NTCIR-18 MedNLP-CHAT Determining Medical, Ethical and Legal Risks in Patient-Doctor Conversations: Task Overview

Eiji Aramaki[1], Shoko Wakamiya[1], Shuntaro Yada[1], Shohei Hisada[1], Tomohiro Nishiyama[1], Lenard Paulo Tamayo[1], Jingnan Xiao[1], Axalia Levenchaud[1], Pierre Zweigenbaum[2], Christoph Otto[3], Jerycho Pasniczek[3], Philippe Thomas[3], Nathan Pohl[4], Wiebke Duettmann[4], Lisa Raithel[5], Roland Roller[3]

1 NAIST, Japan 2 Université Paris-Saclay, France 3 DFKI Berlin, Germany 4 Charité, Germany 5 BIFOLD, TU Berlin, Germany

https://sociocom.naist.jp/mednlp-chat/

Important Problem: Medically correct, but inappropriate for a chatbot to say

About MedNLP-CHAT Task

To determine whether a chatbot's answer to a medical question is appropriate from multiple perspectives

INPUT: A pair of a patient's question and a chatbot answer

OUTPUT:

- **Objective evaluation** by a specialist: Binary class, "risk" (1) and "no risk" (0)
  Medical risk, Ethical risk, and Legal risk
- **Subjective evaluation** by the general public (Japanese dataset only):
  A probability distribution of evaluations on a 5-point scale, "very non-fluent"(-2) to "very fluent" (+2)
  Fluency, Helpfulness, and Harmlessness

Datasets

Japanese dataset (ja, en, fr)

- 226 Q&A pairs (100 for training and 126 for testing)
- Questions were collected through crowdsourcing and answers were created by various chatbots (e.g., GPT4o, ChatGPT, etc.)
- Objective evaluation was done by a specialist and subjective evaluation was done through crowdsourcing

German dataset (de, en, fr)

- 212 Q&A pairs (100 for training and 112 for testing)
- All Qs pertain to nephrology and were curated by nephrology specialists

Each Team's Approaches

Team	LLM(s) Used	Methods
AITOK	GPT-3.5 Turbo (baseline), GPT-4o (proposed)	Two types of LLMs used
IMNTPU	GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Flash, Mistral Small Latest	Zero-shot and 3-shot prompts
NAISTym	Gemini-1.5-Flash, GPT-4o	Chain-of-thought and few-shot prompts
TMU2025	ClinicalBERT (same size as BERT-base)	Transformer-based neural classifier (6 transformer blocks) for word embeddings
TMUNLPG2	BERT-based classification system (bert-base-japanese-v3, japanese-sentiment-analysis), Llama3.1-8B for data augmentation (DA)	BERT-based classification and LLM for data augmentation
TUSNLP	JMedRoBERTa (encoder model), GPT-3.5, GPT-4-mini (decoder models), Llama3.1-8B for DA	Back-translation, data summarization via ChatGPT, Manbyo Dictionary for medical terms, Wikipedia articles for RAG
UEM24	No LLMs used	Pre-processing (tokenization, n-gram extraction, lemmatization), Logistic Regression (LR), combination of two datasets via English language
UPxSocio	Gemini-1.5-Flash	Similarity-based RAG with *k*-nearest and *k*-spread strategies, few-shot prompting (generate support statement, predicted risk)
UTSolve	BioBERT v1.1, MedBERT, ClinicalBERT	Fine-tuning of BioBERT, evaluation of MedBERT and ClinicalBERT models

Results

The best-performing system from each team based on macro F1 score and joint accuracy score

BASELINEs refer to GPT-4o in zero/few/10-shot settings, respectively

Japanese Subtask

Team	Risk	Language: JA	EN	FR	Multi
$\text{BASELINE}_{ja}^{zero-shot}$	Medical Risk	0.387	0.445	0.366	-
	Ethical Risk	0.485	0.431	0.481	-
	Legal Risk	0.446	0.460	0.442	-
$\text{BASELINE}_{ja}^{few-shot}$	Medical Risk	0.568	0.637	0.585	-
	Ethical Risk	0.674	0.687	0.711	-
	Legal Risk	0.635	0.636	0.470	-
AITOK	Medical Risk	0.557 (sys3)	0.522 (sys1)	**0.571 (sys3)**	-
	Ethical Risk	0.579 (sys3)	0.513 (sys1)	**0.590 (sys2)**	-
	Legal Risk	0.531 (sys1)	0.511 (sys1)	0.560 (sys1)	-
	Joint Accuracy (%)	41.27 (sys3)	34.13 (sys1)	34.92 (sys1)	-
IMNTPU	Medical Risk	0.567 (sys2)	0.572 (sys3)	0.548 (sys1)	**0.519 (sys2)**
	Ethical Risk	0.516 (sys2)	0.583 (sys1)	0.539 (sys3)	**0.422 (sys3)**
	Legal Risk	0.642 (sys3)	0.681 (sys1)	**0.594 (sys1)**	**0.576 (sys3)**
	Joint Accuracy (%)	54.76 (sys2)	**54.76 (sys1)**	**50.79 (sys3)**	**15.08 (sys2)**
NAISTym	Medical Risk	**0.569 (sys2)**	-	-	-
	Ethical Risk	0.610 (sys2)	-	-	-
	Legal Risk	0.588 (sys1)	-	-	-
	Joint Accuracy (%)	**55.56 (sys3)**	-	-	-
TMUNLPG2	Medical Risk	0.531 (sys2)	-	-	-
	Ethical Risk	**0.662 (sys2)**	-	-	-
	Legal Risk	**0.741 (sys1)**	-	-	-
	Joint Accuracy (%)	46.03 (sys1)	-	-	-
TMU2025	Medical Risk	-	0.547 (sys3)	-	-
	Ethical Risk	-	0.470 (sys3)	-	-
	Legal Risk	-	0.384 (sys3)	-	-
	Joint Accuracy (%)	-	20.63 (sys3)	-	-
TUSNLP	Medical Risk	0.435 (sys1)	-	-	-
	Ethical Risk	0.610 (sys2)	-	-	-
	Legal Risk	0.524 (sys3)	-	-	-
	Joint Accuracy (%)	41.27 (sys2&3)	-	-	-
UEM24	Medical Risk	-	0.500 (sys1)	-	-
	Ethical Risk	-	0.590 (sys1)	-	-
	Legal Risk	-	0.440 (sys1)	-	-
	Joint Accuracy (%)	-	44.44 (sys1)	-	-
UPxSocio	Medical Risk	-	**0.603 (sys1)**	-	-
	Ethical Risk	-	0.436 (sys2)	-	-
	Legal Risk	-	0.416 (sys2)	-	-
	Joint Accuracy (%)	-	19.05 (sys2)	-	-
UTSolve	Medical Risk	-	0.416 (sys1)	-	-
	Ethical Risk	-	**0.653 (sys1)**	-	-
	Legal Risk	-	**0.725 (sys1)**	-	-
	Joint Accuracy (%)	-	40.48 (sys1)	-	-

German Subtask

Team	Risk	Language: DE	EN	FR	Multi
$\text{BASELINE}_{de}^{zero-shot}$	Medical Risk	0.430	0.445	0.384	0.411
	Ethical Risk	0.567	0.569	0.569	0.564
	Legal Risk	0.576	0.569	0.590	0.581
$\text{BASELINE}_{de}^{10-shot}$	Medical Risk	0.543	0.563	0.562	0.605
	Ethical Risk	0.752	0.668	0.669	0.642
	Legal Risk	0.644	0.648	0.610	0.587
AITOK	Medical Risk	**0.660 (sys3)**	-	-	-
	Ethical Risk	**0.612 (sys3)**	-	-	-
	Legal Risk	**0.667 (sys3)**	-	-	-
	Joint Accuracy (%)	41.96 (sys3)	-	-	-
IMNTPU	Medical Risk	0.548 (sys1)	**0.626 (sys2)**	**0.606 (sys1)**	**0.499 (sys3)**
	Ethical Risk	0.604 (sys1)	0.613 (sys1)	**0.613 (sys1)**	**0.526 (sys2)**
	Legal Risk	0.604 (sys1)	**0.672 (sys1)**	**0.672 (sys1)**	**0.551 (sys3)**
	Joint Accuracy (%)	**49.11 (sys3)**	**50.89 (sys1)**	**50.00 (sys1)**	**12.50 (sys3)**
TMU2025	Medical Risk	-	0.349 (sys3)	-	-
	Ethical Risk	-	0.408 (sys1)	-	-
	Legal Risk	-	0.356 (sys1,2&3)	-	-
	Joint Accuracy (%)	-	16.96 (sys1)	-	-
UEM24	Medical Risk	-	0.594 (sys1)	-	-
	Ethical Risk	-	0.619 (sys1)	-	-
	Legal Risk	-	0.658 (sys1)	-	-
	Joint Accuracy (%)	-	33.04 (sys1)	-	-
UPxSocio	Medical Risk	-	0.614 (sys1)	-	-
	Ethical Risk	-	**0.678 (sys1&2)**	-	-
	Legal Risk	-	0.591 (sys1)	-	-
	Joint Accuracy (%)	-	37.50 (sys1)	-	-

Discussions

- **Legacy Machine Learning V.S. LLMs:** While LLMs dominate current NLP tasks, classical methods still represent viable, resource-efficient alternatives to LLMs in certain scenarios
- **Difficulty of thee risks:** Medical risk is the most challenging category due to the complexity and variability of clinical contexts, requiring nuanced reasoning and domain-specific knowledge
- **Contribution of data augmentation (DA):** DA proved effective in addressing data imbalance, especially in ethical and legal risks, by enhancing model robustness and offering a competitive edge without relying on large models or extensive external resources