LREC-COLING 2024

CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, Alice Oh

KAIST

Introduction

It becomes important to develop culturally-aware LLM. [1]

[1] SHI, Weiyan, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. arXiv preprint arXiv:2404.15238, 2024.

Introduction

It becomes important to develop culturally-aware LLM. [1]

In the Korean context, there was a lack of focus on the culturally-aware evaluation of LLM.

[1] SHI, Weiyan, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. arXiv preprint arXiv:2404.15238, 2024.

Introduction

It becomes important to develop culturally-aware LLM. [1]

In the Korean context, there was a lack of focus on the culturally-aware evaluation of LLM.

Existing Korean benchmarks are either too easy or are translated from English [2,3] **,**
not reflecting specific Korean cultural and linguistic nuances.

[1] SHI, Weiyan, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. arXiv preprint arXiv:2404.15238, 2024.
[2] PARK, Sungjoon, et al. Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680, 2021.
[3] HAM, Jiyeon, et al. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289, 2020.

Introduction

It becomes important to develop culturally-aware LLM. [1]

In the Korean context, there was a lack of focus on the culturally-aware evaluation of LLM.

Existing Korean benchmarks are either too easy or are translated from English [2,3] **,**
not reflecting specific Korean cultural and linguistic nuances.

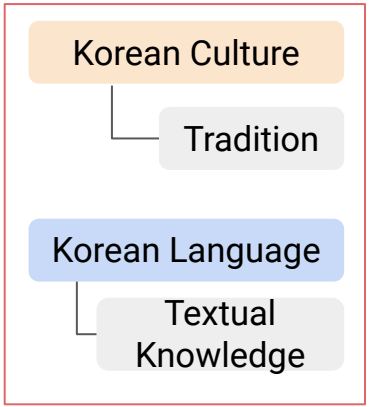
[1] SHI, Weiyan, et al. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. arXiv preprint arXiv:2404.15238, 2024.
[2] PARK, Sungjoon, et al. Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680, 2021.
[3] HAM, Jiyeon, et al. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289, 2020.

CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean

1. 11 fine-grained categories of Korean culture and language
2. Multiple choice questions ranging from everyday life to specific subject areas

CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean

1. **11 fine-grained categories of Korean culture and language**
2. Multiple choice questions ranging from everyday life to specific subject areas



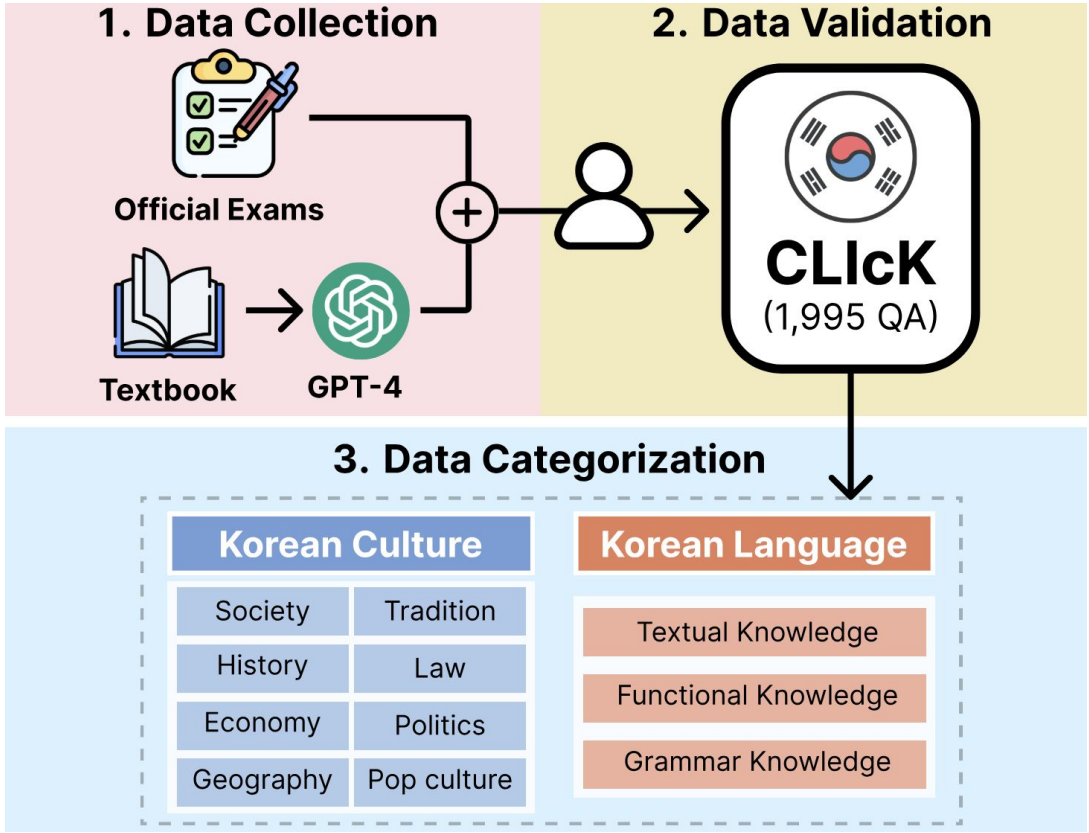
추석의 대표적인 음식은 무엇인가요?
What is the representative food of Chuseok?
1. 김치 Kimchi
2. 비빔밥 Bibimbap
3. ✓ 송편 Songpyeon
4. 떡국 Tteok-guk

다음은 무엇에 대한 글인지 고르십시오.
What is the following text about?
비타민이 가득! 하루 필요한 야채가~ 이 한병에!
Full of vitamins! All the vegetables you need for a day~ in this one bottle!
1. ✓ 주스 Juice
2. 콜라 Coke
3. 우유 Milk
4. 커피 Coffee

CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean

1. 11 fine-grained categories of Korean culture and language
2. **Multiple choice questions ranging from everyday life to specific subject areas**

CLIcK Construction

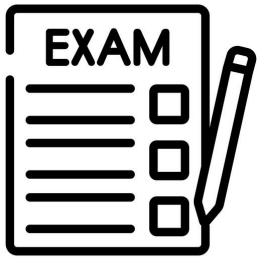


Data Collection - Two Approaches

1. Select questions from standardized Korean exams
2. Using GPT-4 to generate new questions based on textbook

Data Collection - Two Approaches

1. Select questions from standardized Korean exams
2. Using GPT-4 to generate new questions based on textbook



Source	Code	Type	Subject
College Scholastic Ability Test of Korea	CSAT	Exam	Language Geography
Test of Proficiency in Korean	TOPIK	Exam	Language
National Public Service Examination - Grade 9	PSE	Exam	Language History
Public Service Aptitude Test	PSAT	Exam	Constitution
Korean History Exam-Basic	KHB	Exam	History
Test of Teaching Korean as a Foreign Language	Kedu	Exam	Language Culture

Six examinations from Korean institutions

CLOVA OCR

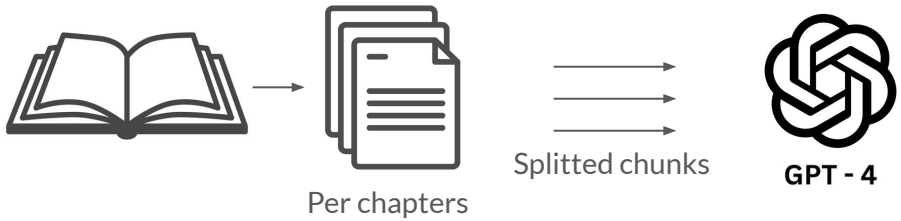
Convert exam files using OCR

Data Collection - Two Approaches

1. Select questions from standardized Korean exams

2. Using GPT-4 to generate new questions based on textbook

We use this approach to introduce novel cultural questions, which are not covered in the exams



Textbook for Korean Immigration and Integration Program(KIIP)

English Translated Prompt

Read the following passage and create 10 multiple-choice questions based on it.
Ensure that the content of each question does not overlap and format them in JSON format including the following elements.
The question_id should start from {current_cnt} and increase by 1 thereafter.
"cite": The sentence from the passage used to create the question
"question_id": {Question number}
"question": {Question}
"choices": {Options}
"answer": {Answer}
"Passage": {content}

Data Validation

- Only validate the GPT-4 generated questions
- Validated by four of the authors, who are Korean native speakers
- The initial dataset reduced to 1,245 samples (62.9% of the original)

Validation Criterion
1. Questions are solely based on the given text. 2. Information in Questions remains consistent over time. 3. Questions should centrally relate to Korea. 4. Questions should be objective and free from bias.

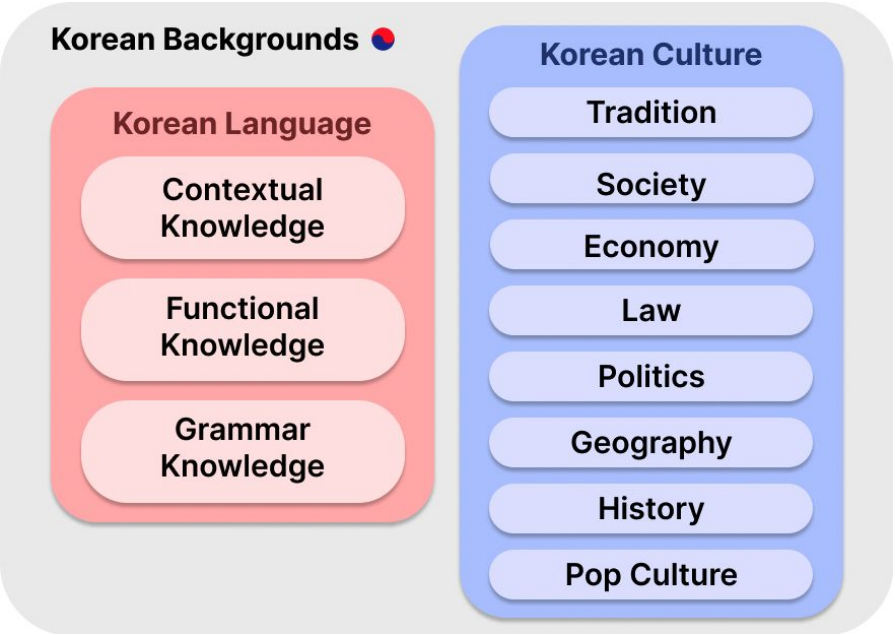
Data Category generation

Korean Language

- Follow definitions of **linguistic knowledge** from Bachman and Palmer (1996)

Korean Culture

- Adopt eight subcategories based on the KIIP textbook



Data Categorization

Case 1 - No further annotation

- GPT4-generated Questions from the textbook
- Questions from Exams that focus on a single subject (i.e. History)

Case 2- Exams contain more than two categories

- A single annotator validates the label assignment based on the provided solutions

CLIcK Dataset

Category		# of Samples		
		Textbook	Exams	Total
Korean Culture	Society	284	25	309
	Tradition	161	61	222
	History	0	280	280
	Law	51	168	219
	Politics	79	5	84
	Economy	57	2	59
	Geography	39	92	131
	Pop culture	15	26	41
Korean Language	Textual	0	285	285
	Functional	0	133	133
	Grammar	0	232	232
	Total	1245	750	1995

Experimental Setting

Model

Type	Model	
API-based LLMs	GPT-3.5-turbo, Claude-2	
Open-source LMs	Multilingual	Korean-specialized
	LLaMA2-chat (7B,13B)	LLaMA2-Ko(7B), KULLM-v2, KoAlpaca Polyglot-Ko(1.3B,3.8B,5.8B,12.8B)

Evaluation Methodology

API-based LLMs	Analyze and select the most likely answer based on the output probabilities of option ID tokens
Open-source LMs	Compare the generated response with the labeled answer

*Cyclic permutation applied to each question to reduce the effect of option ordering bias
*Three different prompts used

Experimental Results

		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

Experimental Results

		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

Claude-2 and GPT-3.5 surpass closed-source models in most categories.

Experimental Results

		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

Their performance remains similar in History, Economy, and Functional Knowledge.

Experimental Results

		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

Claude-2 outperforms GPT-3.5 in almost all categories.

Experimental Results

		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

Open-source models exhibit a low accuracy in the range of 10-50%.

Experimental Results

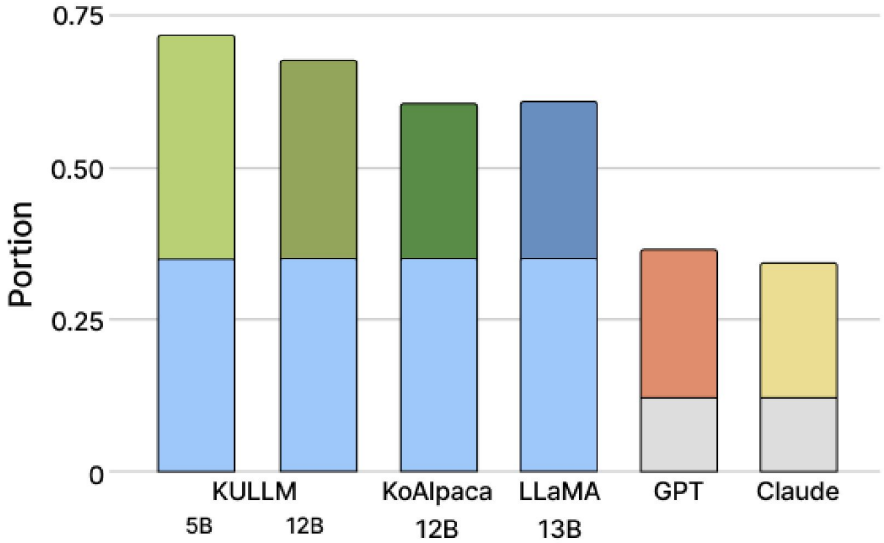
		Polyglot-Ko				KULLM		KoAlpaca		LLaMA-Ko	LLaMA		GPT-3.5	Claude2
		1.3B	3.8B	5.8B	12.8B	5.8B	12.8B	5.8B	12.8B	7B	7B	13B		
Korean Culture	History	26.30	24.71	25.52	24.43	26.48	25.07	26.05	25.84	26.38	30.75	30.73	31.32	**35.00**
	Geography	30.18	28.72	29.06	30.12	27.21	28.66	28.53	30.01	33.21	23.10	25.20	**45.42**	43.30
	Law	38.44	40.16	40.70	43.44	41.67	41.90	40.67	40.13	40.02	45.13	44.12	55.31	**57.09**
	Politics	30.53	32.00	27.74	27.15	26.96	22.68	23.42	28.79	36.03	27.31	26.43	47.75	**60.89**
	Society	34.69	34.31	35.95	37.37	35.95	37.37	33.33	36.44	32.10	39.48	40.93	60.48	**62.43**
	Tradition	32.48	34.01	34.97	33.96	35.86	34.63	32.80	35.45	33.60	33.88	36.12	50.16	**52.10**
	Economy	42.54	42.62	42.25	45.03	43.86	44.08	43.35	43.79	45.32	45.83	46.27	47.59	**53.62**
	Pop culture	29.77	33.68	32.64	29.59	33.60	32.76	34.02	32.63	27.20	33.45	36.41	**68.61**	59.56
	Average	32.71	32.90	33.14	33.40	33.79	33.51	32.33	33.80	33.26	35.44	36.22	49.30	**51.72**
Korean Language	Textual	23.44	23.57	23.27	22.96	24.52	24.65	23.07	24.19	26.75	24.73	24.29	53.19	**55.86**
	Functional	23.77	21.67	22.64	19.84	20.06	19.38	24.76	20.50	26.31	27.04	30.50	32.62	**32.88**
	Grammar	21.87	21.79	23.64	23.05	24.69	25.67	24.03	22.05	23.04	29.32	26.52	38.85	**43.95**
	Average	22.88	22.38	23.27	22.24	23.50	23.78	23.87	22.42	25.69	27.17	26.71	42.32	**45.39**

No trends in performance related to model scale and Korean corpus scale.

Analysis - Overall Difficulty

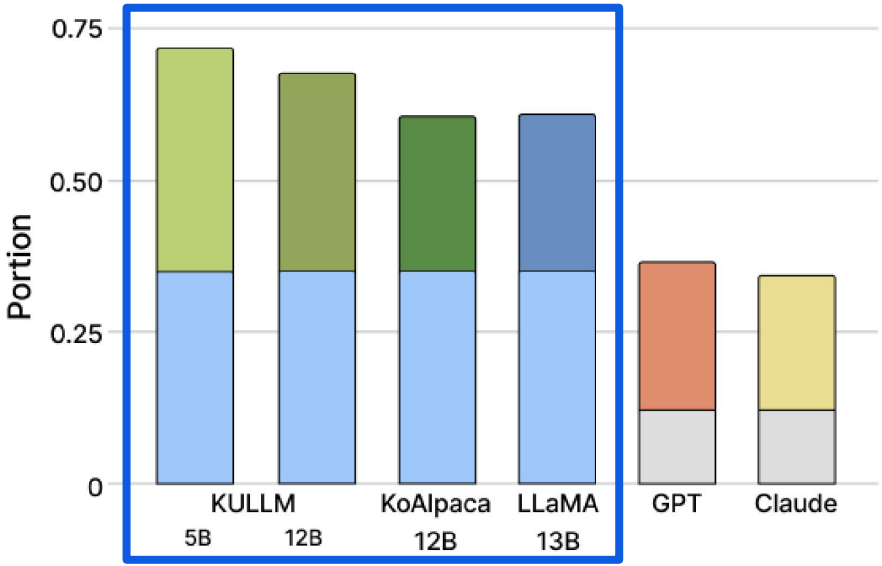
[Definition] *Challenging Problem*

- The problems with an accuracy below random selection threshold



Portions of challenging samples

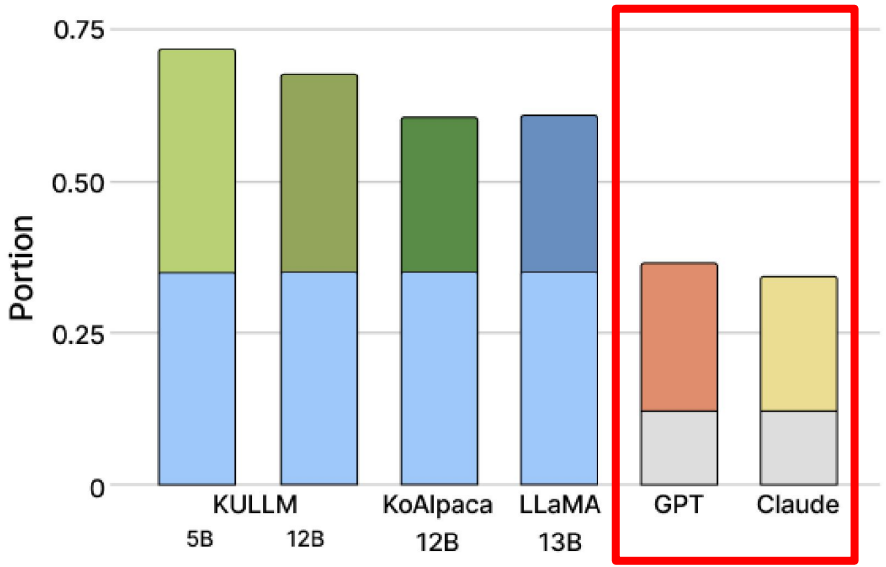
Analysis - Overall Difficulty



Portions of challenging samples

Open-source models have difficulty with over 60% of our dataset.

Analysis - Overall Difficulty



Portions of challenging samples

GPT-3.5 and Claude-2 face challenges with over 30%.

Analysis - Why do model struggle?

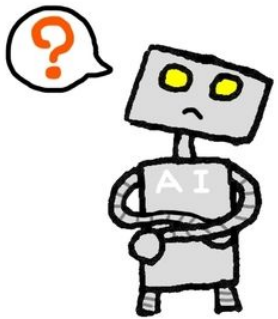
[Definition] *Uncertainty score; opposite meaning of confidence score*

$$\textit{Uncertainty score} = -\frac{1}{\log N} \sum_{i \in \text{options}} p_i \log p_i \quad \textit{(shannon entropy)}$$

Analysis - Why do model struggle?

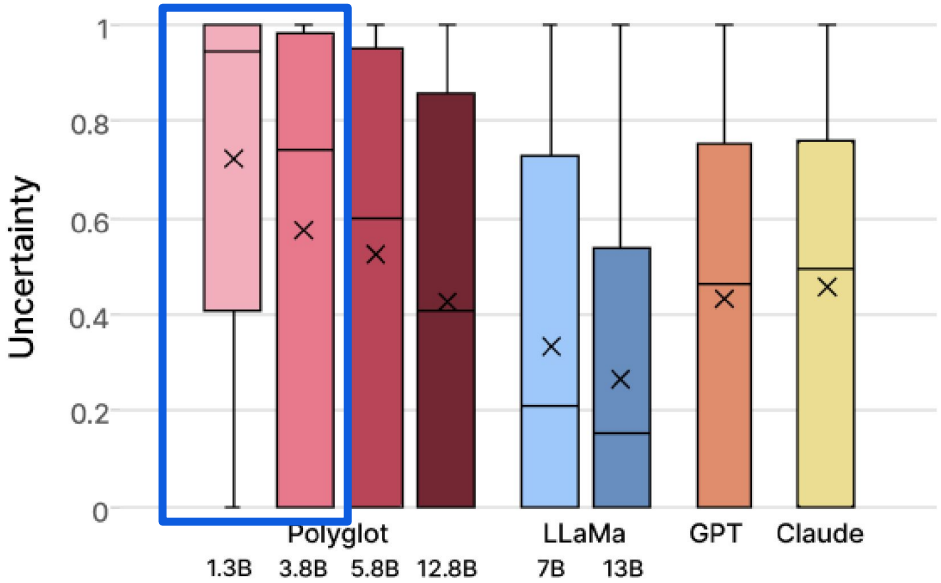
[Definition] *Uncertainty score; opposite meaning of confidence score*

$$\textit{Uncertainty score} = -\frac{1}{\log N} \sum_{i \in \text{options}} p_i \log p_i \quad \textit{(shannon entropy)}$$



Golden Answer	(A)			
Model's Answer	(B)	(C)	(D)	→ **High** Uncertainty
Models' Answer	(B)	(B)	(B)	→ **Low** Uncertainty

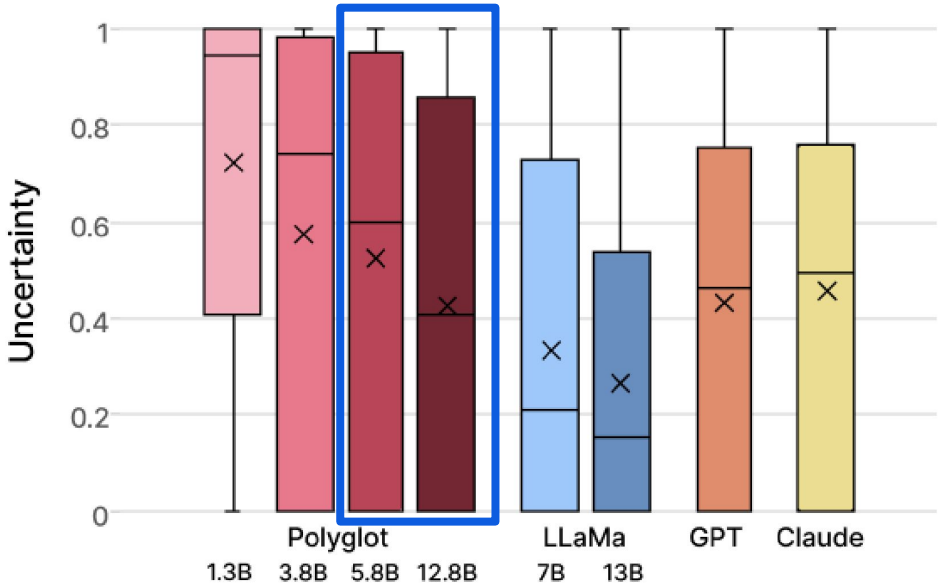
Analysis - Why do model struggle?



Uncertainty score on model's challenging problem

Smaller models tend to randomly select answers without consistency

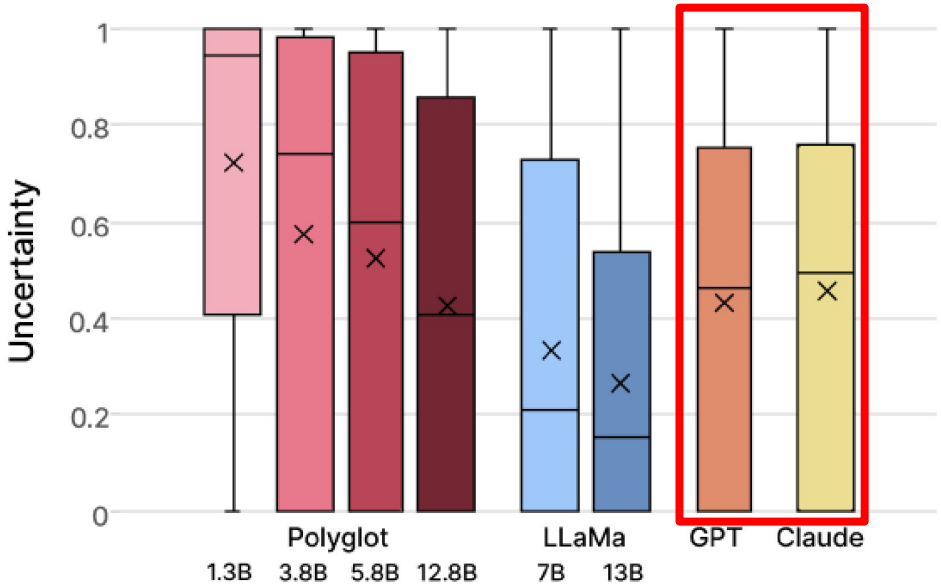
Analysis - Why do model struggle?



Uncertainty score on model's challenging problem

As the size increases, models tend to choose wrong answers consistently.

Analysis - Why do model struggle?



Uncertainty score on model's challenging problem

Ambiguous patterns in GPT-3.5 and Claude-2

Conclusion

- Introduced CLIcK, a **Korean-centric benchmark dataset** derived from official examinations and textbooks, categorized into two main and 11 sub-categories for detailed evaluation.

Conclusion

- Introduced CLIcK, a Korean-centric benchmark dataset derived from local examinations and textbooks, categorized into two main and 11 sub-categories for detailed evaluation.
- Findings indicate that **five open-source models struggle with over 60% of the dataset**, whereas proprietary **LLMs perform better but still require improvements.**

Conclusion

- Introduced CLIcK, a Korean-centric benchmark dataset derived from local examinations and textbooks, categorized into two main and 11 sub-categories for detailed evaluation.
- Findings indicate that five open-source models struggle with over 60% of the dataset, whereas proprietary LLMs perform better but still require improvements.
- Increasing **model size or adding more Korean corpora** does not necessarily enhance understanding of Korean cultural and linguistic aspects.

Take Home Message

Despite the emphasized importance of culturally aware LLMs,

LLMs still lack coverage of linguistic and cultural aspects of non-English languages, and therefore more research is needed on evaluation and modeling techniques.

Thank you!

✉️ kes0317@kaist.ac.kr