

KORonKOR: KORean model ON KORean text2sql

2022-13586 Suyeon Woo



Intro

1. Observation

- Llama (non-Korean model) performed poorly on Korean DB
- Even with RAG, accuracy was low
- Why does it fail to parse Korean queries effectively?

2. Insight

- Llama is not optimized for Korean → struggles with semantic parsing
- Text2SQL requires deep semantic understanding
- So, my hypothesis is...

“Korean-targeted models will perform better on Korean databases”



Methods: Experimental Design

1. Dataset

- [Natural Language to SQL Query Generation Data, AIHub](#)
- Realistic Database setting: English schema + Korean values with Korean query
- Selected 2 domains for Experiment
 - Culture DB: 15 tables with 250 queries
 - Education DB: 16 tables 306 queries



Methods: Experimental Design

2. Metrics

- Intent-based SQL Accuracy
 - Focuses on whether the user's intent is correctly understood
 - **SELECT**: Checks if the requested columns are included
 - **WHERE**: Checks if the same constraints as the ground truth are applied

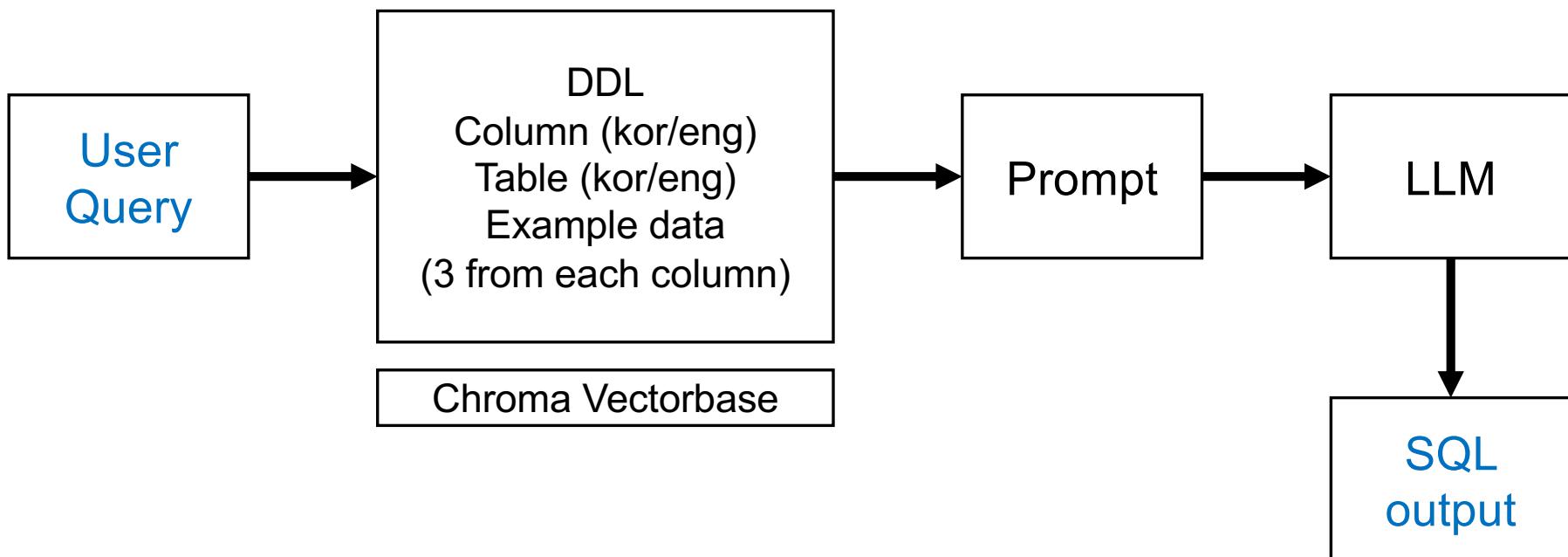
3. Models

- Llama-3.2-3B-Instruct (*baseline)
- Llama-3.2-Korean-Blossom-3B (fine-tuned on Korean data)
- **EXAONE-3.5-2.4B-Instruct (Korean-first LLM)**



Methods: Implementation

- RAG implemented using Vanna framework





Results: Experimental Results

- Overall accuracy is low
- Llama < Blossom < EXAONE : Consistent with the hypothesis

	Culture	Education
Llama-3.2-3B-Instruct	7.6%	17.6%
Ilama-3.2-Korean-Blossom-3B	9.6%	20.9%
EXAONE-3.5-2.4B-Instruct	12.8%	35.0%



Results: Experimental Results

- Hardness provided by dataset
- Consistent with the hypothesis across all Hardness levels

	Culture			Education			
	Easy	Medium	Hard	Easy	Medium	Hard	Extra Hard
Llama-3.2-3B-Instruct	8%	9.5%	5.7%	31.7%	21.0%	9.8%	22.2%
llama-3.2-Korean-Blossom-3B	14%	9.5%	7.6%	41.5%	24.2%	12.1%	11.1%
EXAONE-3.5-2.4B-Instruct	16%	13.7%	10.5%	48.8%	40.3%	25.8%	33.3%



Discussion: Limitation & Error analysis

- Model size may affected SQL generation: due to resource limitation
- Vector search: Errors may stem from generation step or **search** step



Conclusion

- Text2SQL differs fundamentally from other NLP tasks
- Overall low performance: Korean Text2SQL still has a long way to go
- Importance of developing Korean-focused models



Thank you