

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

Jun-Yu Wu
Leisure and Sport Management,
National Taipei University
New Taipei City, Taiwan
s411084021@gm.ntpu.edu.tw

Cheng-Yun Wu
Information Management,
National Taipei University
New Taipei City, Taiwan
s711336101@gm.ntpu.edu.tw

Bor-Jen Chen
Information Management,
National Taipei University
New Taipei City, Taiwan
s711336103@gm.ntpu.edu.tw

Wen-Hsin Hsiao
Information Management,
National Taipei University
New Taipei City, Taiwan
s711336107@gm.ntpu.edu.tw

Min-Yuh Day*
Information Management,
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

Abstract

The IMNTPU team presents a multilingual evaluation of Agentic AI for chatbot risk classification in the NTCIR-18 MedNLP-CHAT task. Our framework integrates fine-tuned small models, optimized few-shot prompting with GPT-4o, and multi-agent aggregation via majority and trust-weighted voting. Results show that Agentic AI enhances decision consistency, especially in subjective tasks like ethical risk, but yields limited gains in structured domains such as medical and legal assessment. Language-specific outcomes reveal that annotation quality and linguistic complexity jointly affect model performance, with Japanese systems showing the most stability. Confidence analysis highlights a decoupling between model certainty and accuracy, underscoring the need for adaptive trust and calibration strategies. Building on these insights, we propose a Trust-Guided Agentic AI architecture featuring self-consistency filtering, dynamic trust updating, and Chain-of-Thought prompting to further improve reliability in safety-critical AI systems.

Keywords

Medical Chatbots, Large Language Models (LLMs), Natural Language Processing (NLP), Agentic AI, Risk Assessment, Generative AI Risk Analysis, Trustworthy AI in Healthcare

Team Name

IMNTPU

Subtasks

Japanese subtask (JA, EN, FR, Multi)
German subtask (DE, EN, FR, Multi)

1 Introduction

As artificial intelligence (AI) and natural language processing (NLP) continue to advance, medical chatbots have emerged as valuable tools to support healthcare professionals by offering preliminary medical advice and reducing their workload. [17]. However, concerns regarding their safety, accuracy and compliance with medical, legal, and ethical standards remain critical. [6, 8, 11]. The NTCIR-18 MedNLP-CHAT task aims to address these concerns by evaluating chatbot responses across Japanese and German datasets [5].

While previous research in medical NLP has extensively explored areas such as clinical documentation, patient communication and clinical research [10], there has been little focus on classifying risks associated with multilingual chatbot reply. This study seeks to fill this gap by developing a robust classification approach capable of evaluating chatbot interactions across multiple languages and legal systems.

In this study, we used fine-tuned Transformer-based models, generative AI prompt engineering, and ensemble learning strategies to train, evaluate, and classify chatbot responses. Transformer-based models were fine-tuned on domain-specific datasets, while generative AI models (GPT-4o, Claude 3.5, Gemini 1.5, and Mistral Small Latest) were evaluated through prompt engineering. Additionally, an Agentic AI ensemble approach, which incorporates majority voting and weighted scoring mechanisms, was introduced to enhance the performance of risk classification.

This study contributes significantly to ensuring the trustworthy deployment of AI in healthcare, providing medical professionals, policymakers, and AI developers with practical methodologies for safe and ethical chatbot implementations.

The remainder of this paper is structured as follows: Section 2 provides a review of related work, then Section 3 describes our proposed methodology. Section 4 presents results and performance analysis across different languages and risk categories. Section 5 concludes with key findings and future directions for improving medical chatbot risk assessment.

2 Related Work

In this section, we review key methodologies relevant to medical chatbot risk assessment. We discuss large language models (LLMs) and generative AI models. Then we explore fine-tuning pretrained models and prompt engineering for optimizing AI responses. Finally, we examine agentic AI, focusing on techniques like majority voting and weighted scoring to improve risk classification.

2.1 Large Language Models (LLMs)

Recent studies have demonstrated that Large Language Models (LLMs) have significantly advanced natural language processing (NLP) by scaling transformer architectures on extensive datasets

[20]. These advancements have facilitated the deployment of LLMs in safety-critical domains such as medical chatbots, where contextual understanding and accurate risk identification are essential. It has been suggested that increasing model size enhances performance across various tasks, leading to emergent capabilities such as in-context learning and zero-shot generalization. Unlike earlier neural language models, LLMs, such as GPT [14, 19] or LLaMA [2], utilize self-attention mechanisms and large-scale training to achieve superior language understanding and generation. Research has also focused on optimizing LLMs through pre-training, fine-tuning, and scaling strategies [21].

2.2 Generative AI Models

Generative AI has transformed various fields, such as the medical field, by enabling content generation, decision support, and human-AI collaboration [9]. To address risk-sensitive decision-making in healthcare NLP, recent studies have examined a range of state-of-the-art generative AI models characterized by their multimodal capabilities, efficiency, and adaptability. Prominent examples of these models include LLaMA 3.2 3B, GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Flash, and Mistral Small Latest. Each model offers unique advantages for NLP applications relevant to medical chatbot development. LLaMA 3.2 3B, a lightweight variant of Meta’s LLaMA 3 series, is designed for multilingual commercial and research applications [2]. GPT-4o, an advanced iteration of OpenAI’s model, features end-to-end multimodal capabilities, improving fluency, contextual awareness, and cost-efficiency [19]. Claude 3.5 Sonnet demonstrates strong performance in reasoning and data processing, contributing to improved chatbot reliability [4]. Gemini 1.5 Flash, optimized for high efficiency and multimodal integration, enhances AI agent performance in real-time applications [18]. Lastly, Mistral Small Latest focuses on lightweight, high-performance NLP and is designed for low-latency, high-accuracy tasks. It excels in local deployment, function execution, and fine-tuning for domain-specific applications [3].

2.3 Fine-tuning Pretrained Models

Fine-tuning pretrained models enables large language models (LLMs) to adapt to specific tasks by updating their parameters, but it is computationally expensive and resource-intensive. Low-Rank Adaptation (LoRA) addresses this issue by freezing pretrained weights and injecting trainable low-rank matrices, significantly reducing the number of trainable parameters while maintaining performance. Studies show that LoRA achieves results comparable to full fine-tuning in tasks like text generation and summarization, with lower computational costs and no additional inference latency. Its efficiency makes it well-suited for scalable, domain-specific NLP applications, particularly in resource-constrained settings [12]. In addition to parameter-based adaptation, recent work also explores non-parametric optimization techniques such as prompt engineering to elicit better responses from frozen models.

2.4 Prompt Engineering for LLMs

Prompt engineering is the practice of designing and refining inputs to guide LLMs toward producing desired outputs. It plays a crucial role in optimizing model performance across various applications,

including question-answering, summarization and reasoning tasks. A well-crafted prompt helps improve accuracy, coherence, and contextual relevance by providing clear instructions and structured input formats [15].

Among the key techniques in prompt engineering, Few-shot learning enhances performance by providing task demonstrations, enabling LLMs to generalize without extensive fine-tuning. This approach improves performance in tasks like classification and translation, allowing efficient adaptation to domain-specific applications [16].

2.5 Agentic AI for Consensus-Based Decision Making

Agentic AI refers to autonomous systems capable of pursuing complex goals with minimal human intervention. Unlike traditional AI, which relies on structured instructions, Agentic AI adapts dynamically to evolving environments through advanced decision-making and resource management. These systems play a crucial role in domains such as healthcare, finance, and customer service [1].

Multi-agent AI systems leverage collaborative decision-making to enhance evaluation accuracy and reliability. One approach is majority voting, where multiple AI agents independently assess the same input, and the most common response is selected to reduce bias. Another approach, scoring, calculates the final output by averaging the scores from multiple AI agents, ensuring impartiality and consistency [7].

3 Methods

3.1 Task Overview and System Design

3.1.1 Task Overview. The NTCIR-18 MedNLP-CHAT task focuses on the comprehensive evaluation of medical chatbot responses across three critical dimensions: medical, legal, and ethical perspectives. The task utilized Japanese and German datasets, where each question-answer (QA) pair was systematically annotated with binary indicators for Medical Risk (MR), Ethical Risk (ER), and Legal Risk (LR). Additionally, the Japanese dataset incorporated subjective metrics (fluency, helpfulness, and harmlessness) evaluated by non-expert annotators.

The datasets were initially developed in accordance with Japanese and German medical and legal standards, and were subsequently translated into English and French by professional translators. Detailed information on dataset construction and task parameters is provided in the NTCIR-18 MedNLP-CHAT Task Overview Paper [5].

3.1.2 Submission Systems. We submitted three systems for all language tracks in both the Japanese and German subtasks. These systems were developed through an integrated approach combining prompt engineering on state-of-the-art pretrained models, fine-tuned architectures, and advanced agentic AI methodologies.

The systems are characterized as follows:

- **System 1:** Implemented optimal LLM selection based on training dataset accuracy metrics, incorporating three-shot prompting techniques.
- **System 2:** Developed an Agentic AI Majority Voting framework, wherein multiple LLMs conduct independent QA pair

evaluations, with final decisions determined through consensus.

- **System 3:** Utilized Agentic AI Weighted Scoring, where models with better training performance were given higher weights in final risk classification.

For Multi-language track, we employed a differentiated approach:

- **System 1 (Zero-shot Fine-tuned Model):** Utilized a directly fine-tuned LLaMA 3.2 3B model without providing examples.
- **System 2 (Three-shot Fine-tuned Model - English Only):** Implemented three-shot prompting exclusively with English examples.
- **System 3 (Three-shot Fine-tuned Model - Multilingual):** Deployed three-shot prompting incorporating multilingual examples (EN, JA, FR).

3.2 Model Development

3.2.1 Fine-tuning Pretrained Models. While large-scale LLMs demonstrate remarkable capabilities in general-purpose tasks, their efficacy in specialized domains such as medical risk assessment warrants further investigation. To address this uncertainty, we conducted a systematic evaluation of fine-tuned smaller-scale language models. Specifically, we fine-tuned the LLaMA 3.2 3B model using the MedNLP-2 Japanese and German datasets, subsequently assessing its performance under various prompting conditions, including zero-shot and three-shot configurations. Our primary objective was to investigate whether a fine-tuned smaller model, augmented with few-shot prompting strategies, could achieve comparable or superior performance relative to larger state-of-the-art models.

3.2.2 Prompt Engineering. Our experimental framework encompassed comprehensive evaluations of state-of-the-art LLMs, including GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Flash, and Mistral Small Latest. We conducted systematic analyses across prompting configurations ranging from zero-shot to five-shot scenarios to determine optimal performance parameters. Through rigorous evaluation of the training set, we established that three-shot prompting consistently yielded superior performance metrics across all tested configurations (Figure 1). To ensure experimental consistency, prompt examples were randomly selected from the dataset. Furthermore, to facilitate multilingual compatibility, all prompts were systematically translated into target languages (JA, FR, DE) utilizing GPT-4o's translation capabilities.

3.3 Agentic AI

Inspired by recent advances in multi-agent LLM collaboration [13], we hypothesized that aggregating multiple model judgments through Agentic AI techniques could enhance risk classification stability and performance compared to individual few-shot prompting. Therefore, two agentic systems — majority voting and trust-weighted voting — were designed to test this hypothesis.

3.3.1 Agentic AI - Majority Voting. To enhance the robustness of our evaluation framework, we developed a collaborative multi-model architecture that transcends single-model limitations. This system orchestrates four independent LLMs to conduct parallel evaluations of patient-doctor conversations, maintaining evaluation independence to prevent cross-influence. The final decision is

Figure 1: Prompt for LLMs (Few-shot)

```

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:
1. Medical Risk:
TRUE: The response provides incorrect, harmful, or
misleading medical advice that could negatively
impact the patient's health.
FALSE: The response adheres to established medical
guidelines and does not pose any harm.
2. Ethical Risk:
TRUE: The response violates ethical principles such
as patient confidentiality, informed consent,
or respect for autonomy.
FALSE: The response is ethically sound and respects
medical ethics.
3. Legal Risk:
TRUE: A label (TRUE or FALSE) that objectively
evaluates the presence or absence of legal risk
in a response, where TRUE means that the
response contains legally incorrect information
.
FALSE: The response complies with the legal
standards of the relevant region (Japanese or
German system).

Please output 1 if the chatbot's response has a risk
and 0 if the chatbot's response has no risk.

Example: {Could have multiple Examples}
"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}

```

derived through a majority voting mechanism, where individual model votes are aggregated to determine the consensus outcome. To address potential voting ties, we implemented a resolution protocol based on model performance metrics from the training dataset, systematically excluding the lowest-performing model's vote. This

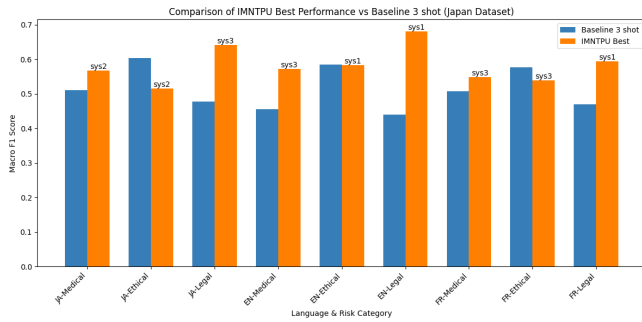


Figure 2: Macro F1 Comparison Between IMNTPU Best System and 3 Shot BASELINE in Japan Dataset.

methodology effectively simulates a panel-based expert review process, mitigating individual model biases while enhancing decision reliability.

3.3.2 Agentic AI - Weighted Scoring. Through comprehensive analysis of the training dataset, we identified distinct performance variations across models for different task types. To capitalize on these performance differentials, we implemented a sophisticated weighted scoring system that assigns proportional influence based on demonstrated model accuracy in the development set. Each LLM generates a confidence score within a normalized range of 0 (indicating no risk) to 1 (indicating high risk probability). The final risk assessment is computed through a weighted aggregation of individual model scores, with weights derived from each model’s historical accuracy metrics on comparable tasks.

To ensure scoring consistency across the model ensemble, we implemented a standardized three-shot prompting protocol. This advanced methodology enables more nuanced risk assessment compared to binary classification approaches, particularly in scenarios involving complex risk evaluation where traditional majority voting systems may prove insufficient.

4 Experimental Results and Discussion

4.1 Overall Performances

Our research team submitted 24 systems across multiple languages for both Japanese and German subtasks, the largest number among all teams. Performance evaluation was conducted using four fundamental metrics: Accuracy, Macro F1, Precision, and Recall. The comprehensive results for the Japanese subtask are presented in Table 4, while German subtask outcomes are documented in Table 5. Supplementary subjective evaluation metrics for the Japanese subtask are detailed in Table 6. To facilitate comparative analysis, Figures 2 and 3 illustrate the Macro F1 performance differential between our optimal system configuration and the official baseline (few-shot GPT-4o) across both Japanese and German datasets, while Figures 4 and 5 present corresponding accuracy comparisons.

4.1.1 Comparison with Baseline. Comparative analysis of Figures 2 and 3 demonstrates that our best-performing system achieved superior Macro F1 scores compared to baseline models across most categories, with two notable exceptions: Ethical Risk assessment in Japanese (JA) and French (FR) datasets within the Japanese subtask,

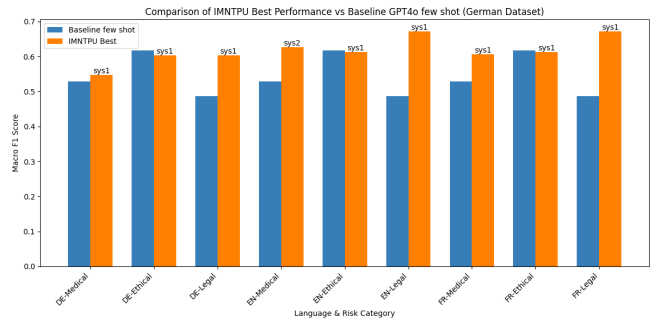


Figure 3: Macro F1 Comparison Between IMNTPU Best System and Few Shot BASELINE in German Dataset.

and Ethical Risk evaluation in the German (DE) subtask. Further examination of Figures 4 and 5 reveals consistent accuracy improvements over the baseline across the Japanese dataset. In the German dataset, while our system demonstrated higher accuracy in Medical and Legal Risk categories, it showed marginally lower performance in Ethical Risk assessment.

These empirical results suggest that our integrated approach of prompt engineering and agentic AI techniques significantly enhanced model performance in Medical and Legal Risk tasks. However, the Ethical Risk category presented persistent challenges, with our models showing lower Macro F1 scores compared to the baseline, despite achieving modest accuracy improvements (+0.02 on average) in the Japanese dataset. Conversely, in the German dataset, our system’s accuracy in Ethical Risk assessment was approximately 0.01 lower than the baseline.

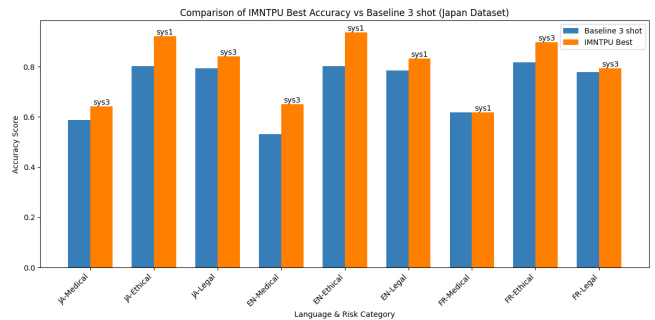


Figure 4: Accuracy Comparison Between IMNTPU Best System and 3 Shot BASELINE in Japan Dataset.

4.1.2 Rationale for Baseline Selection. For performance comparison, we selected the 3-shot prompting configuration as the baseline, aligning with our own system architecture. While the official baseline aggregates results across multiple prompting strategies, our use of a fixed 3-shot configuration ensures a controlled and fair comparison under consistent experimental conditions. This design choice reflects the constraints faced by submitted systems, which are evaluated based on a single-shot configuration, and supports a rigorous one-to-one evaluation of model capabilities. By mirroring

our system’s setup in the baseline, we minimize potential confounding variables and better isolate the contribution of our Agentic AI strategies.

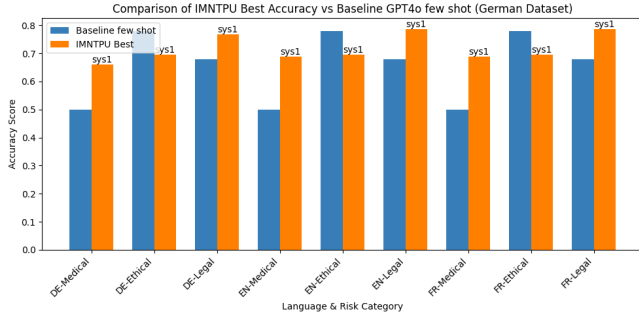


Figure 5: Accuracy Comparison Between IMNTPU Best System and Few Shot BASELINE in German Dataset.

4.2 Performance Analysis of Agentic AI and Fine-tuned Models

Our analysis focuses on the comparative effectiveness of Agentic AI methodologies and fine-tuned small models across different risk categories and languages, providing insights into both the capabilities and limitations of these approaches. Our findings reveal that while Agentic AI demonstrates promising benefits under optimal weight calibration conditions, the effectiveness of fine-tuned small models remains constrained by current data limitations. The following detailed analysis examines these outcomes and suggests future research directions for enhancing weight adjustment methodologies and expanding dataset resources to maximize the potential of both approaches.

4.2.1 Impact of Agentic AI on Different Risks and Languages. Using System 1 as a baseline, we observed varying performance shifts under Agentic AI across languages and risk types (Tables 1 & 2). Japanese systems showed the most consistent improvements, especially in Medical and Ethical Risks, where both System 2 and System 3 outperformed System 1. This suggests that high-quality native-language annotations and prompt alignment significantly enhance the effectiveness of Agentic AI. In contrast, English and French systems showed more variable results, while German models achieved minimal or negative gains. Despite likely high annotation quality, German’s complex syntax may have hindered prompt interpretation and reduced the benefits of aggregation.

Risk category also influenced system performance. Medical and Legal Risks showed relatively minor improvements, likely because their structured, rule-based nature suits LLMs without the need for aggregation. In these tasks, additional voting or weighting may introduce complexity without clear gains. In contrast, Ethical Risk exhibited the largest shifts, reflecting the greater impact of Agentic AI in subjective, context-dependent settings where model confidence and aggregation play a more central role.

Overall, Agentic AI performs best in tasks with higher ambiguity and where dataset quality supports confident reasoning. Structured domains show limited benefit, while subjective tasks gain more

from adaptive aggregation. Future work should investigate dynamic calibration strategies to further optimize performance across both linguistic and task-specific dimensions.

Table 1: IMNTPU System Performance Comparison on Japanese Dataset

Risk & Lang	S1 (Base)	S2	S3
Medical-EN	0.545	(-0.049)	(+0.027)
Medical-FR	0.548	(-0.023)	(-0.048)
Medical-JA	0.502	(+0.065)	(+0.041)
Legal-EN	0.681	(-0.168)	(-0.117)
Legal-FR	0.594	(-0.042)	(-0.059)
Legal-JA	0.564	(+0.006)	(+0.078)
Ethical-EN	0.583	(-0.011)	(-0.011)
Ethical-FR	0.516	(-0.005)	(+0.023)
Ethical-JA	0.479	(+0.037)	(-0.013)

Table 2: IMNTPU System Performance Comparison on German Dataset

Risk & Lang	S1 (Base)	S2	S3
Medical-EN	0.597	(+0.029)	(-0.085)
Medical-DE	0.548	(-0.068)	(-0.030)
Medical-FR	0.606	(-0.106)	(-0.094)
Legal-EN	0.672	(-0.055)	(-0.091)
Legal-DE	0.604	(0.000)	(-0.023)
Legal-FR	0.672	(-0.075)	(-0.045)
Ethical-EN	0.613	(-0.042)	(-0.042)
Ethical-DE	0.604	(-0.119)	(-0.050)
Ethical-FR	0.613	(-0.059)	(-0.042)

4.2.2 Evaluation of Fine-tuned Small Models vs. Large LLMs. To evaluate the competitive potential of fine-tuned small models (LLaMA 3.2 3B) against large LLMs (GPT-4o few-shot methods), we conducted a comprehensive comparison across multiple system configurations, as detailed in Table 3. Our analysis encompassed zero-shot baseline performance, three-shot baseline implementation using GPT-4o, our optimized System 1 with enhanced GPT-4o prompting, and multi-version fine-tuned IMNTPU models based on LLaMA 3.2 3B.

Results indicate that while fine-tuned small models achieve competitive performance in specific domains, particularly Legal Risk assessment, they fail to consistently surpass the GPT-4o three-shot

Table 3: 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

baseline. This performance limitation can be primarily attributed to the constrained size of our training dataset, comprising only 212 samples across both subtasks, which inhibits effective model generalization. Despite the long-term potential of fine-tuning approaches, our findings suggest that few-shot prompting with advanced LLMs currently provides superior performance, particularly given GPT-4o’s robust generalization capabilities across risk categories and languages.

4.2.3 Decision Dynamics Shift in Agentic AI Systems. To better understand the tradeoffs observed in our results, we analyzed how agentic aggregation altered risk prediction behavior. Mostly, both System 2 (majority voting) and System 3 (trust-weighted voting) demonstrated increased precision but decreased recall. This indicates a shift toward conservative classification. The ensemble systems became more hesitant to classify borderline cases as risky, potentially due to uncertainty dilution introduced by weaker LLMs such as Mistral Small.

In System 2, all models were treated as equally reliable, resulting in noisy decisions influenced by less accurate contributors. In contrast, System 3 incorporated prior performance-based trust scores, selectively amplifying the influence of more reliable models. This yielded a consistent performance improvement over System 2, reinforcing the importance of model-aware weighting in multi-agent setups.

Additionally, we observed language-specific inconsistencies in precision and recall. For example, German outputs showed simultaneous drops in both metrics, likely due to certain LLMs exhibiting hallucination or random guessing behaviors when uncertain. These findings emphasize the need to evaluate agentic systems not only on average metrics but also on behavioral shifts under uncertainty.

4.3 Confidence Analysis of Agentic AI system

Beyond raw accuracy, understanding the confidence dynamics of agentic systems provides valuable insights into how aggregation

mechanisms affect internal decision certainty. In high-stakes domains like medical risk assessment, stable and well-calibrated confidence is as crucial as prediction correctness. A model that fluctuates in its certainty—even when accurate—may lead to overcautious or erratic decision-making. This section analyzes how our Agentic AI systems influence prediction confidence across different tasks and languages, and how confidence levels correlate with actual performance.

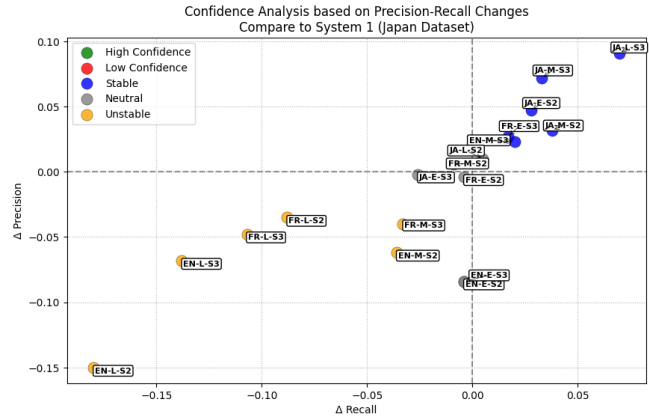


Figure 6: Confidence Analysis Based on Precision-Recall Changes Compare to System 1 on Japan Dataset.

4.3.1 Variability Across Risk Categories. Analysis of Figures 6 and 8 reveals distinct confidence patterns across risk categories. Ethical Risk assessments exhibit relatively stable confidence shifts, while Medical and Legal Risks show greater fluctuations—even within the same language-risk pair.

This likely reflects task structure. Medical and Legal Risks rely on strict domain rules, making them more sensitive to changes in

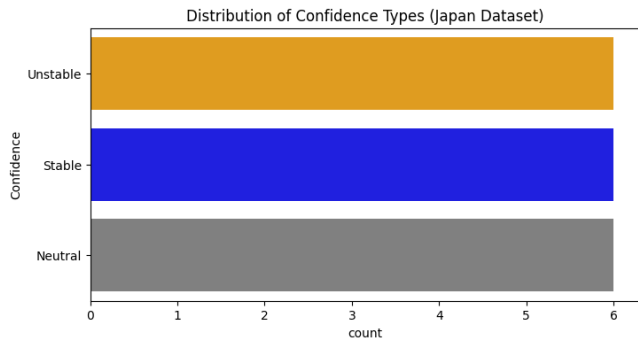


Figure 7: Distribution of Confidence Types on Japan Dataset.

prompt interpretation and model weighting—especially in agentic systems, where aggregation can amplify minor inconsistencies. In contrast, Ethical Risk is inherently ambiguous; slight shifts in output rarely alter the final classification. This suggests Agentic AI performs more reliably when decision boundaries are flexible and multiple interpretations can be reconciled.

Our neutral confidence threshold (± 0.1) avoids overinterpreting minor fluctuations. Still, Figures 7 and 9 show that Medical and Legal tasks more frequently exhibit larger confidence shifts, reinforcing the need for finer-grained calibration. Notably, System 3 shows fewer extreme shifts than System 2, indicating that trust-weighted aggregation helps reduce volatility in more rigid contexts.

In sum, Agentic AI is well-suited to ambiguous tasks like Ethical Risk but may require more refined weighting strategies to maintain stability in rule-bound domains like Medical and Legal Risk.

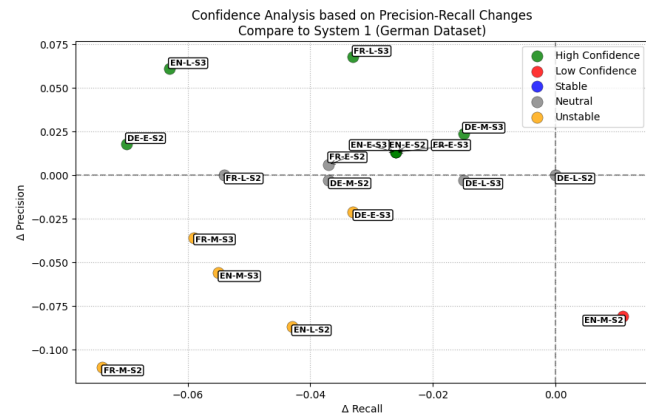


Figure 8: Confidence Analysis Based on Precision-Recall Changes Compare to System 1 on German Dataset.

4.3.2 *Language-Specific Confidence Shifts.* A notable finding from Figures 6 and 8 is that Japanese and German systems consistently exhibited the most stable confidence levels across their respective subtasks. This supports our hypothesis that high-quality native-language annotations enhance model stability under the Agentic

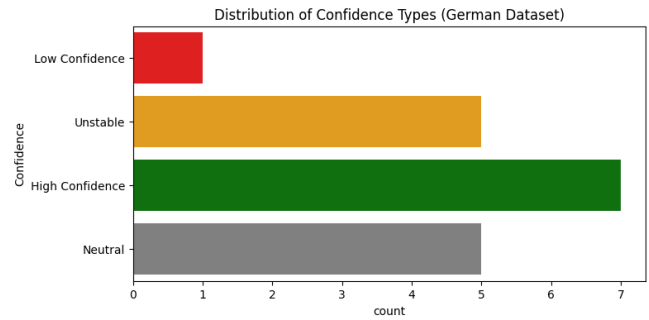


Figure 9: Distribution of Confidence Types on German Dataset.

AI framework. In the Japanese dataset, four systems improved in confidence while the others remained neutral; similarly, all but one German system also showed improvement or stability—reinforcing the role of high-quality annotations in confidence calibration.

However, these gains in confidence did not always translate into accuracy—particularly for German systems. As discussed in Section 4.2.2, German’s complex syntax likely hindered prompt interpretation, limiting performance gains despite stable confidence. This highlights that while strong annotations promote internal consistency, linguistic complexity may still challenge model reasoning.

By contrast, English and French systems showed more volatile confidence levels, suggesting that Agentic AI may be less effective when annotations are translated rather than native. Notably, System 3 consistently exhibited narrower confidence variation than System 2 in these languages, indicating that trust-weighted aggregation mitigates volatility even in lower-quality datasets. These findings underscore the interplay between annotation quality, linguistic structure, and aggregation strategy in shaping both confidence and performance.

4.3.3 *Confidence vs. Accuracy: Are They Correlated?* Our analysis reveals that model confidence and prediction accuracy do not always align. For example, as shown in Figure 8, the DE-E-S3 system exhibits high confidence yet fails to outperform others in accuracy. This suggests that while Agentic AI improves internal decision consistency, it does not guarantee correctness.

A likely explanation is that our current weighting mechanism assigns fixed trust scores based on prior development accuracy. This static approach may overemphasize models that were previously reliable, even if their performance deteriorates under certain task or language conditions. As a result, the system may exhibit inflated confidence without improved outcomes—especially in borderline or ambiguous cases.

Furthermore, histograms of confidence distribution (Figures 7 and 9) show frequent confidence shifts across models and tasks. However, these fluctuations do not consistently correspond to accuracy changes, highlighting a misalignment between internal certainty and external correctness.

These findings underscore a key limitation of our current Agentic AI framework: trust-weighted aggregation may stabilize confidence but can inadvertently amplify overconfidence. Addressing this issue

will require dynamic, task-aware calibration strategies that adjust trust based on ongoing performance rather than static priors.

5 Conclusion

This study provides a large-scale evaluation of Agentic AI in multilingual risk assessment, demonstrating its potential to enhance decision consistency while highlighting its limitations across structured and subjective tasks. Our results indicate that Agentic AI significantly improves performance in subjective assessments, such as Ethical Risk, but has a more variable impact on structured tasks like Medical and Legal Risk, where predefined decision frameworks limit its effectiveness.

Language and dataset quality played a crucial role in performance variability. Japanese systems benefited from high-quality native-language annotations, leading to improved confidence calibration and accuracy, while German models, despite strong annotations, struggled with complex syntactical structures that hindered effective prompt adaptation. Fine-tuned small models showed competitive performance in legal risk assessment but did not consistently outperform GPT-4o 3-shot prompting, underscoring the need for larger, high-quality training datasets for small model optimization.

Confidence analysis revealed that higher model confidence does not always correspond to improved accuracy, particularly in Ethical Risk assessment, where over-reliance on weight-based adjustments may lead to systematic errors. This suggests the need for adaptive confidence calibration mechanisms to ensure stability without introducing false certainty.

Due to the limited scale of the NTCIR-18 dataset, particularly in multilingual subsets, our observations should be interpreted with caution. Larger-scale evaluations across diverse domains will be necessary to confirm the generalizability of our findings.

Building on these insights, future work will focus on three core extensions. First, we plan to integrate self-consistency filtering, a mechanism that excludes internally inconsistent models prior to aggregation. Second, we will implement dynamic trust updating, where model reliability scores are adjusted based on recent task-specific feedback. Third, we aim to incorporate Chain-of-Thought (CoT) prompting into each agent's reasoning process to improve inference completeness, particularly in ethical or ambiguous cases.

These enhancements have been formally proposed and explored in our follow-up work, Trust-Guided Multi-Agent Risk Classification, which directly extends the framework and design principles introduced in this NTCIR submission.

Acknowledgements

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.

References

- [1] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* (2025).

- [2] Meta AI. 2024. Llama 3.2 Lightweight Model Card. Retrieved from https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.
- [3] Mistral AI. 2025. Mistral Small 3. Retrieved from <https://mistral.ai/news/mistral-small-3>.
- [4] Anthropic. 2024. Introducing the Next Generation of Claude. Retrieved from <https://www.anthropic.com/news/claude-3-family>.
- [5] Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, Shohei Hisada, Tomohiro Nishiyama, Lenard Paulo Velasco Tamayo, Jingnan Xiao, Axalia Levenchaut, 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*.
- [6] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* 183, 6 (2023), 589–596.
- [7] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. arXiv:2308.07201 [cs.CL] <https://arxiv.org/abs/2308.07201>
- [8] Dariush D Farhud and Shaghayegh Zokaei. 2021. Ethical issues of artificial intelligence in medicine and healthcare. *Iranian journal of public health* 50, 11 (2021), i. <https://doi.org/10.18502/ijph.v50i11.7600>
- [9] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. 2024. Generative ai. *Business & Information Systems Engineering* 66, 1 (2024), 111–126.
- [10] Rachit Garg and Anshul Gupta. 2024. A Systematic Review of NLP Applications in Clinical Healthcare: Advancement and Challenges. In *Advances in Data-Driven Computing and Intelligent Systems*, Swagatam Das, Snehanshu Saha, Carlos A. Coello Coello, and Jagdish C. Bansal (Eds.). Springer Nature Singapore, Singapore, 31–44.
- [11] Rachel S Goodman, J Randall Patrinely, Cosby A Stone, Eli Zimmerman, Rebecca R Donald, Sam S Chang, Sean T Berkowitz, Avni P Finn, Eiman Jahangir, Elizabeth A Scoville, et al. 2023. Accuracy and reliability of chatbot responses to physician questions. *JAMA network open* 6, 10 (2023), e2336483–e2336483.
- [12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.
- [13] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 14165–14178. <https://doi.org/10.18653/v1/2023.acl-long.792>
- [14] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Aleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- [15] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*. Springer, 387–402.
- [16] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–7.
- [17] Prakhhar Srivastava and Nishant Singh. 2020. Automated Medical Chatbot (Medibot). In *2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC)*. 351–354. <https://doi.org/10.1109/PARC49193.2020.236624>
- [18] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] <https://arxiv.org/abs/2403.05530>
- [19] OpenAI Team. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] <https://arxiv.org/abs/2410.21276>
- [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [21] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL] <https://arxiv.org/abs/2303.18223>

A Formal Run Full Results

This appendix provides additional tables that show all the formal run results of IMNTPU.

Table 4: Performance of IMNTPU in Japan Subtask

Language	Risk	Metric	System 1*	System 2*	System 3*
EN	Ethical Risk	Accuracy	0.937	0.929	0.929
		F1-score	0.583	0.572	0.572
		Precision	0.722	0.638	0.638
		Recall	0.558	0.554	0.554
	Legal Risk	Accuracy	0.833	0.802	0.833
		F1-score	0.681	0.513	0.564
		Precision	0.671	0.521	0.603
		Recall	0.694	0.514	0.556
	Medical Risk	Accuracy	0.635	0.603	0.651
		F1-score	0.545	0.496	0.572
		Precision	0.634	0.572	0.657
		Recall	0.571	0.535	0.591
JA	Ethical Risk	Accuracy	0.921	0.865	0.873
		F1-score	0.479	0.516	0.466
		Precision	0.468	0.515	0.466
		Recall	0.492	0.520	0.466
	Legal Risk	Accuracy	0.802	0.810	0.841
		F1-score	0.564	0.570	0.642
		Precision	0.570	0.580	0.661
		Recall	0.560	0.565	0.630
	Medical Risk	Accuracy	0.611	0.635	0.643
		F1-score	0.502	0.567	0.543
		Precision	0.590	0.622	0.662
		Recall	0.542	0.580	0.575
FR	Ethical Risk	Accuracy	0.865	0.857	0.897
		F1-score	0.516	0.511	0.539
		Precision	0.515	0.511	0.542
		Recall	0.520	0.516	0.537
	Legal Risk	Accuracy	0.738	0.786	0.794
		F1-score	0.594	0.552	0.535
		Precision	0.588	0.553	0.540
		Recall	0.639	0.551	0.532
	Medical Risk	Accuracy	0.619	0.619	0.595
		F1-score	0.548	0.525	0.500
		Precision	0.597	0.603	0.557
		Recall	0.564	0.555	0.531
Multi	Ethical Risk	Accuracy	0.460	0.563	0.595
		F1-score	0.350	0.405	0.422
		Precision	0.481	0.494	0.498
		Recall	0.421	0.476	0.493
	Legal Risk	Accuracy	0.619	0.746	0.714
		F1-score	0.434	0.559	0.576
		Precision	0.458	0.555	0.575
		Recall	0.431	0.574	0.625
	Medical Risk	Accuracy	0.508	0.524	0.460
		F1-score	0.482	0.519	0.458
		Precision	0.483	0.524	0.466
		Recall	0.483	0.525	0.465

Table 5: Performance of IMNTPU in German Subtask (Objective Evaluation)

Language	Risk	Metric	System 1*	System 2*	System 3*
DE	Medical Risk	Accuracy	0.661	0.634	0.652
		F1-score	0.548	0.480	0.518
		Precision	0.731	0.728	0.755
		Recall	0.585	0.548	0.570
	Ethical Risk	Accuracy	0.696	0.643	0.670
		F1-score	0.604	0.485	0.554
		Precision	0.795	0.813	0.774
		Recall	0.626	0.556	0.593
	Legal Risk	Accuracy	0.768	0.768	0.759
		F1-score	0.604	0.604	0.581
		Precision	0.876	0.876	0.873
		Recall	0.606	0.606	0.591
EN	Medical Risk	Accuracy	0.688	0.679	0.643
		F1-score	0.597	0.626	0.512
		Precision	0.760	0.679	0.704
		Recall	0.618	0.629	0.563
	Ethical Risk	Accuracy	0.696	0.679	0.679
		F1-score	0.613	0.571	0.571
		Precision	0.768	0.781	0.781
		Recall	0.630	0.604	0.604
	Legal Risk	Accuracy	0.786	0.750	0.759
		F1-score	0.672	0.617	0.581
		Precision	0.812	0.725	0.873
		Recall	0.654	0.611	0.591
FR	Medical Risk	Accuracy	0.688	0.625	0.643
		F1-score	0.606	0.500	0.512
		Precision	0.740	0.630	0.704
		Recall	0.622	0.548	0.563
	Ethical Risk	Accuracy	0.696	0.670	0.679
		F1-score	0.613	0.554	0.571
		Precision	0.768	0.774	0.781
		Recall	0.630	0.593	0.604
	Legal Risk	Accuracy	0.786	0.759	0.777
		F1-score	0.672	0.597	0.627
		Precision	0.812	0.812	0.880
		Recall	0.654	0.600	0.621
Multi	Medical Risk	Accuracy	0.500	0.464	0.500
		F1-score	0.492	0.460	0.499
		Precision	0.495	0.466	0.512
		Recall	0.495	0.465	0.513
	Ethical Risk	Accuracy	0.500	0.536	0.527
		F1-score	0.487	0.526	0.502
		Precision	0.488	0.527	0.502
		Recall	0.487	0.528	0.502
	Legal Risk	Accuracy	0.607	0.661	0.652
		F1-score	0.499	0.511	0.551
		Precision	0.501	0.534	0.559
		Recall	0.501	0.521	0.550

Table 6: Subjective Evaluation Results for IMNTPU (Japan Subtask)

Metric	System 1*	System 2*	System 3*
Fluency	0.026	0.026	0.025
Harmlessness	0.017	0.021	0.020
Helpfulness	0.026	0.028	0.028