K-HALU: MULTIPLE ANSWER KOREAN HALLUCINATION BENCHMARK FOR LARGE LANGUAGE MODELS

Jaehyung Seo
Korea University
Seoul, Republic of Korea
seojae777@korea.ac.kr

Heuiseok Lim*
Korea University
Seoul, Republic of Korea
limhseok@korea.ac.kr

ICLR 2025
International Conference On Learning Representations

https://j-seo.github.io/

1. Introduction

Threats to Open-source LLM Reliability

Limited parameter sizes and constrained training data
→ unfaithfulness, unsupported, unverifiable outputs

However, the lack of publicly available Korean hallucination benchmarks poses a significant threat to thoroughly evaluating the reliability of open-source LLMs.

2. K-HALU

Table 1: Examples of K-HALU benchmark according to the instruction type for selecting hallucinated statements and faithful statements. This table has been translated from Korean into English for the convenience of non-Korean speakers. (Refer to the Korean version in Appendix H).

Hallucinated statements selection type — Society Domain
#Publish Date: January 11, 2021
#Document: The Korean Association for Public Administration was established in 1956 ... Red tape is a symbol of bureaucratic formalism.
#Instruction: Select the hallucinated statements that differ from or are unsupported in relation to the content of the given document. Note that there can be multiple hallucinated statements.
#1: Professor Park Soon-ae is working to expand women's participation in the public sector.
#2: Professor Park Soon-ae went to the United States to pursue a Ph.D. while raising two children.
#3: Professor Park Soon-ae declares an intention to reform the bureaucratic field in the 3G era.
#4: Professor Park Soon-ae points out that administrative convenience is an issue in Korea.
#5: Professor Park Soon-ae states that civil servants prefer maintaining regulations over deregulation.
#Answers: [2, 3]
Faithful statements selection type — International Domain
#Publish Date: August 19, 2015
#Document: Google is launching a new 'Android One' smartphone in six African countries ... It is one of the major projects being pursued.
#Instruction: Select the faithful statements that correspond to the information identifiable from the given document. Note that there may be multiple correct statements.
#1: Google is selling the 'Hot 2' model in six North American countries, including the Canada and Mexico.
#2: The price of the newly launched model is under 100 dollars.
#3: Google is launching a new 'iPhone One' smartphone in six African countries.
#4: The 'Hot 2' model, produced by Infinix, features a 10-inch touchscreen and a 5GHz quad-core processor.
#5: A satellite internet service project is underway in Kampala, the capital of Uganda.
#Answers: [2]

2. K-HALU

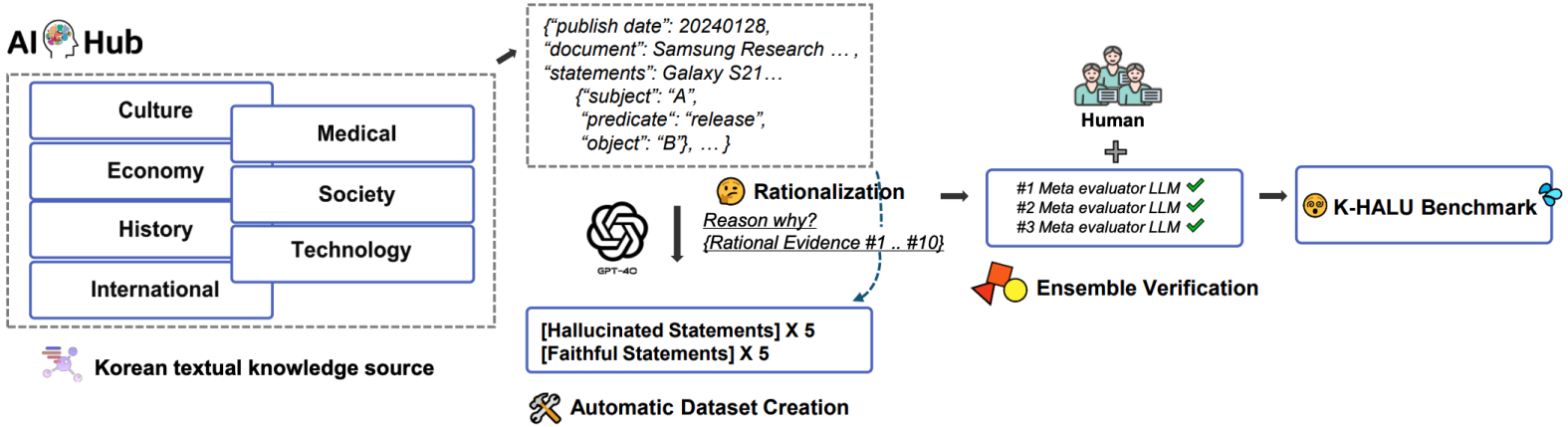

Figure 1: Overview of K-HALU benchmark construction pipeline.

Table 2: Number of K-HALU Benchmark examples according to knowledge domain, instruction type, and multiple-answer questions.

K-HALU Benchmark	Culture	Economy	History	Internation	Medical	Society	Technology	Total
- # examples	300	299	300	329	325	317	300	2,170
Instruction type								
- # hallucination select	150	150	150	165	162	158	150	1,085
- # fact select	150	150	150	164	162	159	150	1,085
Multiple-answers								
- # single-answer	180	179	180	197	196	190	180	1,302
- # two-answers	90	90	90	100	98	95	90	653
- # three-answers	30	30	30	32	31	32	30	215

3. Experiments Setup

GPT-3.5 Turbo
GPT-4 Turbo
GPT-4 omni

Log Probability | Exact Match | LLM-as-a-Judgement

$$P(x) = \frac{1}{|x|} \sum_{i=1}^{|x|} \log \mathbb{P}(x_i | S : x_{<i})$$

4. Results

Baseline Accuracy

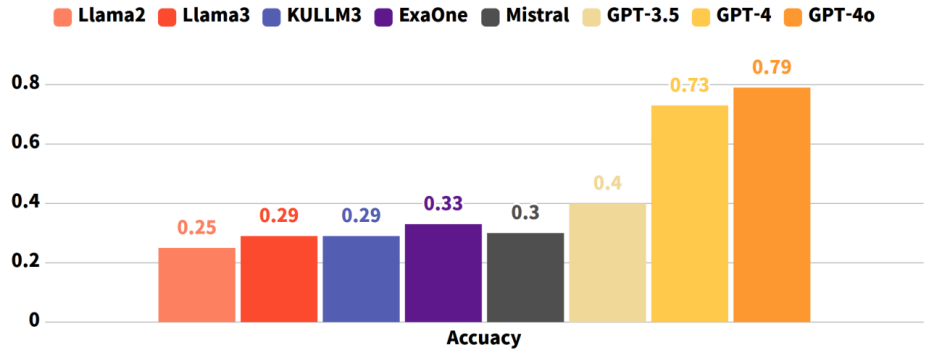

Figure 2: Baseline accuracy of models on the K-HALU test set. Open-source LLMs show lower accuracy, with ExaOne performing the best among them, while GPT-4 omni achieves the highest overall performance in comparison to other models.

Table 9: Performance of LLMs on K-HALU evaluated by Accuracy, F1, Precision, and Recall.

Model	Accuracy	F1	Precision	Recall
Llama2 (Touvron et al., 2023)	0.2475	0.3939	0.4152	0.3979
Llama3 (AI@Meta, 2024)	0.2880	0.4332	0.4503	0.4356
KULLM3 (Kim et al., 2024)	0.2862	0.4156	0.4323	0.4174
ExaOne (Research et al., 2024)	**0.3295**	**0.4770**	**0.5117**	**0.4760**
Mistral-NEMO (MistralAI, 2024)	0.3014	0.4393	0.4530	0.4417
GPT-3.5 Turbo (Ouyang et al., 2022)	0.4046	0.7237	0.6476	0.8260
GPT-4 Turbo (OpenAI, 2023)	0.7346	0.8538	0.8211	0.8899
GPT-4 omni (OpenAI, 2023)	**0.7866**	**0.8732**	**0.8552**	**0.8924**

4. Results

Domain Analysis

Table 4: Model performance across seven knowledge domains, with accuracy evaluated using log probabilities. **Bold** indicates the domain where each model achieved the highest performance, while underline represents the domain where each model recorded the lowest performance.

Models	Culture	Economy	History	International	Medical	Society	Technology
Llama2 (Touvron et al., 2023)	0.2800	0.2642	0.2533	0.2219	0.2185	0.2177	**0.2833**
Llama3 (AI@Meta, 2024)	**0.3200**	0.2843	0.2733	0.3100	0.2923	0.2618	0.2733
KULLM3 (Kim et al., 2024)	**0.3333**	0.2876	0.2833	0.2736	0.2862	0.2744	0.2667
ExaOne (Research et al., 2024)	0.3367	0.3077	0.3033	**0.3526**	0.3385	0.3186	0.3467
Mistral-Nemo (MistralAI, 2024)	**0.3200**	0.3110	0.2933	0.3100	0.3077	0.2808	0.2867
GPT-3.5 Turbo (Ouyang et al., 2022)	0.3800	0.4013	0.3533	**0.4316**	0.4154	0.4164	0.4300
GPT-4 Turbo (OpenAI, 2023)	**0.7667**	0.7157	0.6967	0.7325	0.7384	0.7476	0.7433
GPT-4 omni (OpenAI, 2023)	0.7867	0.7659	0.7633	0.7994	0.7938	0.7950	**0.8000**

4. Results

Multiple Answers

Table 5: Model performance across seven domains based on the number of multiple answers with accuracy evaluated using log probabilities. (1/2/3) indicates the model's performance according to the number of labels. **Bold**: highest, underline: lowest performance.

Models	Culture (1/2/3)	Economy (1/2/3)	History (1/2/3)	International (1/2/3)	Medical (1/2/3)	Society (1/2/3)	Technology (1/2/3)	Total (1/2/3)
Llama2	0.32 / 0.20 / 0.30	0.30 / 0.21 / 0.23	0.29 / 0.21 / 0.13	0.28 / 0.15 / 0.09	0.26 / 0.15 / 0.19	0.26 / 0.13 / 0.22	0.28 / 0.30 / 0.23	**0.28** / 0.19 / 0.20
Llama3	0.36 / 0.26 / 0.30	0.32 / 0.22 / 0.23	0.32 / 0.23 / 0.23	0.33 / 0.28 / 0.31	0.31 / 0.26 / 0.32	0.31 / 0.18 / 0.25	0.29 / 0.23 / 0.27	**0.32** / 0.24 / 0.26
KULLM3	0.38 / 0.26 / 0.30	0.31 / 0.27 / 0.20	0.33 / 0.24 / 0.10	0.29 / 0.24 / 0.25	0.30 / 0.25 / 0.32	0.35 / 0.13 / 0.25	0.26 / 0.26 / 0.33	**0.32** / 0.23 / 0.25
ExaOne	0.35 / 0.30 / 0.37	0.30 / 0.32 / 0.30	0.35 / 0.27 / 0.13	0.33 / 0.37 / 0.44	0.31 / 0.36 / 0.45	0.37 / 0.22 / 0.31	0.38 / 0.31 / 0.27	**0.34** / 0.31 / 0.33
Mistral	0.35 / 0.27 / 0.30	0.34 / 0.27 / 0.30	0.34 / 0.24 / 0.17	0.32 / 0.27 / 0.38	0.31 / 0.30 / 0.35	0.33 / 0.19 / 0.25	0.27 / 0.30 / 0.33	**0.32** / 0.26 / 0.30
GPT-3.5	0.20 / 0.69 / 0.53	0.30 / 0.56 / 0.53	0.26 / 0.58 / 0.23	0.30 / 0.63 / 0.59	0.25 / 0.69 / 0.58	0.27 / 0.62 / 0.69	0.27 / 0.64 / 0.73	0.27 / 0.63 / **0.56**
GPT-4	0.62 / 0.97 / 1.0	0.59 / 0.87 / 1.0	0.57 / 0.87 / 0.90	0.62 / 0.88 / 0.97	0.61 / 0.92 / 0.97	0.65 / 0.88 / 0.91	0.62 / 0.94 / 0.90	0.61 / 0.90 / **0.95**
GPT-4o	0.66 / 0.98 / 1.0	0.66 / 0.89 / 1.0	0.64 / 0.93 / 0.97	0.69 / 0.96 / 1.0	0.67 / 0.98 / 0.97	0.70 / 0.93 / 0.97	0.69 / 0.97 / 0.97	0.67 / 0.95 / **0.98**

4. Results

Instruction Types

Table 6: Model performance in hallucination and fact selection types across the number of multiple answers with accuracy evaluated using log probabilities. **H_Selection** refers to the hallucination selection type and **F_Selection** refers to the fact selection type. A number in () represents the number of labels, and **_Avg.** denotes the average score for each type. **Bold**: highest performance.

Models	H_Selection (1)	F_Selection (1)	H_Selection (2)	F_Selection (2)	H_Selection (3)	F_Selection (3)	H_Avg.	F_Avg.
Llama2	0.1656	0.4015	0.0736	0.3089	0.0373	0.3611	0.1253	**0.3696**
Llama3	0.1825	0.4554	0.0859	0.3884	0.0280	0.4815	0.1382	**0.4378**
KULLM3	0.1656	0.4723	0.0644	0.4006	0.0186	0.4815	0.1207	**0.4516**
ExaOne	0.1488	0.5338	0.0644	0.5505	0.0467	0.6019	0.1134	**0.5456**
Mistral	0.1626	0.4815	0.0767	0.4465	0.0467	0.5463	0.1253	**0.4774**
GPT-3.5	0.3098	0.2215	0.5920	0.6697	0.4860	0.6296	**0.4120**	0.3972
GPT-4	0.6043	0.6231	0.8834	0.9266	0.9626	0.9352	0.7235	**0.7456**
GPT-4o	0.6288	0.7185	0.9479	0.9480	0.9813	0.9815	0.7594	**0.8138**

4. Results

ICL & COT

Table 7: Model performance by # shots and CoT with accuracy evaluated using log probabilities.

Models	Zero-shot	3-shot	3-shot + CoT
Llama2 (Touvron et al., 2023)	**0.2475**	0.2410	0.2429
Llama3 (AI@Meta, 2024)	0.2880	0.2880	**0.2926**
KULLM3 (Kim et al., 2024)	0.2862	**0.3018**	0.2959
ExaOne (Research et al., 2024)	**0.3295**	0.2922	0.2991
Mistral-Nemo (MistralAI, 2024)	0.3014	0.3060	**0.3115**

Exact Match & LLM-as-a-Judge

Table 8: Performance comparison of models using Exact Match and LLM-as-a-Judge.

Models	Exact Match			LLM-as-a-Judge		
	Zero-shot	3-shot	3-shot + CoT	Zero-shot	3-shot	3-shot + CoT
Llama2 (Touvron et al., 2023)	0.0106	0.0051	0.0060	0.0101	0.0115	0.0189
Llama3 (AI@Meta, 2024)	0.1203	0.2396	0.1760	0.0484	0.0203	0.0212
KULLM3 (Kim et al., 2024)	0.0194	0.0083	0.0070	**0.1134**	0.1005	0.0797
ExaOne (Research et al., 2024)	0.0484	0.1885	0.1839	0.0922	**0.1276**	0.1023
Mistral-NEMO (MistralAI, 2024)	**0.1240**	**0.2668**	**0.2770**	0.0995	0.1138	**0.1106**

4. Results

Multilingual Extending

Table 10: Performance of Models on K-HALU (English and Malay Versions).

Models	Accuracy	F1	Precision	Recall
K-HALU (English ver.)				
Llama2 (Touvron et al., 2023)	0.2757	0.4124	0.4261	0.4131
Llama3 (AI@Meta, 2024)	0.2727	0.4012	0.4139	0.4038
KULLM3 (Kim et al., 2024)	0.2846	0.4082	0.4235	0.4091
ExaOne (Research et al., 2024)	0.2638	0.4079	0.4288	0.4106
Mistral-Nemo (MistralAI, 2024)	**0.2906**	**0.4354**	**0.4496**	**0.4364**
K-HALU (Malay ver.)				
Llama2 (Touvron et al., 2023)	0.2620	0.4317	0.4393	0.4314
Llama3 (AI@Meta, 2024)	0.2715	0.4443	0.4528	0.4452
KULLM3 (Kim et al., 2024)	0.2600	0.4268	0.4373	0.4274
ExaOne (Research et al., 2024)	0.2505	0.4171	0.4297	0.4189
Mistral-Nemo (MistralAI, 2024)	**0.2772**	**0.4491**	**0.4591**	**0.4476**

5. Conclusion

- We proposed **K-HALU**, a benchmark specifically designed to evaluate hallucination detection in Korean
- K-HALU leverages Korean textual documents from **seven diverse domains** to rigorously assess the ability of LLMs to detect hallucinations.
- The benchmark utilizes a strict **multiple-answer evaluation framework**, requiring models to identify **all possible correct answers**, enhancing insights into their reliability.
- Analysis shows that **open-source LLMs significantly underperform** in hallucination detection compared to closed API models.
- Future work will focus on **improving datasets and developing advanced models** to further reduce hallucinations in Korean language processing

Our Code & Dataset is available!!

README MIT license

K-HALU

K-HALU: Multiple Answer Korean Hallucination Benchmark for Large Language Models [Paper]

Jaehyung Seo and Heuiseok Lim

NLP & AI Lab, Korea University

Dataset Download

The **K-HALU** dataset will be available on AI-HUB in **March 2025**.
Dataset Link (Coming Soon)

- The K-HALU dataset is available for download and use through the provided link upon **agreeing to the usage policy and submitting a usage application**.
- K-HALU has been developed and released in compliance with the **National Information Society Agency of Korea (NIA)**'s data usage policies and registration procedures.

📌 **GitHub** https://github.com/J-Seo/K-HALU

Thank You