NRAG: A Knowledge-Enhanced LLM Framework for Interpretable Neurosurgical Disease Diagnosis in Outpatient and Emergency Settings

1. Prompt templates

This study leverages custom-designed prompt templates to construct a clinical neurosurgical instruct-tuning dataset, specifically tailored for neurosurgical diseases, formalized as:

"Conversations":

[

"role":"user", "content":"*If you are a doctor in a neurosurgery outpatient department, based on the patient's clinical medical record data and knowledge graph, you can infer the possible diagnosis of the patient. Clinical medical record*: {***patient information (chief complaint, medical history, allergy history, and physical examination)***}. *The following are possible diagnoses*: {***KG-enhanced information***},

"role:": assistant", "content":"{***Diagnose***}"

]

2. BootStrap sampling

We used the Bootstrap sampling method (repeated 1000 times) to calculate the 95% confidence intervals for the F1 scores of each model. The confidence interval for NRAG is [0.7720, 0.8780], while the second-best model, DeepSeek, has an interval of [0.7635, 0.8627]. The overall performance distribution of NRAG is higher than that of DeepSeek. The t-test result is 8.1428, with $P < 0.0001$. Statistically, there is a significant performance difference between NRAG and the strongest baseline model.

3. Comparison experiment

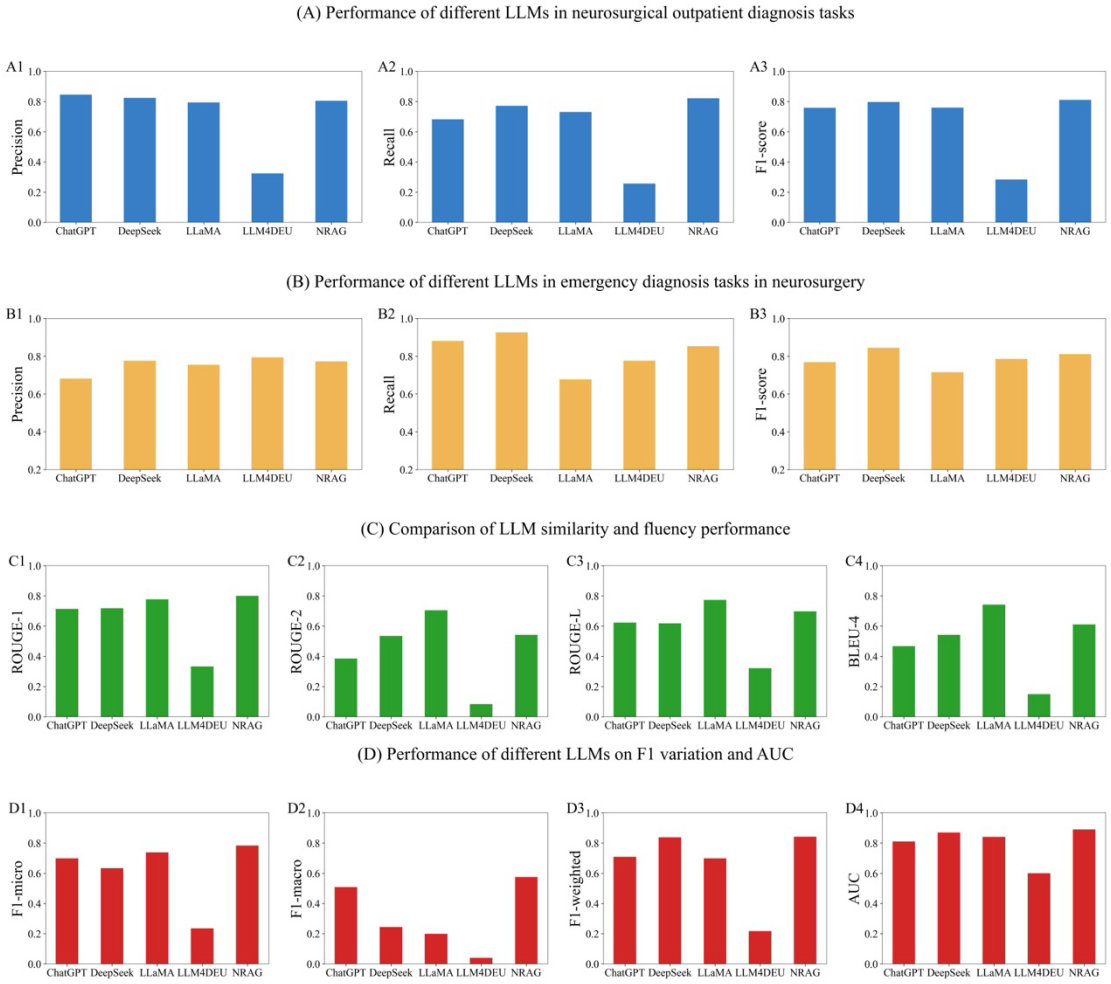

Fig. 1 Comparison experiment about NRAG. (A) Performance of different LLMs in neurosurgical outpatient diagnosis tasks, including precision, recall, and F1-score. (B) Performance of different LLMs in emergency diagnosis tasks in neurosurgery. (C) Comparison of LLM similarity and fluency performance, including ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4. (D) Performance of different LLMs on F1_micro, F1_macro, F1_weighted, and AUC.

4. Case analysis

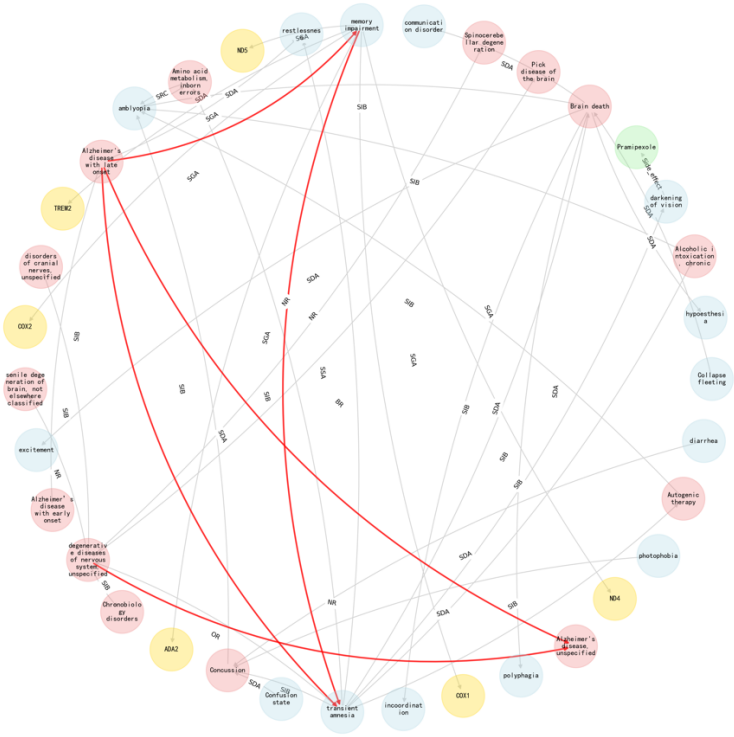

Fig .2 Case analysis of reasoning path. The red connections represent the key connection relationships identified during the information retrieval process, which are also the basis for the inference path.

The results of related diseases retrieved from the graph in the case of symptoms are shown in Fig. 2. We searched for subnets in the graph as displayed, where red nodes represent diseases, blue nodes represent symptoms, green nodes represent drugs, and yellow nodes represent genes (relation details are shown in Table 2). The connections between nodes represent the existence of correlation relationships during the period, and the red connections represent the key connection relationships identified during the information retrieval process, which are also the basis for the inference path. The identified symptoms of “Transient forgetfulness and Impaired Memory” can be retrieved from the knowledge graph as Alzheimer's disease and other nervous system diseases, thus indicating the improvement of interpretability of large models in the knowledge graph.

Table 2 Names and abbreviations of edges

Name	Abbreviations
Broader Relationship	BR
Narrow Relationship	NR
Sibling Relationship	SR
Side_effect	SE
Source-Related Correlation	SRC
Symptom_Disease_association	SDA
Symptom_Gene_association	SGA
Symptom_Symptom_association	SSA
Other Relationship	OR

Only the edges involved in the figure are shown in the table. For other types of edges, please refer to the knowledge graph.