ISPOR 2024

Leveraging Large Language Models (LLMs) for Classifying Peer Reviewed Publications for Literature Review

INTUITIVE

MSR17

INTRODUCTION: Literature reviews are crucial for clinical research and evidence-based decision medicine. However, they are labor intensive and take a long time to complete. The use of Large Language Models (LLMs) might help reduce the burden, due to their groundbreaking context awareness.

AIM: This project aims to investigate the use of LLMs for classifying peer-reviewed publications. By leveraging LLMs, we aim to develop a system that can accurately extract metadata from publications and present it in a structured format.

METHODS: We utilized **OpenAI's GPT-4 model**, **Llama-index library**, and the **Pydantic library** to extract structured output from publication abstracts. The LLM model, in combination with user-defined Pydantic models, enables more consistent and reliable structured outputs (Figure 1). The program was developed and validated using Python 3.12.

Figure 1. Code example

```
class Procedure(BaseModel):
    Procedure_name: str = Field(..., description="Procedure name from the abstract. ie. Lobectomy")

class Abstract(BaseModel):
    """Data model for an abstract."""
    Procedures: List[Procedure]

program = OpenAIPydanticProgram.from_defaults(
    output_cls=Abstract,
    llm=OpenAI(temperature=0, model="gpt-4"),
    prompt_template_str=(
        "Please extract the following abstracts into a structured data according to: {input_str}."),
    verbose=True)

df_input = df.abstract[0]
response_obj = program(input_str=df_input).model_dump()
```

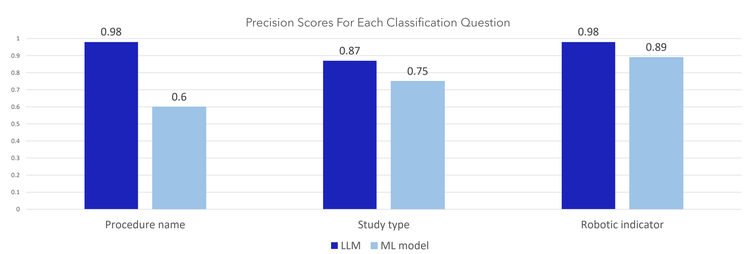
```
{'Procedures': [{'Procedure_name': 'minimally invasive radical cystectomy (MIRC)'},
  {'Procedure_name': 'open radical cystectomy (ORC)'},
  {'Procedure_name': 'robot-assisted radical cystectomy'},
  {'Procedure_name': 'laparoscopic radical cystectomy'}]}
```

COMPARISON TO MACHINE LEARNING MODELS: We compared traditional machine learning models (Logistic Regression, Decision Trees, and Gradient Boosting) to the LLM approach. These models were trained on a dataset of 35,000 labeled publications. We selected the best-performing model for each type of publication metadata and compared it to the LLM approach. To evaluate data extraction accuracy, we created three classification questions: **surgical procedure name**, **study type**, and whether **robotic surgery** was studied. These questions were tested on a dataset of 118 abstracts with different procedure and study types (Table 1). The results were evaluated by two independent researchers.

Table 1: Distribution of Procedure Name, Study Type, and Robotic-Assisted Surgery Indicator in Testing Dataset

Procedure name	N	Procedure name	N
Prostatectomy	20	HPB	7
Hysterectomy	14	Cholecystectomy	2
Cystectomy	8	Lymphadenectomy	7
Sacrocolpopexy	5	Cardiac surgery	2
Colorectal	13	Lobectomy	2
Nephrectomy	13	Esophagectomy	2
Gastrectomy	4	TORS	3
Hernia repair	1	Other	19
Study type	**N**	**Study type**	**N**
Systematic review	30	Economic modelling	10
Randomized controlled trials	20	Non-systematic literature reviews	9
Retrospective comparative studies	20	Single arm or case series studies	9
Prospective non-randomized comparative studies	10	Other	10
Robotic assisted surgery indicator	**N**	**Robotic assisted surgery indicator**	**N**
Robotic surgery was studied	93	Minimal invasive surgery only	25

Figure 2: Comparison of LLM and ML Models Evaluation



RESULTS: Three questions were assessed individually. The precision of procedure name, study type, and robotic surgery indicator was 0.98, 0.87, and 0.98, respectively. Traditional machine learning models had classification results of 0.60, 0.75, and 0.89, respectively (Figure 2). LLM showed significantly better results in procedure name and slightly better results in study type and robotic indicators, compared to traditional machine learning methods. The evaluation for LLM and the reasoning for misclassification against true answers are presented in Table 2.

Table 2: Types of Misclassifications and Their Associated Reasons

Misclassifications	Reason	N
Procedure name	Missing one or more concomitant procedures	2
Study type	Literature review misclassified as systematic review	3
Study type	Systematic review misclassified as literature review	3
Study type	Randomized control trial on another factor not surgical modality	1
Study type	Prospective data collection with retrospective analysis	3
Study type	Single arm study, no cohort comparison was performed	2
Robotic-assisted surgery indicator	Robotic surgery was mentioned, but no data specific to robotic surgery was studied	2

CONCLUSIONS: This study demonstrates that LLMs can accelerate the extraction and analysis of vast amounts of information, increasing productivity and optimizing the literature review process. Additionally, LLMs are more versatile and generalizable compared to their traditional Machine Learning counterparts. Clinical librarians can leverage these tools to shift their role from manually labeling and extracting data to validating model output.