AITOK at the NTCIR-18 MedNLP-CHAT to Identify Medical, Ethical and Legal Risks in Patient-Doctor Conversations

Hiroki Tanioka
Tokushima University, Japan

Introduction

Background and Motivation: Artificial Intelligence (AI) is revolutionizing healthcare by enabling advanced data processing, clinical decision support, and patient interaction tools. Among recent innovations, Generative AI — especially Large Language Models (LLMs) — has emerged as a powerful approach for enhancing medical dialogue systems. However, the use of AI in healthcare raises critical concerns regarding accuracy, ethical appropriateness, and legal responsibility. Misjudgment in AI-generated responses can result in medical errors, misinformation, or breaches of patient trust. Hence, evaluating AI's capacity to manage these dimensions is essential before real-world deployment.

Keywords: Generative AI, Large Language Models, Medical Risk, Ethical Risk, Legal Risk, GPT-4o, Prompt Engineering, NTCIR-18

Task Overview

The NTCIR-18 MedNLP-CHAT task [1] focuses on evaluating the ability of AI systems to detect medical, ethical, and legal risks in patient-doctor dialogues. Participants were required to analyze simulated conversational data in both Japanese and German.

AITOK Models

The AITOK team submitted three models:

- **AITOK1:** A control baseline generating responses based on random probabilities (Table 1) derived from the training dataset.
- **AITOK2:** Utilized OpenAI's GPT-3.5 Turbo, selected for its balance of efficiency and performance.
- **AITOK3:** Applied GPT-4o, the latest high-performance LLM with stronger contextual understanding and inference capabilities.

These methods allowed a comparative analysis across model sophistication levels and multilingual applicability.

Methodology

Our approach focused on developing a prompt-based evaluation framework tailored to each risk category. The methodology comprised the following key components:

- **Prompt Structure:** Each prompt followed a standardized template including: (1) role assignment (e.g., medical, ethical, or legal expert), (2) patient query, (3) doctor response, (4) decision criteria, and (5) a clear output format. Prompts were crafted in Japanese and translated into German and French using DeepL© to ensure consistency across languages.
- **Model Execution:** Prompts were submitted to LLMs (GPT-3.5 Turbo and GPT-4o) via OpenAI API. Each input case was evaluated five times to reduce variance, and majority voting was used to assign the final TRUE/FALSE classification.
- **Risk Calibration:** To align model outputs with the class distributions of the training data, we calibrated the expected proportion of "TRUE" responses based on the original occurrence rates (e.g., 32% for medical risk). This minimized overestimation and false positives.
- **Evaluation Criteria:** For each risk type:

Medical
- Emphasis on scientific accuracy, safety, and urgency.

Ethical
- Focused on appropriateness of communication, empathy, and respect.

Legal
- Assessed based on likelihood of liability under severe misadvise conditions.

This methodology ensured a reproducible and interpretable framework for risk assessment in AI-mediated medical consultations.

Table 1: Percentage of risk occurrence in Training data.

Risk Type	TRUE	FALSE	Percent at Risk
Medical Risk	32	68	32%
Ethical Risk	20	80	20%
Legal Risk	24	76	24%

Evaluation Metrics

To quantitatively assess performance, the following metrics were used:

- **Accuracy (Acc):** Overall rate of correct classifications
- **F1 Macro (F1m):** Harmonic mean of precision and recall across classes
- **Precision (Pre):** Rate of true positive classifications among all positives
- **Recall (Rec):** Rate of true positive classifications among actual positives
- **False Positive Rate (FPR):** Proportion of false positives among all negatives and calculate false positive according to Eq. (1) based on Table 2.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{1}$$

Table 2: Confusion matrix calculated from actual and predicted conditions.

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

These metrics enabled comprehensive evaluation of model performance in identifying nuanced conversational risks.

Prompt Engineering

Effective prompt engineering was critical to guiding the LLMs to produce contextually accurate and risk-sensitive outputs. Each prompt was constructed with the following considerations (Fig. 1).

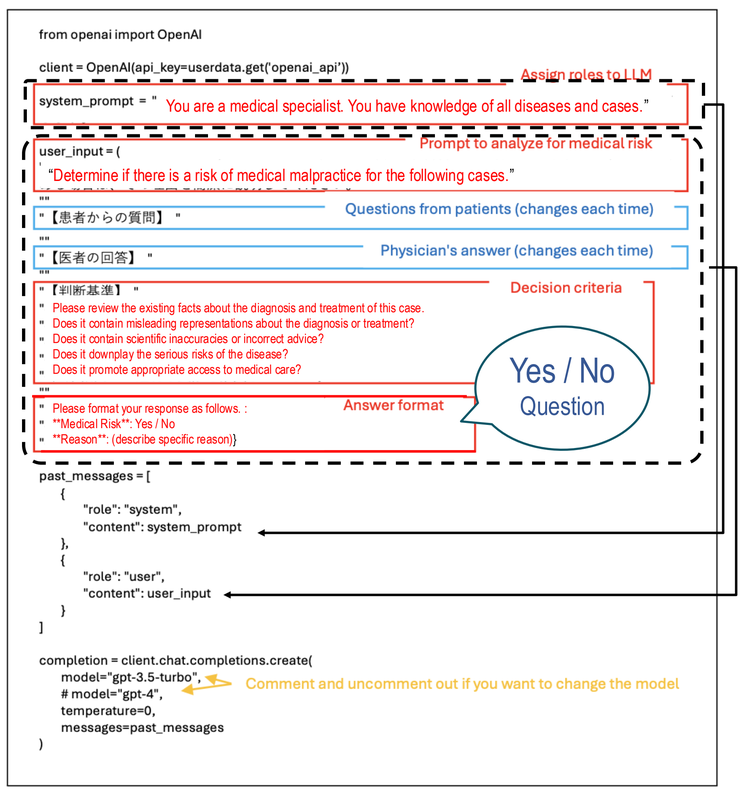

Figure 1: Prompts for checking medical context in Japanese.

- **Role Definition:** Clearly specifying the system's role (e.g., medical expert, legal advisor) helped align the model's reasoning path.
- **Decision Framework:** Detailed criteria were provided within the prompt to focus model attention on relevant aspects, such as medical accuracy, patient safety, and ethical tone.
- **Structured Output Format:** All responses followed a binary format (e.g., "Medical Risk: Yes/No" with a justification) to facilitate evaluation and aggregation.
- **Temperature and Repetition:** Temperature was set to zero to reduce randomness. Five prompt trials per case were used, with majority voting to enhance robustness.
- **Language Adaptation:** Prompts were tested in Japanese, German, and French to validate multilingual robustness through consistent structural translation.

In order to align the model's output with the class distribution of the training data, the same prompt was run multiple times (10 times in this case), and a "TRUE" response was assigned when the percentage of "Yes" responses exceeded a threshold value.

This disciplined approach to prompt design enabled scalable, interpretable, and reproducible evaluations across different risk dimensions and languages.

Results Summary

Japanese Subtask

In Japanese subtask, GPT-4o achieved the highest accuracy in medical and ethical risk detection, while the random baseline unexpectedly outperformed LLMs in legal risk, highlighting challenges in legal judgment modeling (Table 3-5).

Table 3: Results of medical risk for Japanese subtask in Japanese (ja).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 ja	0.571	0.538	0.293	0.543	0.540
AITOK 2 ja	0.389	0.380	0.573	0.384	0.380
AITOK 3 ja	**0.651**	**0.557**	**0.067**	**0.674**	**0.584**

Table 4: Results of ethical risk for Japanese subtask in Japanese (ja).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 ja	0.746	0.481	0.220	0.505	0.515
AITOK 2 ja	0.413	0.360	0.627	0.549	0.686
AITOK 3 ja	**0.794**	**0.579**	**0.195**	**0.574**	**0.715**

Table 5: Results of legal risk for Japanese subtask in Japanese (ja).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 ja	**0.706**	**0.531**	**0.231**	**0.534**	**0.551**
AITOK 2 ja	0.659	0.513	0.296	0.527	0.546
AITOK 3 ja	0.651	0.451	0.269	0.467	0.449

German Subtask

In German subtask, GPT-4o showed the best performance across all risk types in the German subtask, confirming its multilingual strength, with legal risk detection notably more accurate than in the Japanese subtask (Table 6-7).

Table 6: Results of legal risk for German subtask in German (de).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 de	0.420	0.364	0.403	0.359	0.376
AITOK 2 de	0.580	0.508	**0.194**	0.536	0.525
AITOK 3 de	**0.679**	**0.660**	0.239	**0.664**	**0.658**

Table 7: Results of legal risk for German subtask in German (de).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 de	0.589	0.514	0.179	0.548	0.533
AITOK 2 de	0.616	0.494	**0.075**	0.602	0.540
AITOK 3 de	**0.616**	**0.612**	0.403	**0.616**	**0.621**

Table 8: Results of legal risk for German subtask in German (de).

Run Type	Acc	F1m	FPR	Pre	Rec
AITOK 1 de	0.634	0.528	0.215	0.534	0.529
AITOK 2 de	0.438	0.432	0.759	0.598	0.575
AITOK 3 de	**0.750**	**0.667**	**0.114**	**0.698**	**0.655**

Key Insights

GPT-4o showed consistently strong performance in medical and ethical risk detection. Legal risk remained more challenging, with mixed results across languages. Rapid translation, with appropriately designed prompts, proved effective for multilingual tasks.

Conclusion

This study demonstrated that GPT-4o holds substantial promise for use in AI-assisted medical dialogues, particularly in identifying medical and ethical risks. Nevertheless, challenges remain in ensuring legal safety. Future systems should incorporate knowledge bases or rules for legal reasoning and expand prompt templates for nuanced discourse. Controlling False Positive Rates is crucial in sensitive domains like healthcare, and prompt design must reflect clinical realities and regulatory expectations.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP22K12293. We express our gratitude to the NTCIR-18 organizers for the task setup and to DeepL© for providing accurate multilingual translations.

Reference

[1] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, Shohei Hisada, Tomohiro Nishiyama, Lenard Paulo Velasco Tamayo, Jingnan Xiao, Axalia Levenchaud, Pierre Zweigenbaum, Christoph Otto, Jerycho Pasniczek, Philippe Thomas, Nathan Pohl, Wiebke Duettmann, Lisa Raithel, and Roland Roller. 2025. NTCIR-18 MedNLP-CHAT Determining Medical, Ethical and Legal Risks in Patient-Doctor Conversations: Task Overview. In *Proceedings of the NTCIR-18 Conference.*