Evaluation Results of UTUtLB25 Team in NTCIR-18 U4 Task of Table Question Answering of Securities Reports

Long Si[1], Yin Zhang[1], Xiaotian Wang[2], Takehito Utsuro[1] (1 : University of Tsukuba, 2 : University of Tokyo)

Abstract

Purpose : Build a system to participate in the information extraction task from tables in securities reports (NTCIR-18 U4 Task)

Structure of the NTCIR-U4 Task :

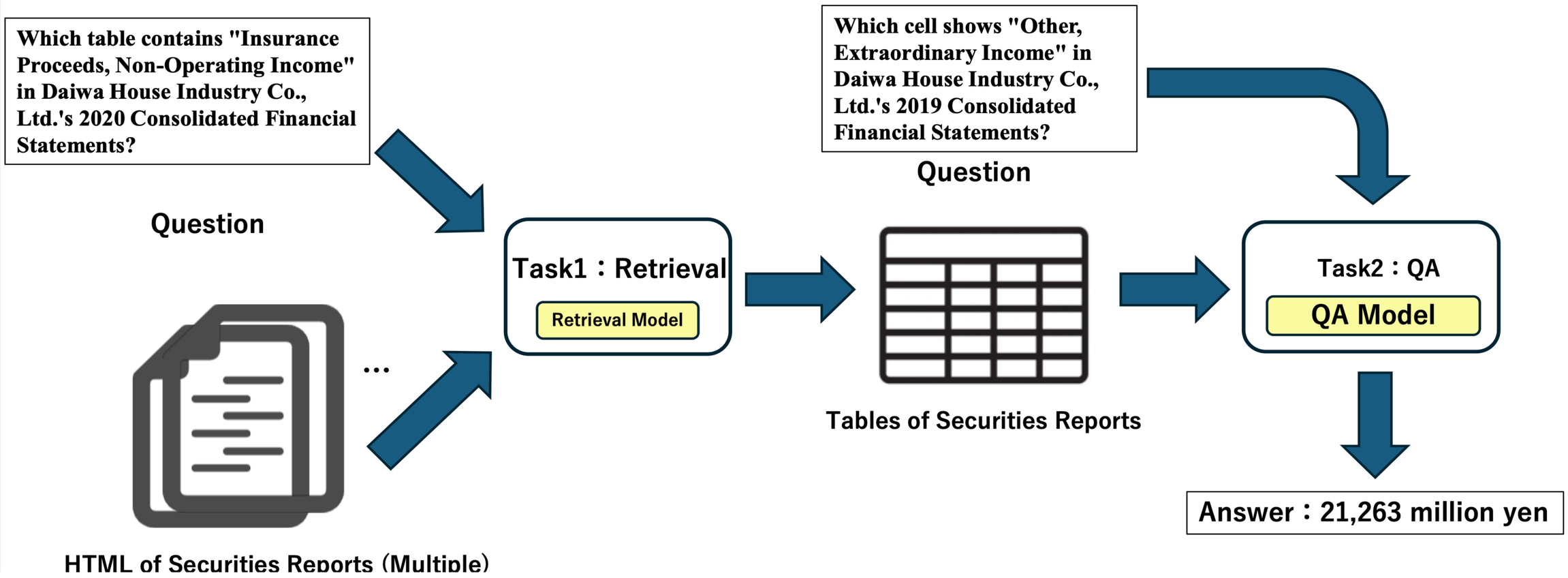
1.TR (Table Retrieval) Task : Retrieve tables containing relevant data

2.TQA (Table Question Answering) Task : Extract the answer to a given question from the retrieved table

Approach :

Apply state-of-the-art **Large Language Models (LLMs)** to the **TQA(Table Question Answering) task**

Framework of NTCIR-18 task



Input, Output, Eval Metrics

	TR Task	TQA Task
Input①	HTML of Securities Reports(Multi)	Tabular data(Table ID)
Input②	Question(question ID)	Question(question ID)
Output	Table(Table ID)	Answer(Cell ID or Value)
Eval Metrics	Accuracy	Accuracy

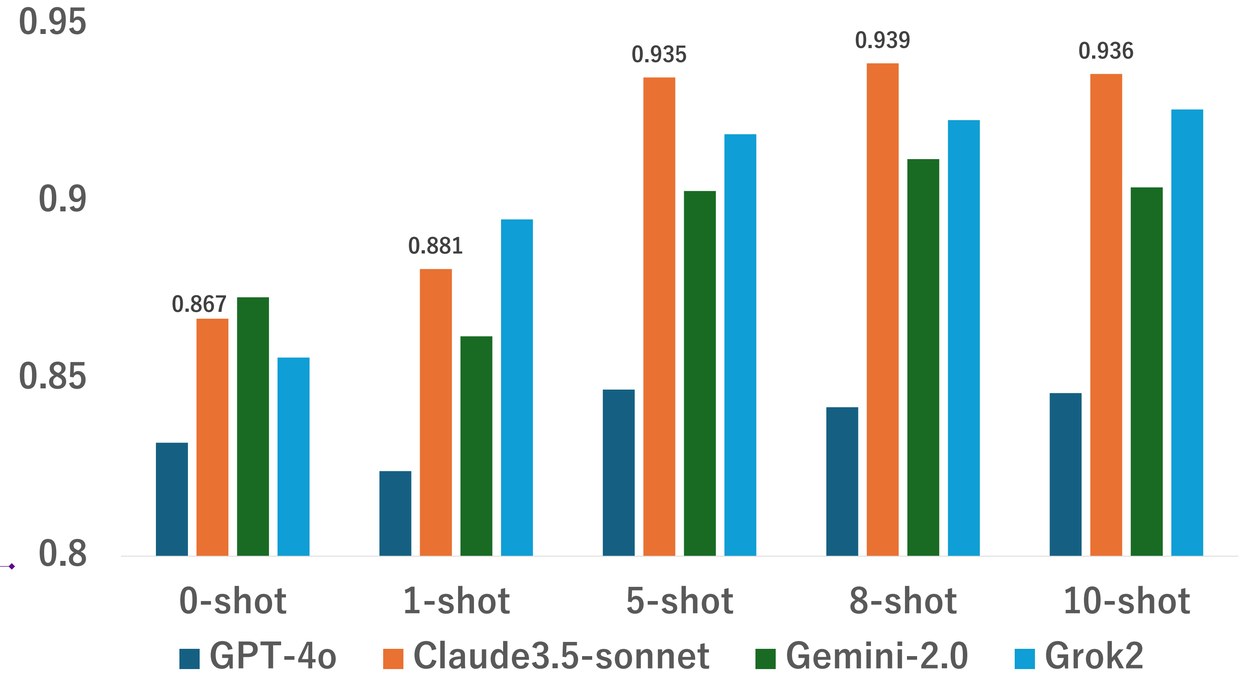
Eval Metrics of TR task :

$$\text{Accuracy} = \frac{Nums\ o\ f\ Correct\ (Table\ ID)}{Total}$$

Eval Metrics of TQA task :

$$\text{Accuracy} = \frac{Nums\ o\ f\ Correct\ (Cell\ ID\ or\ Value)}{Total}$$

Result of LLMs using zero-shot/few-shots

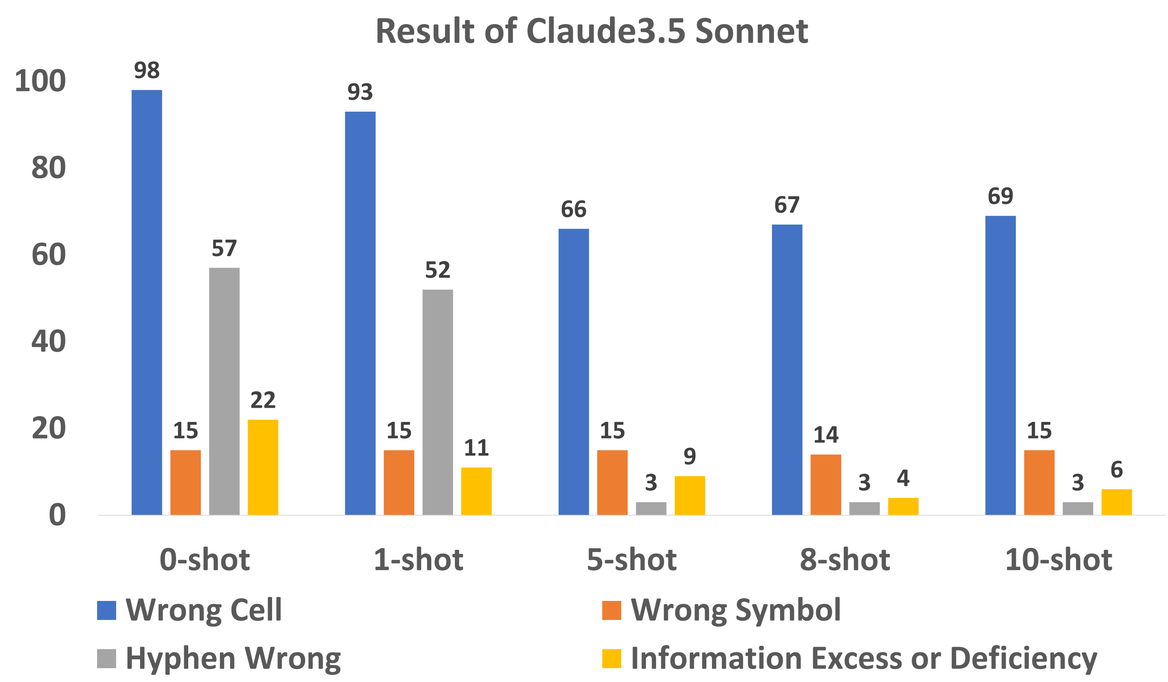


Result of TQA task using LLMs

- Evaluation of validation data
- Cell id : 1-shot Value : 8-shot

Model	Cell id Accuracy	Value Accuracy
GPT-4o	87.3%	84.2%
Claude3.5-Sonnet	93.9%	93.9%
Claude3.5-Haiku	66.9%	70.1%
Gemini2.0-Flash	83.7%	91.2%
Gemini1.5-Flash	61.6%	79.7%
Grok2	85.1%	92.3%
Grok-beta	86.5%	91.1%

Result Analysis of Incorrect Answer



Result of NTCIR-18 U4 Task(Top 5)

Public Score : Part of test data
Private Score : Full data

team	Public cell ID	Public Value	Private cell ID	Private Value
UTUtLB25	0.9309	0.9470	0.9066	**0.9317**
UTUtLB25	0,9309	0.9409	0.9066	0.9266
UTUtLB25	0.9309	0.9286	0.9066	0.9116
airev	**0.9785**	**0.9493**	**0.9473**	0.9047
airev	**0.9785**	**0.9493**	**0.9473**	0.8991

Summary

- Propose a method for **TQA task** of securities reports using **LLMs**
- Experimental results prove the **effectiveness of LLMs**