# UPxSocio at NTCIR-18 MedNLP-CHAT Task: Similarity-Based Few-Shot Example Selection for Prompt-Based Detection

Michael Van Supranes[1,2], Martin Augustine Borlongan[1], Joseph Ryan Lansangan[1], Genelyn Ma. Sarte[1], Shaowen Peng[2], Shoko Wakamiya[2], Eiji Aramaki[2]

[1]Univeristy of the Philippines Diliman [2]Nara Institute of Science and Technology

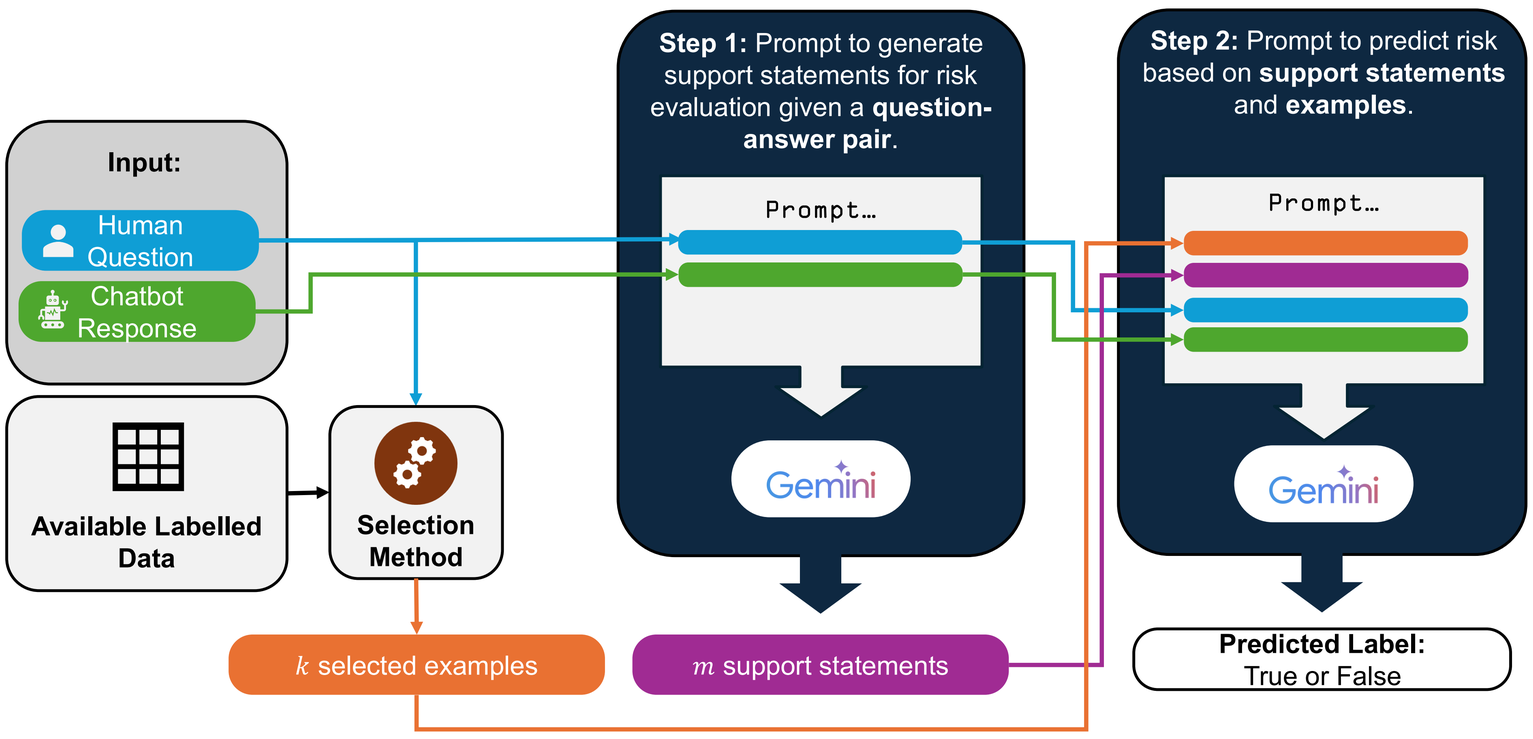
## INTRODUCTION

We present a prompt-based system for detecting medical, ethical, and legal risks in chatbot-generated responses, as part of the MedNLP-CHAT Task at NTCIR-18. **Our two-step method, using Gemini-1.5-flash, first generates support statements to guide reasoning, then integrates them into a few-shot classification prompt**. We submitted systems for English versions of the Japanese and German subtasks, exploring variations in example selection and label distribution. Our results show relatively stronger performance on medical risk detection, while ethical and legal risks remain challenging. Ablation studies across 24 prompt variants reveal trade-offs between recall and precision, influenced by example similarity of selected examples. In conclusion, a well-optimized prompt design can be a good starting point for developing a risk detection system without large-scale model training.

**Keywords:** In-context Learning, Few-shot Example Selection, Prompt Engineering

**Subtask:** Japanese subtask (EN), German subtask (EN)

## METHOD: TWO-STEP FEW-SHOT PROMPTING




### Elements of Prompt Design:

**Number of Support Statements** ( $m$ ) to be generated in the first step.

**Method of Example Selection: $k$-Nearest** selects a set of examples that are semantically similar to the test case. **$k$-Spread** maximizes the variability of examples included. BERTScore (Zhang et al., 2019) was used to measure semantic similarity.

**Number of Few-shot Examples** $(k)$ included in the second step.

**Distribution of Examples: Balanced**, equal number of risk and non-risk cases; **Skewed**, majority of examples are risk cases.

## ANALYSIS OF PROMPT DESIGN

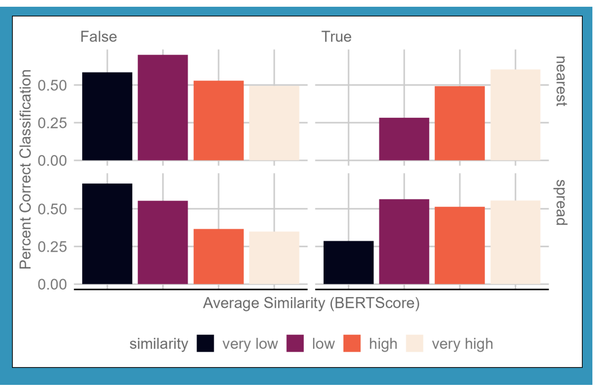
**Different variations of the prompt were tested on the available training data for the medical risk category.** Results were then analyzed to see which elements of the prompt are important. Logistic regression and Chi-square Automatic Interaction Detection were used to identify significant associations with correct detection. Below is a summary of insights:

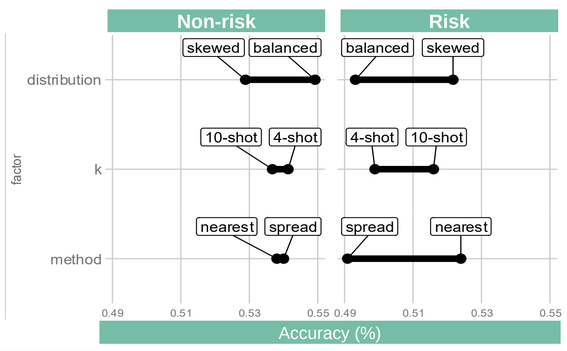
**#1 more similar examples imply better recall of risk.**

Higher similarity is associated with higher correct classification of risk-positive

However, it may increase false positive rate. Using $k$-nearest tend to have a lower false positive rate.

Semantic similarity is a useful metric for example selection.




**#2 the effect of distribution, number, and similarity of examples on accuracy tend to be more evident when detecting risk.** However, they are not as impactful when detecting non-risk cases.

**#3 $k$-Nearest strategy performs better with more examples in the prompt.**

| Factors | | Average Macro F1 | |
|---|---|---|---|
| Method | $k$ | German | Japanese |
| $k$-nearest | 10 | **0.5717** | **0.4854** |
| | 4 | 0.5694 | 0.4613 |

| Average Macro F1 (Both subtasks) | | |
|---|---|---|
| 1-step | 2-step ($m = 5$) | 2-step ($m = 10$) |
| 0.5072 | **0.5286** | 0.5099 |

**#4 A 2-step prompt structure does not have a clear advantage over a single prompt.** Generating 5 support statements is better than 10.

## SYSTEMS AND RESULTS

| Subtask - System | | Few-shot Prompt Design: Method | Distribution | $k$ | $m$ | Macro F1-score: Medical | Ethical | Legal |
|---|---|---|---|---|---|---|---|---|
| Japanese (EN) | 1 | $k$-nearest | Balanced | 10 | 10 | **0.603** | 0.426 | 0.397 |
| | 2 | $k$-spread | Balanced | 10 | 10 | 0.570 | 0.436 | 0.416 |
| German (EN) | 1 | $k$-nearest | Balanced | 10 | 10 | **0.614** | **0.678** | 0.591 |
| | 2 | $k$-spread | Balanced | 10 | 10 | 0.570 | **0.678** | 0.565 |

**Japanese Subtask:**

- ✓ Relatively high macro F1-score in medical risk detection.
- ✓ System 1 performed best in the medical risk category.
- × Both systems struggled with ethical and legal risk detection.

**German Subtask:**

- ✓ Relatively high macro F1-score in medical and ethical risk detection.
- ✓ System 1 ranked top 3 in medical risk detection.
- ✓ Both systems achieved best results for ethical risk detection.
- × Legal risk detection remained a challenge.

## CHALLENGES, AND LIMITATIONS

- **High computational cost** due to example selection requiring pairwise similarity per risk type. It is time consuming. **The method will greatly benefit from a faster algorithm for finding similar examples.**
- Due to time constraints, **only 10-shot prompts** were tested in the official run.
- Currently, the performance tends to be good for medical risk detection, but **correctly detecting all risk types is still a challenge.**
- The quality of the generated statements were not analyzed. **Better statements may be achieved if relevant facts were included in the prompt (e.g., via RAG).**
- The method were only tested using the Gemini model. **Future works may consider other models.**