Improving Research Question Quality with Controlled Natural Languages and Large Language Models

RQ-CNL

Context

Good research questions (RQs) are essential for guiding academic research, yet there are few tools to help formulate or improve RQs. This project explores methods for creating a Controlled Natural Language (CNL) and a RQ Improvement Model to score and enhance research question quality.

RQ Scoring & Improvement Models

Collected 125 RQs annotated by human evaluators

1. **BERT-base**, **Flan-T5**, and **Mistral-7B-Instruct** were used for RQ quality scoring for dimensions of: relevance**(REL)**, fluency**(FLU)**, feasibility**(FEA)**, and clarity**(CLA)**.
2. **Flan-T5** (small & base) and **Mistral-7B-Instruct** were fine-tuned for rewriting bad quality research questions.

Extracted RQ → Human Evaluation → Model Training

Research Question:
"Is task-specific reasoning bad?"

* RQs were extracted from ACL Anthology and UCT honours repository

RQ Quality Scores: REL=2/5, FLU=3/5, FEA=3/5, CLA=5/5

Improved version:
"Do LLMs acquire task-specific or generalizable reasoning skills during pretraining?"

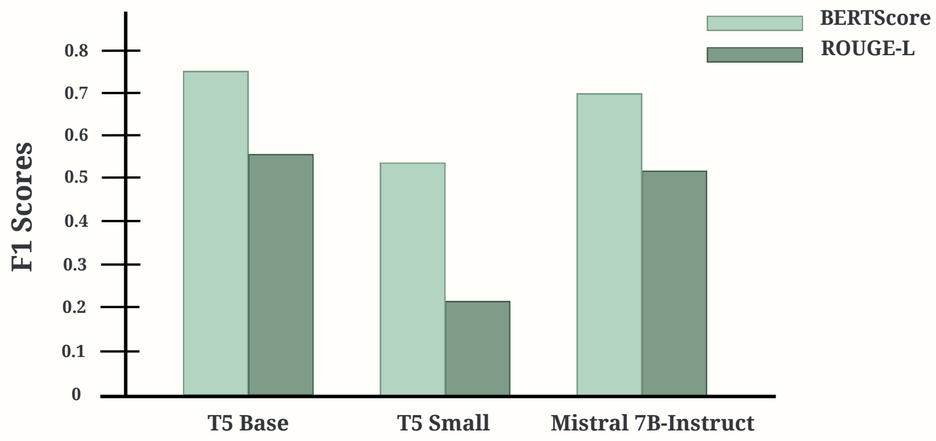
Results

Test Set Size: 25 RQs

T-test and Pearson Correlation Comparison of Score Models

Score Model	t-test p-value	Pearson Correlation (r)
Bert-base	0.148	0.194
T5-base	0.000	-0.232
Mistral	0.306	0.034

BERTScore and ROUGE-L Comparison of Improvement Models



Conclusions

- **Scoring**: BERT-base is most effective.
- Scoring models struggled with rating in a manner similar to human judgement due to data scarcity.
- **RQ Improvement**: Flan-T5-base had the highest scores where generated text was lexically and semantically similar to human improved RQs.

Controlled Natural Languages

Method 1: Implicit RQs

Llama 3.2 and **Mistral 7B**, were prompted to extract *implicit* RQs from research paper *abstracts*.

Method 2: Explicit RQs

BERT and **SciBERT**, were trained on a dataset of question sentences to identify *explicit* RQs from research papers.

Separate CNL template sets were generated by identifying key concepts and actions in the RQ sets (*explicit and implicit*) and replacing them with placeholder slots.

Research Question → CNL Template

How can we quantify bias in pre-trained language models?

How can we **PC1 EC1** in **EC2** ?

Entity Chunks (ECs): Key concepts or objects
Predicate Chunks (PCs): Actions or relations

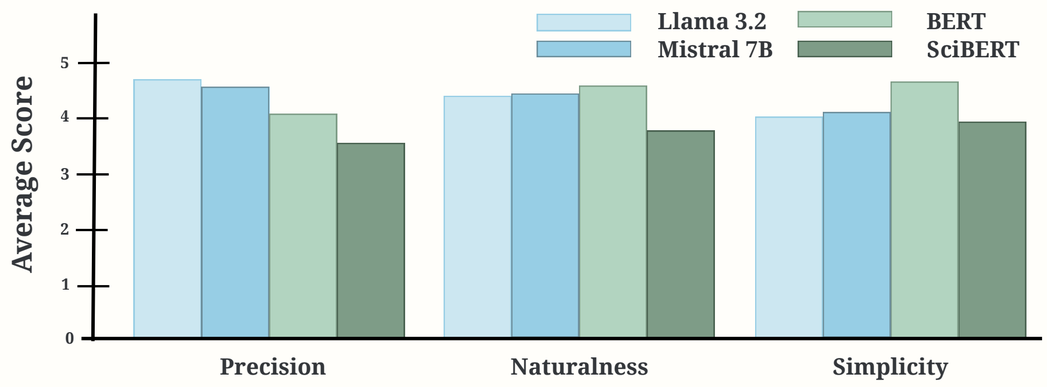
Results

BLEU Scores

(Lexical similarity between CNL templates and test set templates)

	Implicit RQs		Explicit RQs	
Metric	**Llama**	**Mistral**	**BERT**	**SciBERT**
Sentence-level	0.513	0.469	0.411	0.499
Corpus-level	0.894	0.864	0.155	0.189

CNL Quality Evaluation (PENS Framework)



Conclusions

- Explicit RQs produced templates with greater variety, while implicit RQ templates followed more uniform patterns.
- Llama 3.2 is stronger for readability and coverage, whereas Mistral 7B is better suited for precision and evaluation-focused questions.
- SciBERT's training on scientific text makes it better suited to the task of identifying RQs.

Mandikudza Dangwa
DNGMAN001@myuct.ac.za

Rector Ratsaka
RTSREC001@myuct.ac.za

Emma van der Berg
VBREMM005@myuct.ac.za

Supervisor: Zola Mahlaza zmahlaza@cs.uct.ac.za

dept@cs.uct.ac.za

Department of Computer Science