 2020) for EMG table tasks is used to train models such as hierarchical transformers. The model and dataset are both closed source

EMG Diagnostic Report

Input: Examination Tables Age: 32 Gender: Male Height: 178 cm

Electromyogram (EMG)

被检肌肉 Examined Muscle	纤颤 Fibs	正锐 PSW	束颤 Fascics	其他 Others	MUP多相 MUP Polyph	MUP形态 MUP Form	募集相 Recurit
左 指总伸肌 L Ext.Dig.Com	-	1+	-	-	-	> 5mv MUP	干扰相 Interference
左 第一背侧骨间肌 L Dors.Int.I	-	-	-	-	-	-	混合相 Mix
...	...	...	...	...	...	...	...

Nerve Conduction Velocity (NCV)

被检神经 Examined Nerve	项目 Type	刺激 Stim	记录 Rec	潜伏期（左）Lat (L)	潜伏期（右）Lat (R)	波幅（左）Amp (L)	波幅（右）Amp (R)	速度（左）CV (L)	速度（左）CV (R)
正中神经 Median	运动 Motor	腕 Wrist	拇短展肌 APE	4.2		7.0		7.0	
尺神经 Ulnar	感觉 Sensory	中指 Dig III	腕 Wrist	3.4	3.8	15.0	14.2	46.2	44.8
腓总神经 Reroncal	F波 F-wave	踝 Ankle	趾短伸肌 EDB	50.5					
...	...	...	...	...	...	...	...	...	...

Output: Diagnosis

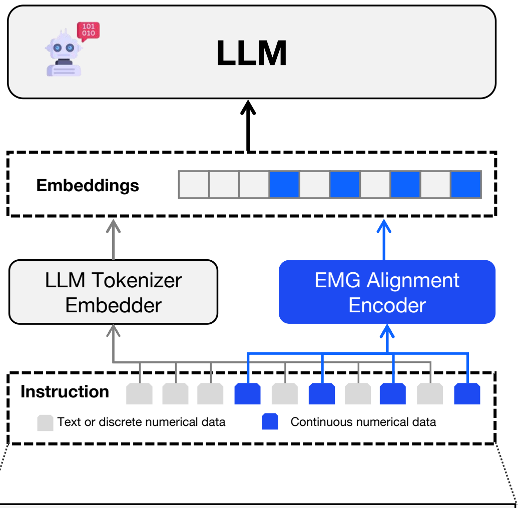
Findings:

- EMG: 被检肌未见明显肌源性或神经源性损害肌电改变。
  (No denervation or reinnervation in the muscles examined.)
- NCV: 左侧正中神经运动传导潜伏期正常上限，感觉神经传导速度轻度减慢；余运动和感觉神经传导速度和波幅正常范围。
  (Mildly prolonged motor distal latency and slightly slowed sensory nerve conduction velocity of left Median Nerve is revealed. The conduction velocity and amplitude of residual motor and sensory nerves are within the normal range.)

Impression:

左侧正中神经轻度损害，CTS可考虑。
(The left median nerve is mildly damaged, and diagosis of Carpal Tunnel Syndrome could be considered)

Figure 1: An EMG diagnostic report example, including EMG tables (EMG and NCV) and their corresponding diagnosis (*Findings* and *Impression*). For our automatic diagosis generation task, the input is the EMG tables and the output is the diagnosis.



Instruction

As a physician, your task is to provide a diagnosis for the patient based on the EMG and NCV examination tables.

EMG Test:

Examined Muscle	Fibs	PSW	Fascics	Others	MUP Polyph	MUP Form	Recurit
L Ext.Dig.Com	-	1+	-	-	-	-	Interference
L Dors.Int.I	-	-	-	-	-	-	Mix

...

NCV Test:

Examined Nerve	Type	Stim	Rec	Lat(L)	Kat(R)	Amp(L)	Amp(R)	CV(L)	CV(R)
Median	Motor	Elbow	APE	**6.2**	**6.1**	**5.8**	**6.4**	**60.5**	**59.2**
Median	Motor	Wrist	APE	**2.4**	**2.3**	**6.1**	**7.0**	NaN	NaN
Median	Sensory	Digit III	Wrist	**2.4**	**2.5**	**25**	**20**	**61.9**	**61.9**
Median	F-wave	Wrist	APE	**23.3**	**23**	NaN	NaN	NaN	NaN
Ulnar	Sensory	ADM	Wrist	**2.4**	**2.5**	**25**	**20**	**61.9**	**61.9**
Ulnar	Motor	Below Elbow	ADM	**5.4**	**5.3**	**6.2**	**6.7**	**58.8**	**59.1**
Ulnar	Motor	Above Elbow	ADM	**7.5**	**7.7**	**6.0**	**6.3**	**57.1**	**58.1**
Radial	Sensory	Forearm	Opisthenar	**2.3**	**1.9**	**22**	**23**	**71.0**	**86.8**

...

Please generate your diagnosis:

Figure 2: EMGLLM Framework. Medical examination tables contain a large amount of continuous data. The numbers marked in blue in the instruction could be encoded by the EMG Alignment Encoder.

and the method is relatively simple, which cannot fully adapt to complex data-to-text tasks. Therefore, this paper aims to explore automatic diagnostic generation based on EMG tables.

3 Method

3.1 Model

EMGLLM is composed of two fundamental components: the EMG Alignment Encoder and the LLM. The EMG Alignment Encoder is a specialized module tailored for understanding medical examination tables such as EMG tables. As illustrated in Figure 2, when an EMG table is input, text and discrete data are tokenized and vectorized by LLM's tokenizer and embedder directly. For continuous numerical data, they are encoded into virtual tokens using the EMG Alignment Encoder. The model's output is the *Findings* and *Impression* of the EMG tables.

The process by which the EMG Alignment Encoder analyzes continuous numerical table cells is analogous to the approach employed by doctors. In EMG examinations, reference values are the most critical criterion for determining whether a test parameter is within normal ranges. The reference range defines the upper and lower limits of normalcy for a specific examination item. The extent to which the test value exceeds the reference range reflects the degree of pathological alteration in the muscle or nerve. In practice, doctors first assess the relative magnitude of test values based on reference values, then make a annotation within the table cell to indicate the degree of abnormality. The EMG Alignment Encoder is designed to emulate this process by comparing the continuous test data with multiple reference ranges and encoding the abnormality degree semantically into virtual tokens that are more interpretable by the LLM.

Reference Value Acquisition In practice, doctors rely on their clinical experience to make appropriate adjustments to reference values for certain individual cases. This process involves strong

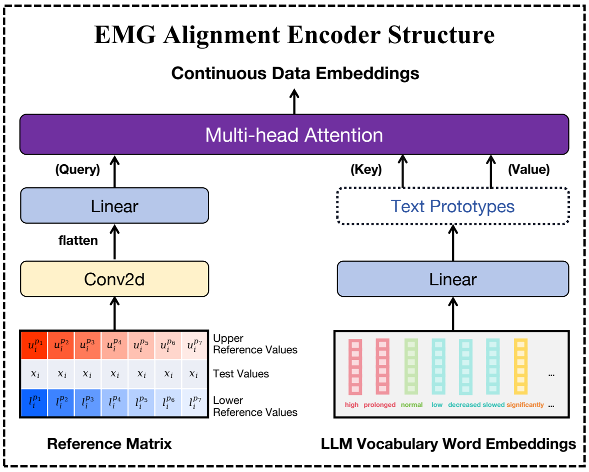


Figure 3: EMG Alignment Encoder Structure. EMG Alignment Encoder regards the reference matrix of the test values as an image, extracts features using the convolutional layer, and aligns continuous data with text through the Attention mechanism.

subjectivity. Therefore, we propose a method for mining reference values based on percentiles from the training dataset. We filter out the completely healthy cases without any abnormalities from the training dataset and statistically analyze the healthy case subset for each item. For a given examination item i, we use the k upper percentiles $u_i^{p_1}, u_i^{p_2}, \cdots, u_i^{p_k}$ and the k lower percentiles $l_i^{p_1}, l_i^{p_2}, \cdots, l_i^{p_k}$ as multiple reference values at different levels, where $p_1, p_2, \ldots, p_k$ represent different percentile thresholds. These percentiles allow us to estimate the boundaries of the reference ranges from data distribution of healthy individuals.

EMG Alignment Encoder Structure The input to the EMG Alignment Encoder for item i is a reference matrix X_i:

$$X_i = \begin{pmatrix} u_i^{p_1} & u_i^{p_2} & \cdots & u_i^{p_k} \\ x_i & x_i & \cdots & x_i \\ l_i^{p_1} & l_i^{p_2} & \cdots & l_i^{p_k} \end{pmatrix} \quad (1)$$

where x_i denotes the continuous test value. The EMG Alignment Encoder views the reference matrix X_i as a form of image, where the pixels represent the arrangement of the examined value and reference ranges, as illustrated in Figure 3. Using a convolutional layer $Conv$ with d_C output channels, the model sequentially compares the test value with the reference values. Subsequently, a linear layer integrates the output vectors of $Conv$ to produce the vector $\hat{X}_i$ representing the feature of the test value x_i:

$$\hat{X}_i = f_1(W_1 \text{flatten}(Conv(X_i)) + b_1) \quad (2)$$

where f_1 and b_1 are the activation function and bias. When $Conv$ outputs m vectors of dimension d_C, $W_1 \in \mathbf{R}^{m\times N}$, where N represents the number of vectors output by the linear layer. Consequently, $\hat{X}_i$ contains N vectors of dimension d_C.

These data features are then aligned with the word embeddings in the LLM's vocabulary. As shown in Figure 3, the alignment process first involves learning a set of text prototypes $E' \in \mathbf{R}^{V'\times D}$ from the LLM's vocabulary $E \in \mathbf{R}^{V\times D}$ through $E' = W_2 E$, where V and V' refers to the size of vocabulary and text prototypes respectively subject to $V' \ll V$, D denotes dimension of the LLM embeddings, and $W_2 \in \mathbf{R}^{V\times V'}$. Text prototypes E' serve as a compressed version of the vocabulary capable of semantically implying health or abnormality in medical diagnosis, such as "prolonged", "slowed", and "decreased". The EMG Alignment Encoder then connects the continuous data features $\hat{X}_i$ in Equation 2 to these text prototypes via a multi-head attention layer:

$$\begin{aligned} &EMGAlignmentEncoder(X_i) \\ &= MultiHeadAttention(Q_i, K, V, n_{head}) \quad (3) \end{aligned}$$

where $Q_i = \hat{X}_i W_Q$, $K = E'W_K$, $V = E'W_V$, n_{head} is the number of heads, $W_Q \in \mathbf{R}^{d_C\times d}$. $W_K, W_V \in \mathbf{R}^{D\times d}$, $d = \lfloor d_C/n_{head} \rfloor$. The output of EMG Alignment Encoder is N data embeddings of dimension D. In Equation 3, Query is computed from the continuous data in tables, while the Key and Value are derived from the LLM embeddings. The EMG Alignment Encoder leverages this Attention mechanism to associate continuous data with text.

The additional reference value information and reasonable continuous data encoding contribute to enhancing the performance of LLMs in data-to-text medical tasks. Another advantage of the EMG Alignment Encoder lies in its continuous function, where similar numeric values are encoded into correspondingly similar virtual tokens. In contrast, standard LLMs tokenize numeric values, a process that discretizes the table's data. For instance, two numerically close values, such as 9.99 and 10.0, may result in significantly different word embeddings in LLMs, which may be not reasonable in data-to-text scenario.

Finally, EMGLLM integrates the EMG Alignment Encoder with the LLM. With the assistance of the EMG Alignment Encoder, the LLM gains better understanding of the EMG table. Combined with the LLM's strong generative capabilities, this enhancement endows EMGLLM with better automated diagnostic abilities.

3.2 Training

EMG Alignment Encoder Pre-training Before training on EMG diagnosis task, we pre-train the EMG Alignment Encoder to ensure it can properly perform data understanding in a single continuous test value. The purpose of pre-training is to help the model understand the underlying medical semantics behind the relative size relationship between a test value with its reference values. Freezing the LLM, two types of pre-training tasks based on one test value are applied: (1) Classification of abnormality. (2) Making LLM generate a diagnostic description of a word. The loss function in pre-training is same as the supervised fine-tuning of LLM.

Figure 4 presents examples of the pre-training data. The instructions for pre-training tasks require EMG Alignment Encoder to provide reasonable virtual tokens so that the base LLM can clearly understand their meaning. In pre-training dataset construction, the test value x_i and the reference values u_i and l_i can be obtained by sampling from the diagnosis generation training dataset, and the output labels can be constructed directly from manually defined rules. For example, if a test value exceeds $u_i^{0.02}$ by 20%, the virtual tokens should convey the meaning of "significantly high". This rule-based approach does not rely on any authoritative reference values from hospitals, but can naturally learn an understanding of reference values from the data distribution of healthy individuals, which has good generality.

Model Fine-tuning Upon the completion of pre-training, we proceed with supervised training for the EMG data-to-text task. In the fine-tuning phase, we further train both the LLM and the EMG Alignment Encoder on EMG train dataset, where LLM is efficiently trained by the Low-Rank Adaptation (LoRA) (Hu et al., 2021) method.

Through the aforementioned steps, the EMG Alignment Encoder's representation of continuous numerical data can be enhanced, enabling the EMGLLM to better understand continuous data and diagnose from EMG tables.

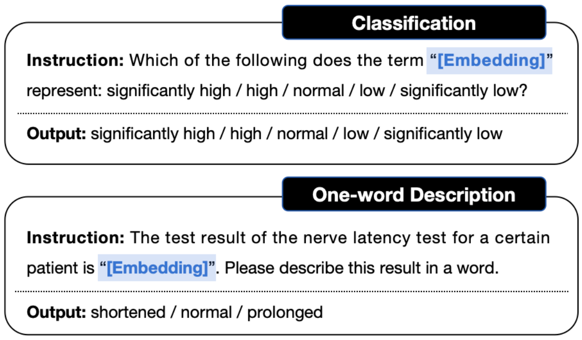


Figure 4: Examples of two pre-training tasks for EMG Alignment Encoder. **[Embedding]** represents the virtual tokens encoded by the EMG Alignment Encoder from a reference matrix X_i in Equation 1, which is required to enable pre-trained LLMs to complete single data diagnosis without fine-tuning EMG data.

4 Experiments

4.1 ETM Dataset

In this section, we introduce a high quality EMG diagnostic report dataset ETM (Electromyogram Table Mart) derived from Huashan Hospital Affiliated to Fudan University [2] with high authenticity, accuracy, and authority, which contains a total of 17,250 diagnostic reports from 2006 to 2013, and each data includes:

- Basic information of real anonymized patients (age, gender, and height).
- EMG tables (EMG and NCV tests) from the real EMG examination in the hospital.
- Diagnosis (*Findings* and *Impression*) personally written by experienced physicians.

The data format is shown in Figure 1. The full dataset is further proportionally divided into training, validation, and testing set, with data volumes of 13800, 1725, and 1725 respectively, which can effectively support medical data-to-text research.

Some statistical information of the ETM dataset is displayed in Table 1 basic statistics for our Some statistics information ETM dataset. The average number of continuous numerical data in tables is 33.14, indicating that the model's input contains dense numerical information. The automatic diagnostic task requires the model to have a sufficient understanding of continuous test values.

[2] https://www.huashan.org.cn/

Measurement	Value
# of Samples	17,250
Avg # of Continous Numerical Data	33.14
Avg Length (*Findings*)	85.04
Avg Length (*Impression*)	26.82

Table 1: Dataset Statistics

4.2 Setup

4.2.1 Baseline Methods

We select various baseline models capable of performing automatic EMG diagnosis, including both general text-to-text generation models and models specifically designed for data-to-text tasks. All models undergo supervised training on the ETM dataset, except for DeepSeek-R1, which is evaluated under the 0-shot and 3-shot settings.

Chinese-Alpaca-2-7B-16K Chinese-Alpaca-2-7B-16K (Cui et al., 2023) is a widely used LLM. It also serves as the base LLM module for EMGLLM. The prompt template for Chinese-Alpaca-2-7B-16K is fully consistent with that of EMGLLM, with the only difference being that Chinese-Alpaca-2-7B-16K directly process the continuous data in textual form. Besides, this model is similarly fine-tuned using the LoRA method, with training hyperparameters consistent with those of EMGLLM. The comparison with EMGLLM can clearly demonstrate the effect of the EMG Alignment Encoder on the automatic generation of diagnostic results.

TableLLM-7B TableLLM (Zhang et al., 2024b) is an LLM specifically designed for tabular data inputs, fine-tuned on a large dataset of table tasks. Since the base model used by TableLLM, CodeLlaMA-7B (Rozière et al., 2023), does not support Chinese, we replicate the training using the official code on Chinese-CodeLlaMA-7B to develop a Chinese version TableLLM , and subsequently fine-tune it on ETM dataset.

Lattice Lattice (Wang et al., 2022) is a data-to-text generation model with a structure-aware self-attention mechanism and a tranformation-invariant positional encoding mechanism improved from T5-base.

DeepSeek-R1 We conduct experiments on DeepSeek-R1 (DeepSeek-AI, 2025) with 671B parameters, one of the most powerful reasoning models currently available. Due to limited computational resources, we do not fine-tune the model; instead, we use data from the ETM training set as examples to prompt DeepSeek-R1 to complete the tasks. As a deep thinking model, it first generates a long chain of thought for reasoning, and then outputs the final answer containing the EMG diagnosis. We only evaluate the final answer part, and design the following three experimental settings without training to test DeepSeek-R1's performance:

- **0-shot:** No complete EMG table few-shot examples are provided. Only 3 random formatted EMG diagnostic outputs in training set are given to ensure the model follows the correct structure, including both *Findings* and *Impression* sections.
- **Random 3-shot**: 3 EMG tables and their corresponding diagnoses are randomly selected from the training set as few-shot examples.
- **RAG 3-shot:** The 3 most similar samples are retrieved from the training set as context. We use the frequency-based BM25 (Manning et al., 2008) retrieval method to compute similarity. The keywords for each retrieved sample are extracted from the first column of the EMG table and the first four columns of the NCV table. Each cell is considered a keyword, including the names of examined muscles, names of examined nerves, examination types, stimulation sites, and recording sites, excluding any numerical information.

4.2.2 Implementation Details

For the implementation of EMGLLM, we first obtain reference values from the ETM training set. From a total of 13,800 samples, we filter out 7,166 (52%) completely healthy samples based on text rules and perform quantile statistics on each examination item i to determine reference values u_i and l_i. Subsequently, 7 quantile thresholds $\{p_1, p_2, p_3, ..., p_7\} = \{0.02, 0.05, 0.08, ..., 0.2\}$ are set to construct the reference matrix $\hat{X}_i$.

In the EMG Alignment Encoder, the output channel number $d_C = 64$, the number of output embeddings $N = 2$, the size of text prototypes $V' = 192$, and the number of heads $n_{head} = 8$.

For the LLM component of EMGLLM, we select the widely-used Chinese-Alpaca-2-7B-16K as base model. Pre-training of EMG Alignment Encoder is conducted for 2000 steps, followed by 5 epochs of fine-tuning, with a batch size of 1 and a gradient accumulation step of 16. Optimization is performed using the Adam optimizer, with

Model		Automatic						Model		Human
		ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	Correctness	Completeness	Human
Overall	DeepSeek-R1 (0-shot)	55.83(0.31)	34.10(0.57)	45.06(0.42)	42.56(0.27)	33.49(0.40)	27.43(0.48)	3.33	3.25	3.68
	DeepSeek-R1 (random 3-shot)	58.61(0.42)	36.26(0.53)	49.61(0.41)	48.11(0.26)	37.61(0.28)	30.39(0.33)	3.31	3.37	3.70
	DeepSeek-R1 (RAG 3-shot)	63.20(0.27)	42.26(0.45)	55.02(0.49)	52.38(0.42)	42.65(0.42)	35.77(0.42)	3.45	3.60	4.00
	Lattice	71.59(0.94)	56.82(1.25)	65.25(0.96)	55.97(0.82)	46.57(0.85)	39.41(0.83)	3.46	3.28	3.50
	TableLLM-7B	74.41(0.42)	58.03(0.80)	67.37(0.55)	65.48(0.56)	58.31(0.67)	53.22(0.79)	3.72	3.70	3.88
	Chinese-Alpaca-2-7B-16K	79.24(0.33)	65.15(0.70)	73.18(0.45)	71.26(0.50)	65.18(0.50)	60.67(0.74)	4.02	3.92	4.21
	EMGLLM (Ours)	**80.44(0.23)**	**66.26(0.50)**	**74.24(0.26)**	**72.86(0.51)**	**66.70(0.58)**	**62.14(0.64)**	**4.11**	**4.09**	**4.38**
Findings	DeepSeek-R1 (0-shot)	54.83(0.41)	33.09(0.46)	44.58(0.36)	36.40(0.35)	28.94(0.37)	23.50(0.35)	3.45	3.42	3.88
	DeepSeek-R1 (random 3-shot)	59.28(0.62)	36.59(0.67)	50.37(0.64)	46.66(0.58)	36.72(0.57)	29.35(0.58)	3.38	3.48	3.88
	DeepSeek-R1 (RAG 3-shot)	64.93(0.42)	43.27(0.56)	56.19(0.50)	51.97(0.56)	42.62(0.59)	35.54(0.61)	3.50	3.74	4.12
	Lattice	71.83(0.73)	56.80(0.71)	65.67(0.71)	54.73(0.56)	46.33(0.56)	39.69(0.53)	3.63	3.41	3.56
	TableLLM-7B	74.19(0.58)	57.45(0.81)	66.93(0.67)	64.55(0.65)	57.35(0.74)	51.92(0.81)	3.85	3.86	3.90
	Chinese-Alpaca-2-7B-16K	79.02(0.66)	64.35(0.77)	72.43(0.66)	70.11(0.73)	63.91(0.78)	59.16(0.80)	4.03	4.00	4.36
	EMGLLM (Ours)	**80.36(0.52)**	**66.03(0.69)**	**73.92(0.53)**	**71.83(0.54)**	**65.75(0.61)**	**61.05(0.66)**	**4.10**	**4.13**	**4.40**
Impressions	DeepSeek-R1 (0-shot)	48.14(0.63)	25.75(0.69)	44.19(0.64)	33.25(0.55)	24.85(0.57)	19.31(0.58)	3.20	3.09	3.48
	DeepSeek-R1 (random 3-shot)	48.70(0.87)	25.35(1.13)	44.92(0.98)	35.27(1.25)	25.84(1.24)	19.40(1.22)	3.23	3.25	3.52
	DeepSeek-R1 (RAG 3-shot)	53.79(0.50)	31.78(0.72)	50.55(0.60)	40.13(0.71)	31.41(0.75)	25.32(0.80)	3.41	3.46	3.88
	Lattice	65.06(0.56)	46.51(0.74)	63.04(0.59)	50.77(0.59)	39.60(0.60)	30.19(0.63)	3.29	3.14	3.43
	TableLLM-7B	70.36(0.54)	51.53(0.79)	67.91(0.60)	62.69(0.66)	53.60(0.72)	46.87(0.81)	3.59	3.54	3.86
	Chinese-Alpaca-2-7B-16K	76.68(0.34)	61.04(0.55)	74.72(0.40)	70.41(0.50)	62.85(0.59)	57.14(0.62)	4.01	3.85	4.06
	EMGLLM (Ours)	**77.21(0.41)**	**61.49(0.86)**	**75.26(0.43)**	**70.91(0.36)**	**63.29(0.64)**	**57.38(0.86)**	**4.13**	**4.05**	**4.36**

Table 2: **Main Results.** Average results (standard deviation) of EMGLLM and baseline models on the ETM test set. All automatic evaluations are tested with 5 random seeds.

Model		Automatic Evaluation						Model Evaluation	
		ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	Correctness	Completeness
Overall	Chinese-Alpaca-2-7B-16K	77.48(0.71)	62.82(0.71)	70.86(0.62)	68.25(0.80)	61.97(0.78)	57.48(0.78)	3.88	3.85
	EMGLLM (Ours)	**79.48(0.65)**	**65.73(0.76)**	**73.51(0.70)**	**71.64(0.72)**	**65.63(0.77)**	**61.25(0.80)**	**3.93**	**3.92**
Findings	Chinese-Alpaca-2-7B-16K	77.63(0.65)	62.81(0.72)	70.48(0.59)	67.72(0,81)	61.54(0.81)	56.83(0.82)	3.88	3.91
	EMGLLM (Ours)	**79.30(0.48)**	**65.36(0.53)**	**72.92(0.46)**	**70.64(0.54)**	**64.71(0.55)**	**60.14(0.56)**	**3.90**	**3.93**
Impressions	Chinese-Alpaca-2-7B-16K	73.36(0.71)	55.17(0.88)	71.16(0.66)	65.82(0.74)	57.07(0.78)	50.92(0.87)	3.87	3.78
	EMGLLM (Ours)	**76.91(1.10)**	**60.56(1.85)**	**74.63(1.20)**	**70.02(1.31)**	**62.16(1.76)**	**56.38(2.10)**	**3.95**	**3.92**

Table 3: **Data-intensive Results.** Average results (standard deviation) of experiment on a subset with larger average amount of continuous values. EMGLLM demonstrates more significant advantages. All automatic evaluations are tested with 5 random seeds.

a learning rate of 5e-5. The LLM is trained using the LoRA method, with a rank of 8, an alpha value of 16, and the training target set to ['q_proj', 'v_proj'].

For the training of baseline models, we preprocess the dataset according to the input and output formats required by the model and employ the recommended hyperparameters of the projects.

4.2.3 Metrics

To comprehensively evaluate the quality of EMG automatic diagnosis, we use multidimensional metrics. The automatic metrics include:

• **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004):** It measures the generation quality by comparing the overlap between texts. ROUGE-1, ROUGE-2, and ROUGE-L are selected as metrics.

• **BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002):** It compares the n-gram match between texts. We use BLEU-1, BLEU-2, and BLEU-3 to evaluate the model's capabilities.

In addition, we introduce model evaluation, using GPT-4o (OpenAI, 2023) as a judge to assess the quality of the model-generated diagnoses. We provide authoritative doctors' ground truth diagnoses as a reference for GPT-4o, simultaneously inputting the model-generated results, allowing GPT-4o to analyze and compare the differences between the two and provide a multidimensional evaluation. GPT-4o's scoring criteria include:

• **Correctness:** evaluate whether a diagnosis falsely reports non-existent abnormalities. A higher score indicates fewer false positives.

• **Completeness:** evaluate whether a diagnosis has missed reporting existing abnormalities. A higher score indicates fewer missed abnormalities.

In model evaluation, GPT-4o evaluates the *Findings* and *Impressions* separately, and we use the average of these two evaluations as the overall score for the diagnosis. The evaluation template for GPT-4o is presented in Appendix A.

Finally, we conduct the human evaluation and provide human scores. We sample 50 examples from the test set and rate from 1 to 5 to the gen-

	Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3
Overall	w/o Reference Value	79.02(0.38)	64.32(0.53)	72.79(0.47)	70.81(0.58)	64.44(0.64)	59.85(0.71)
	w/o Encoder Pre-training	79.24(0.18)	64.98(0.38)	73.02(0.12)	71.44(0.36)	65.26(0.40)	60.77(0.45)
	EMGLLM (rule-based)	79.86(0.28)	65.50(0.62)	73.69(0.30)	72.38(0.43)	66.09(0.54)	61.45(0.62)
	w/o Encoder Fine-tuning	80.07(0.12)	65.92(0.33)	74.09(0.12)	72.66(0.27)	66.47(0.35)	61.91(0.40)
	EMGLLM	**80.44(0.23)**	**66.26(0.50)**	**74.24(0.26)**	**72.86(0.51)**	**66.70(0.58)**	**62.14(0.64)**
Findings	w/o Reference Value	79.03(0.36)	64.26(0.58)	72.36(0.36)	70.03(0.58)	63.81(0.66)	59.05(0.72)
	w/o Encoder Pre-training	79.00(0.45)	64.20(0.59)	72.27(0.46)	70.21(0.52)	63.96(0.57)	59.19(0.63)
	EMGLLM (rule-based)	79.45(0.35)	64.60(0.52)	72.76(0.32)	70.89(0.32)	64.56(0.39)	59.67(0.45)
	w/o Encoder Fine-tuning	79.83(0.26)	65.18(0.31)	73.29(0.28)	71.42(0.36)	65.15(0.37)	60.34(0.41)
	EMGLLM	**80.36(0.52)**	**66.03(0.69)**	**73.92(0.53)**	**71.83(0.54)**	**65.75(0.61)**	**61.05(0.66)**
Impressions	w/o Reference Value	74.88(0.29)	57.68(0.54)	72.62(0.24)	67.93(0.29)	59.56(0.40)	53.37(0.47)
	w/o Encoder Pre-training	76.27(0.29)	60.19(0.48)	74.15(0.38)	69.94(0.47)	62.16(0.53)	56.31(0.59)
	EMGLLM (rule-based)	76.93(0.49)	60.79(0.81)	74.83(0.54)	70.56(0.69)	62.69(0.83)	56.71(0.95)
	w/o Encoder Fine-tuning	**77.45(0.55)**	**61.72(0.70)**	**75.49(0.42)**	**71.15(0.48)**	**63.52(0.60)**	**57.58(0.66)**
	EMGLLM	77.21(0.41)	61.49(0.86)	75.26(0.43)	70.91(0.36)	63.29(0.64)	57.38(0.86)

Table 4: **Ablation Study Results.** All automatic evaluations are tested with 5 random seeds.

erated outputs of each model. These human experts are graduate students responsible for research and development projects in the medical technology field. The criteria for the human evaluation can be found in Appendix A. We ask human experts to score the *Findings* and *Impression* separately based on the following scoring criteria, and the average of these two scores is then taken as the overall score. The evaluations are conducted in a blinded manner, with human raters unaware of the model identities.

4.3 Results

4.3.1 Main Results

Table 2 presents the main results of the EMG automatic diagnosis generation. In automatic, model and human evaluations, it can be observed that EMGLLM outperforms all baseline methods, including data-to-text models such as Lattice and TableLLM-7B. In particular, the comparison between EMGLLM and Chinese-Alpaca-2-7B-16K clearly illustrates the improvement brought by the EMG Alignment Encoder to the LLM in EMG automatic diagnosis. This demonstrates that the EMGLLM framework effectively utilizes test values and reference values to reasonably encode numerical data in medical tables, resulting in higher-quality diagnosis generation.

Additionally, DeepSeek-R1 shows lower performance in three settings without fine-tuning, indicating that a general LLM without specific fine-tuning lacks the knowledge of medical data. This underscores the importance of datasets for medical tables and highlights the contribution of the ETM.

We also observe that for all models in the experiment, the rankings of the model evaluation metrics are basically consistent with those of the human evaluation scores, indicating that GPT-4o can serve as a substitute for human evaluation in our task.

4.3.2 Effectiveness on Data-intensive Input Scenario

To further validate the effectiveness of data encoding method, we extract samples with a relatively large number of continuous numerical data from the ETM dataset, resulting in a data-intensive subset. This subset contains 5,000 training samples and 600 test samples, with an average of 43.49 continuous numerical values per sample, higher than 33.14 shown in Table 1. As shown in Table 3, compared to the results from training and testing on the full dataset in Table 2, the performance gap between EMGLLM and Chinese-Alpaca-2-7B-16K widens, exceeding 3 in overall diagnoses, 2 in *Findings*, and 6 in *Impressions* in terms of ROUGE-2. Therefore, as the data amounts in the tables increase and the task becomes more challenging, EMGLLM demonstrates greater robustness.

4.3.3 Ablation Study

In Section 3.1, we propose a method for obtaining reference values and attempt to compare test values with them using the EMG Alignment Encoder. A natural question arises: once reference values are obtained, is it effective to directly convert continuous numerical data into categorical terms such as ”high”, ”normal”, or ”low” based on rules without the EMG Alignment Encoder? Therefore, we conduct experiment on a rule-based approach for processing data input. Specifically, if the test value for item i exceeds $u_i^{0.05}$, it is denoted as "high"; if it is below $l_i^{0.05}$, it is denoted as "low"; otherwise, it is denoted as "normal". The LLM trained by this rule-based method is denoted as EMGLLM (rule-based) in Table 4. It is shown that replacing the EMG Alignment Encoder with rules leads to a certain degradation in performance. This indicates that the rule-based method is relatively inflexible in handling medical examination tables. Besides, to verify the necessity of introducing reference values, we evaluate EMGLLM without reference values by replacing each detection item's reference values with random numbers from a standard normal distribution during model fine-tuning phase. As shown in Table 4, this leads to a significant performance drop.

We also conduct ablation study over training methods in Section 3.2. As shown in Table 4, pre-training the EMG Alignment Encoder is essential, resulting in a well-calibrated initialization. Fine-tuning the EMG Alignment Encoder in conjunction with the LLM on real EMG diagnostic datasets can further enhance the capabilities.

5 Conclusion

In this paper, we propose EMGLLM, a medical data-to-text model, for the automatic diagnosis generation of Electromyography (EMG) tables. The model framework with the EMG Alignment Encoder can enhance the encoding of continuous numerical data, enabling the model to simulate the process by which physicians compare test values to reference values during diagnosis. This approach facilitates a better model understanding of the degree of health and abnormality reflected by the data. In addition, we construct the ETM dataset, which comprises 17,250 real case examination EMG tables and diagnoses from authoritative doctors, to support the advancement of medical data-to-text research. Finally, experimental results demonstrate that EMGLLM outperforms baseline methods in all automatic, model and human evaluations for EMG diagnosis generation, confirming the effectiveness of the EMGLLM approach in handling medical examination data for automatic diagnosis.

Limitation

EMGLLM is designed to augment the model understanding of continuous numerical data in medical examination tables, without addressing other elements of the tables. At present, our experiments have been conducted solely on the EMG task. We will extend our model to other types of medical examinations, such as complete blood counts and urinalysis tables in future works.

Ethics Statement

We acknowledge the limitations of current LLMs and the ethical implications of their use, including the potential for inaccurate or misleading responses in diagnosis. However, our research has shown an improvement in reliability compared to baseline methods. Future research may explore more robust methods to address these challenges.

Our dataset is constructed from real medical diagnostic reports and contains basic information about patients. However, the dataset we publish is completely anonymous, and we only disclose age, gender, and height information without revealing any other private information. The release of this dataset has been approved for use by Huashan Hospital Affiliated with Fudan University.

All human participants involved in the evaluation of this research were compensated at or above the average local wage rate, ensuring fair remuneration for their time and contributions.

Acknowledgments

This research was supported by *National Natural Science Foundation of China* (Grant No. 62406121), *Natural Science Foundation of Hubei Province, China* (Grant No. 2024AFB189), *National Key R&D Program of China* (Grant No. 2023YFF1204800) and *National Natural Science Foundation of China* (Grant No. 62176058). The project's computational resources are supported by CFFF platform of Fudan University.

References

Andrea J Boon, Kais I Alsharif, C Michel Harper, and Jay Smith. 2008. Ultrasound-guided needle emg of the diaphragm: technique description and case report. *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, 38(6):1623–1626.

Arthur Bureau, Jean-Maxime Le Carpentier, Eric Le Carpentier, Yannick Aoustin, and Yann Péréon. 2021. Décomposition et analyse de tracés emg pour aider au diagnostic des maladies neuromusculaires. *Preprint*, arXiv:2109.14922.

Stanislas Chambon, Mathieu Galtier, Pierrick Arnal, Gilles Wainrib, and Alexandre Gramfort. 2017. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *Preprint*, arXiv:1707.03321.

Navin Cooray, Zhenglin Li, Jinzhuo Wang, Christine Lo, Mahnaz Arvaneh, Mkael Symmonds, Michele Hu, Maarten De Vos, and Lyudmila S Mihaylova. 2022. Automated movement detection with dirichlet process mixture models and electromyography. In *2022 25th International Conference on Information Fusion (FUSION)*, volume 31, page 01–08. IEEE.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Jasper R Daube. 2002. Assessing the motor unit with needle electromyography. *CONTEMPORARY NEUROLOGY SERIES*, 66:293–323.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Mekia Shigute Gaso, Selcuk Cankurt, and Abdulhamit Subasi. 2021. Electromyography signal classification using deep learning. In *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*. IEEE.

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, Francois Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. 2023. xval: A continuous number encoding for large language models. *Preprint*, arXiv:2310.02989.

Qizheng Gu, Cong Nie, Ruixiang Zou, Wei Chen, and Dong Tian. 2020. Automatic generation of electromyogram diagnosis report. *IEEE*.

YiQiu Guo, Yuchen Yang, Ya Zhang, Yu Wang, and Yanfeng Wang. 2024. Dictllm: Harnessing key-value data structures with large language models for enhanced medical diagnostics. *Preprint*, arXiv:2402.11481.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shailza Jolly, Zi Xuan Zhang, Andreas Dengel, and Lili Mou. 2021. Search and learn: Improving semantic coverage for data-to-text generation. *Preprint*, arXiv:2112.02770.

Bashar Katirji. 2002. The clinical electromyography examination: An overview. *Neurologic clinics*, 20(2):291–303.

X. Li, X. Zhang, X. Yi, D. Liu, H. Wang, B. Zhang, B. Zhang, D. Zhao, and L. Wang. 2023. Review of medical data analysis based on spiking neural networks. *Preprint*, arXiv:2212.02234.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Van Hieu Nguyen, Gia Thien Luu, Thien Van Luong, Mai Xuan Trang, Philippe Ravier, and Olivier Buttelli. 2023. After-fatigue condition: A novel analysis based on surface emg signals. *Preprint*, arXiv:2309.04770.

Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. 2020. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. *arXiv preprint arXiv:2010.02467*.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ildar Rakhmatulin. 2024. Pieeg-16 to measure 16 eeg channels with raspberry pi for brain-computer interfaces and eeg devices. *Preprint*, arXiv:2409.07491.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Ellen Tan, Yossef Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Defossez, Jade Copet, Faisal Azhar, Hugo Touvron, Gabriel Synnaeve, Louis Martin, Nicolas Usunier, and Thomas Scialom. 2023. Code llama: Open foundation models for code. Technical report, MetaAI.

Swarnadeep Saha, Xinyan Velocity Yu, Mohit Bansal, Ramakanth Pasunuru, and Asli Celikyilmaz. 2022. Murmur: Modular multi-step reasoning for semi-structured data-to-text generation. *Preprint*, arXiv:2212.08607.

Mandar Sharma, Ajay Gogineni, and Naren Ramakrishnan. 2024. Innovations in neural data-to-text generation: A survey. *Preprint*, arXiv:2207.12571.

Niklas Smedemark-Margulies, Yunus Bicer, Elifnur Sunger, Tales Imbiriba, Eugene Tunik, Deniz Erdogmus, Mathew Yarossi, and Robin Walters.

2023. Fast and expressive gesture recognition using a combination-homomorphic electromyogram encoder. *Preprint*, arXiv:2311.14675.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sam Wiseman, Arturs Backurs, and Karl Stratos. 2021. Data-to-text generation by splicing together nearest neighbors. *Preprint*, arXiv:2101.08248.

Ruslan Yermakov, Nicholas Drago, and Angelo Ziletti. 2021. Biomedical data-to-text generation via fine-tuning transformers. *arXiv preprint arXiv:2109.01518*.

Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. 2024a. Tablellama: Towards open large generalist models for tables. *Preprint*, arXiv:2311.09206.

Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2024b. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *Preprint*, arXiv:2403.19318.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over tabular and textual data. *Preprint*, arXiv:2401.13223.

A Model and Human Evaluation Details

In Sections 4.3.1 and 4.3.2, we introduce a GPT-4o-based model evaluation with 0.1 model temperature for stability. GPT-4o is tasked with scoring both the *Findings* and *Impression* sections, and we use the average of these two scores as the overall score of a diagnosis. The template is shown in Figure 5, where we provide the full scoring criteria for Correctness and Completeness, allowing GPT-4o to reference the authoritative doctor's diagnosis when assigning scores.

In the human evaluation, we ask human experts to score based on the following scoring criteria.

5 - The generated diagnosis is completely identical to the real diagnosis. Not only the conclusion but also the detailed descriptions are fully consistent.

4 - The generated diagnosis and the real diagnosis have identical conclusions, and most of the detailed descriptions are accurate. There may be minor omissions or incomplete descriptions in certain details, but these discrepancies do not affect the overall diagnostic conclusion.

3 - The generated diagnosis and the real diagnosis have the same direction, and the conclusions are generally consistent, but there are 1 to 2 notable discrepancies and slight inaccuracies in details.

2 - The generated diagnosis is largely inconsistent with the real diagnosis, with only a few minor details agreeing.

1 - The generated diagnosis is completely opposite to the real diagnosis. The conclusion is significantly erroneous, with a fundamentally incorrect assessment of the condition, which does not meet medical standards.

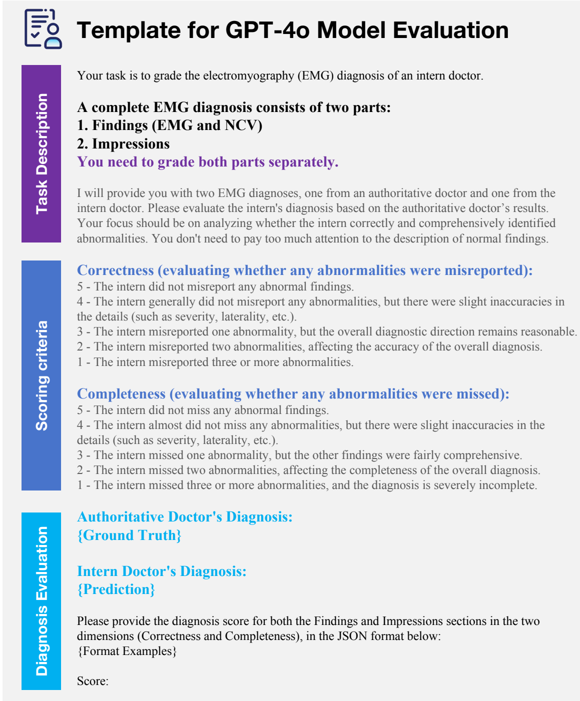


Figure 5: Template for GPT-4o evaluation of EMG diagnosis generation