Supplementary Materials for

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}***,

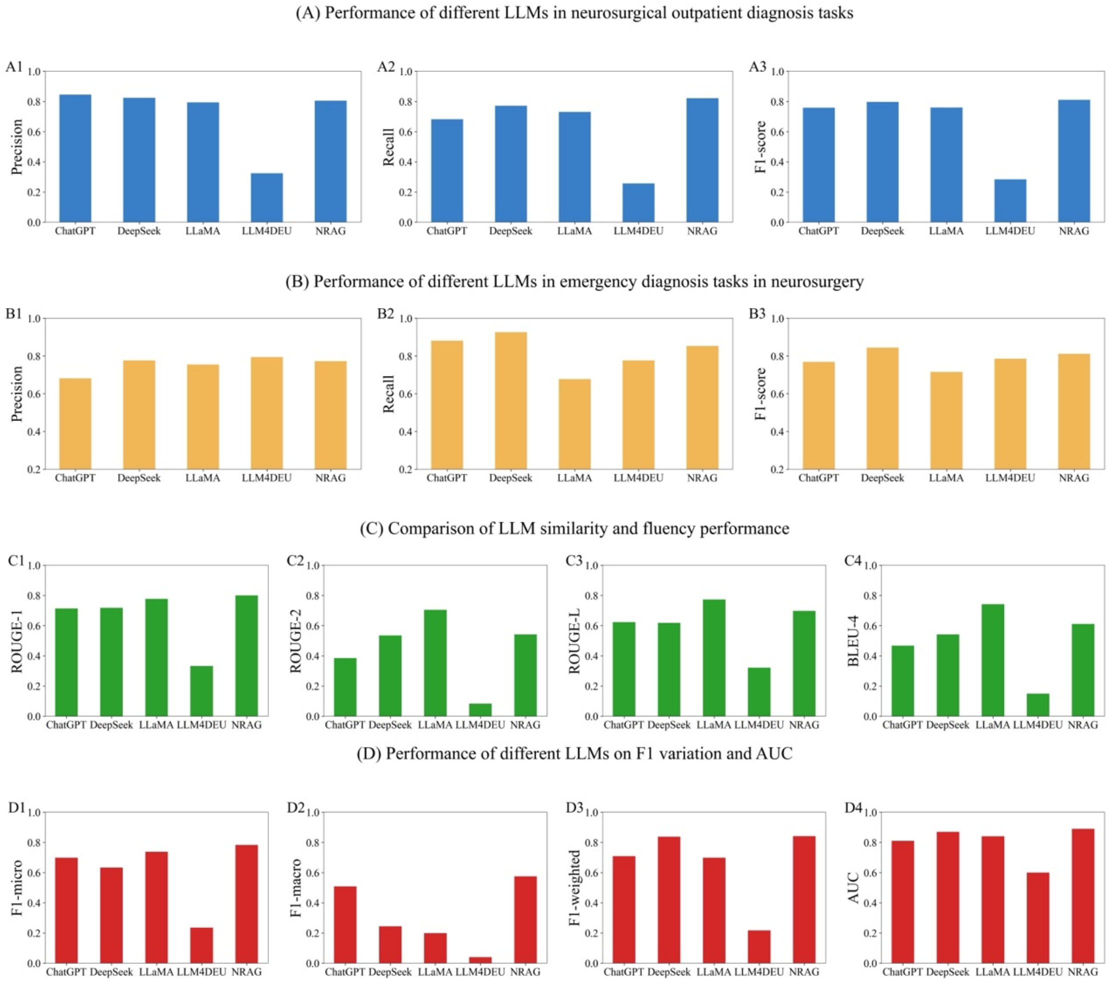
“role:”: assistant”, “content”:“***{Diagnose}***”

]

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



(A) Performance of different LLMs in neurosurgical outpatient diagnosis tasks

(B) Performance of different LLMs in emergency diagnosis tasks in neurosurgery

(C) Comparison of LLM similarity and fluency performance

(D) Performance of different LLMs on F1 variation and AUC

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. Ablation experiments on ED dataset

Table 1 Results of ablation experiments on ED dataset

	ED		
Model	**Precision**	**Recall**	**F1-score**
ChatGLM3$_{KG}$	0.2301±0.0387	0.1933±0.0418	0.2101±0.0420
ChatGLM3$_{KG+shots}$	0.3503±0.0183	0.3275±0.0215	0.3384±0.0195
NRAG	0.7723±0.0247	**0.8531±0.0203**	**0.8107±0.0223**
NRAG$_{PGL}$	0.7442±0.0306	0.8469±0.0152	0.7922±0.0241
NRAG$_{NPI}$	**0.7802±0.0225**	0.8375±0.0230	0.8078±0.0213
NRAG$_{RDA}$	0.4448±0.0231	0.6467±0.0368	0.5271±0.0283
NRAG$_{w/o\ KGI}$	0.4683±0.0293	0.6875±0.0563	0.5570±0.0383

ChatGLM3$_{KG}$: ChatGLM3-6B with KG Information (w/o fine-tuning).

ChatGLM3$_{KG+shots}$: ChatGLM3-6B with KG Information and few-shot prompts (w/o fine-tuning).

NRAG$_{PGL}$: NRAG with Partial Gold Label.

NRAG$_{NPI}$: NRAG with Negative Prompt Inclusion.

NRAG$_{RDA}$: NRAG with Random Disease Assumption.

NRAG$_{w/o\ KGI}$: NRAG without KG Information.

NRAG$_{KGPI}$: NRAG with KG Path format Information.

The experimental results show that, on the ED dataset, the complete NRAG model also achieves optimal performance (F1=0.8107). It is noteworthy that modifying the prompt (NRAG$_{NPI}$, F1=0.8078) and partially removing the gold labels (NRAG$_{PGL}$, F1=0.7922) only led to slight performance degradation, further demonstrating the robustness of our method in the presence of prompt interference or partial knowledge absence. Moreover, both of these strategies performed significantly better than injecting a random candidate disease list (F1=0.5570), which further confirms that high-quality knowledge graph information is crucial for the model's accurate reasoning.

In conclusion, the ablation experiments conducted on both the OD and ED datasets consistently validate the effectiveness of the core components of the NRAG framework and its strong generalizability.

5. Case analysis

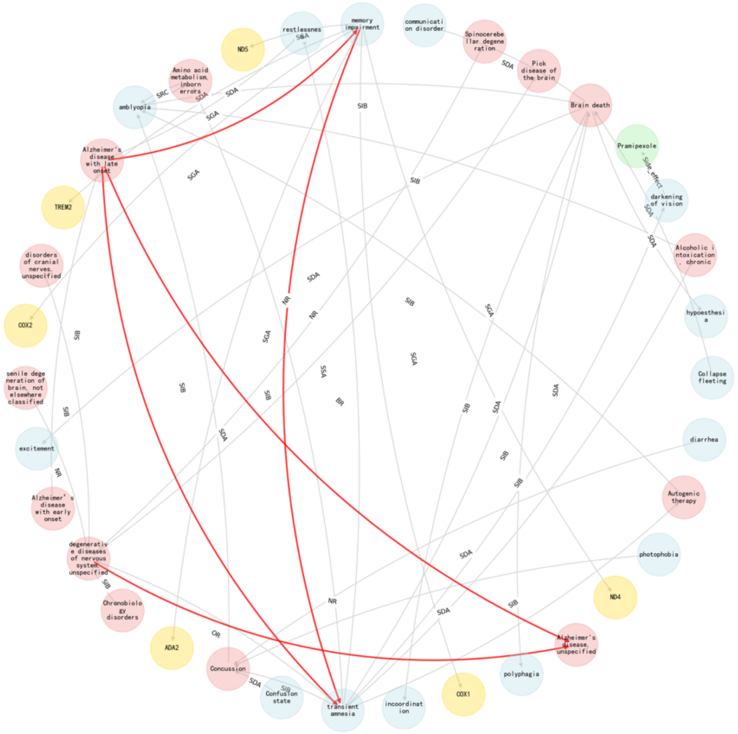

memory impairment
communication disorder
Spinocerebellar degeneration
Pick disease of the brain
Brain death
Pramipexole
Side_effect
darkening of vision
Alcoholic intoxication, chronic
hypoesthesia
Collapse fleeting
diarrhea
Autogenic therapy
photophobia
ND4
Alzheimer's disease, unspecified
polyphagia
COX1
incoordination
transient amnesia
Confusion state
Concussion
ADA2
Chronobiology disorders
degenerative diseases of nervous system, unspecified
Alzheimer's disease with early onset
excitement
senile degeneration of brain, not elsewhere classified
COX2
disorders of cranial nerves, unspecified
TREM2
Alzheimer's disease with late onset
amblyopia
Amino acid metabolism, inborn errors
ND5
restlessness
SDA
SGA
SIB
SRC
SSA
NR
BR
OR

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.