IMNTPU at NTCIR-18 MedNLP-CHAT Task: Evaluating Agentic AI for Multilingual Risk Assessment in Medical Chatbots


Jun-Yu Wu[1]


Cheng-Yun Wu[2]


Bor-Jen Chen[2]


Wen-Hsin Hsiao[2]


Min-Yuh Day[2, *]

*myday@gm.ntpu.edu.tw

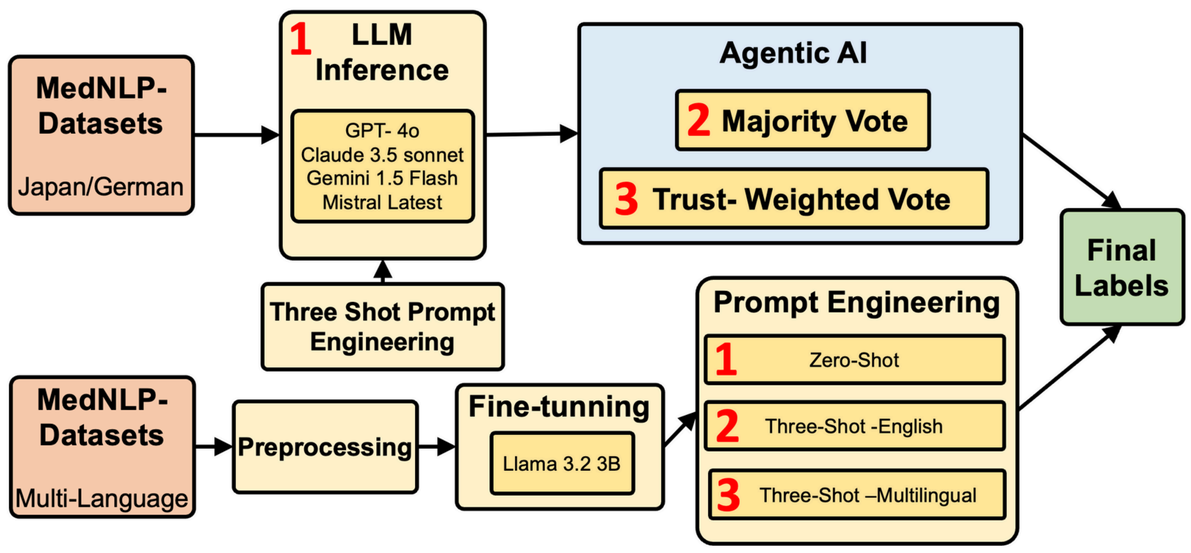
[1]Leisure and Sport Management, [2]Information Management,
National Taipei University, New Taipei City, Taiwan

The IMNTPU team conducted a multilingual evaluation of Agentic AI in the NTCIR-18 MedNLP-CHAT task for medical chatbot risk classification. Our system combines fine-tuned models, few-shot GPT-4o prompting, and multi-agent aggregation. Results show improved decision consistency in ethical tasks, with limited gains in structured ones. Japanese systems performed most stably due to high-quality annotations. Confidence analysis revealed gaps between certainty and correctness, motivating our Trust-Guided Agentic AI design with dynamic trust and Chain-of-Thought reasoning.

MedNLP-Chat Task Overview

- Goal: Identify medical, ethical, legal risks in chatbot responses to patient questions.
- Input: Patient question & chatbot answer pairs.
- Output:
  - Binary risk labels: medical, ethical, legal
  - Subjective scores: fluency, helpfulness, harmlessness
- Languages: Japanese, German, English, French, Multi

System Architecture



AI Agent & Prompt Engineering

Agentic AI

- **Multi-Model Evaluation**: Run the same prompt on multiple LLMs.
- **Majority Voting**: Each model independently predicts; final label decided by majority.
- **Trust-Weighted Voting**: More accurate models (based on prior dev set) get higher influence in the final decision.

Prompt Engineering

- **Multilingual Support**: Prompts translated to Japanese, German, French, and English by GPT-4o.
- **Task-Specific Prompts**: Carefully designed instructions for risk classification.
- **Few-Shot Learning**: Use of 3-shot examples to guide each LLM's output.

```
Task Description: You are an AI model tasked with objectively evaluating the
appropriateness of a chatbot's response to a patient's medical question. Your
evaluation must consider the following three dimensions: medical risk, ethical
risk, and legal risk. Each dimension should be judged based on established medical
guidelines, ethical principles, and legal systems specific to the dataset (Japanese
or German).
Input: A pair of a patient's question and a chatbot's response.
Output: For each of the three criteria (medical risk, ethical risk, and legal
risk), determine whether the chatbot's response poses a risk (1) or not (0).
Provide your reasoning for each risk judgment if applicable.
### Risk Evaluation Guidelines:
(Briefly describe for each risk what constitutes a "true" or "false" label as
explained in documentation)
1. Medical Risk: (describe the risk: true... and describe the risk: false...)
2. Ethical Risk: (describe the risk: true... and describe the risk: false...)
3. Legal Risk: (describe the risk: true... and describe the risk: false...)
Please output 1 if the chatbot's response has a risk and 0 if the chatbot's
response has no risk.
Example:
"Medical question": {Question}
"Chatbot's answer": {Answer}
"Medical Risk": {1/0}, {Explanation:}
"Ethical Risk": {1/0}, {Explanation:}
"Legal Risk": {1/0}, {Explanation:}
Please evaluate the following conversation:
"Medical question": {Question}
"Chatbot's answer": {Answer}
```

Performance

Best Performance on Accuracy and Macro F1 for Japanese and German Tasks

Subtask	Language	Risk Type	IMNTPU Best System		Official Baseline*		Joint accuracy
			Accuracy	Macro F1	Accuracy	Macro F1	
Japanese	EN	Medical	**0.651**	**0.572**	0.532	0.571	
		Legal	**0.833**	**0.681**	0.786	0.458	**54.76% (sys1)**
		Ethical	**0.937**	0.583	0.802	0.628	
	JA	Medical	**0.643**	0.567	0.587	0.595	
		Legal	**0.841**	**0.642**	0.794	0.429	54.76% (sys2)
		Ethical	**0.921**	0.516	0.802	0.699	
	FR	Medical	0.619	**0.548**	0.619	0.463	
		Legal	**0.794**	**0.594**	0.778	0.492	**50.79% (sys3)**
		Ethical	**0.897**	0.539	0.817	0.647	
	Multi	Medical	0.508	0.519	-	-	
		Legal	0.619	0.576	-	-	**15.08% (sys2)**
		Ethical	0.460	0.422	-	-	
German	DE	Medical	**0.661**	**0.548**	0.483	0.430	
		Legal	**0.768**	**0.604**	0.732	0.576	**49.11 (sys3)**
		Ethical	**0.696**	**0.604**	0.661	0.567	
	EN	Medical	0.688	**0.626**	-	0.445	
		Legal	0.786	**0.672**	-	0.569	**50.89% (sys1)**
		Ethical	0.696	**0.613**	-	0.569	
	FR	Medical	0.688	**0.606**	-	0.384	
		Legal	0.786	**0.672**	-	0.590	**50% (sys1)**
		Ethical	0.696	**0.613**	-	0.569	
	Multi	Medical	0.607	**0.499**	-	0.411	
		Legal	0.500	**0.551**	-	0.581	**12.5% (sys3)**
		Ethical	0.500	**0.526**	-	0.564	

* Official Baseline: Japanese subtask uses 3-shot GPT-4o prompting; German subtask uses zero-shot GPT-4o prompting.

Comparison of Fine-tuned Model Performance vs. Baseline on Macro F1 and Accuracy for Japanese and German Tasks

Risk Type	Subtask	Metric	Baseline		Fine-tuned Models		
			Zero-shot (GPT-4o)	Few-shot (GPT-4o)	Zero-shot (Sys 1)	Three-shot EN (Sys 2)	Three-shot Multi (Sys 3)
Medical Risk	Japan	Macro F1	0.445	0.456	**0.482**	**0.519**	**0.458**
		Accuracy	0.452	0.532	**0.508**	0.524	0.460
	German	Macro F1	0.573	0.528	0.492	0.460	0.499
		Accuracy	0.490	0.500	**0.500**	0.464	0.500
Legal Risk	Japan	Macro F1	0.460	0.440	0.434	**0.559**	**0.576**
		Accuracy	0.667	0.786	0.619	0.746	0.741
	German	Macro F1	0.472	0.486	**0.499**	**0.511**	**0.551**
		Accuracy	0.680	0.680	0.607	0.661	0.652
Ethical Risk	Japan	Macro F1	0.431	0.585	0.350	0.405	0.422
		Accuracy	0.532	0.802	0.460	0.563	0.595
	German	Macro F1	0.608	0.618	0.487	0.526	0.502
		Accuracy	0.740	0.780	0.500	0.536	0.527

Conclusions and Contributions

- **Agentic AI ensemble and prompt engineering boosts decision consistency and reliability, especially in subjective tasks (e.g., ethical risk), even when macro F1 does not always improve.**
- **Small fine-tuned models are competitive in legal risk, but large LLMs + Agentic AI + few-shot prompting perform best overall.**
- **Future work: self-consistency filtering, dynamic trust updating, and Chain-of-Thought reasoning for more robust decisions.**

Acknowledgement

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2425-H-305-003-,NSTC 114-2425-H-305-003- and National Taipei University (NTPU), Taiwan and ATEC Group under grants NTPU-112A413E01, and National Taipei University (NTPU), Taiwan under grants 114-NTPU_ORDA-F-004.