Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators

Liang Chen, Yang Deng, Yayao Bian, Bingzhe Wu,

Tat-Seng Chua, Kam-Fai Wong

The Chinese University of Hong Kong, National University of Singapore,
Tencent AI Lab

❖ Motivation & Key Result

➢ Motivation

- The community is concerned about the reliability and potential implications of using uncensored LLM-generated knowledge.
- Existing evaluations focus on single aspects, such as factuality, hindering a comprehensive understanding of LLM-generated knowledge.

➢ Key Result

- Unearthed key factors influencing reliability in generated knowledge, like long-tail topics, long-form generation and model capacity.
- Revealed a surprising insight: lower factuality in generated knowledge doesn't significantly hamper downstream tasks.
- Demonstrated that output relevance and coherence outweigh minor factual errors.

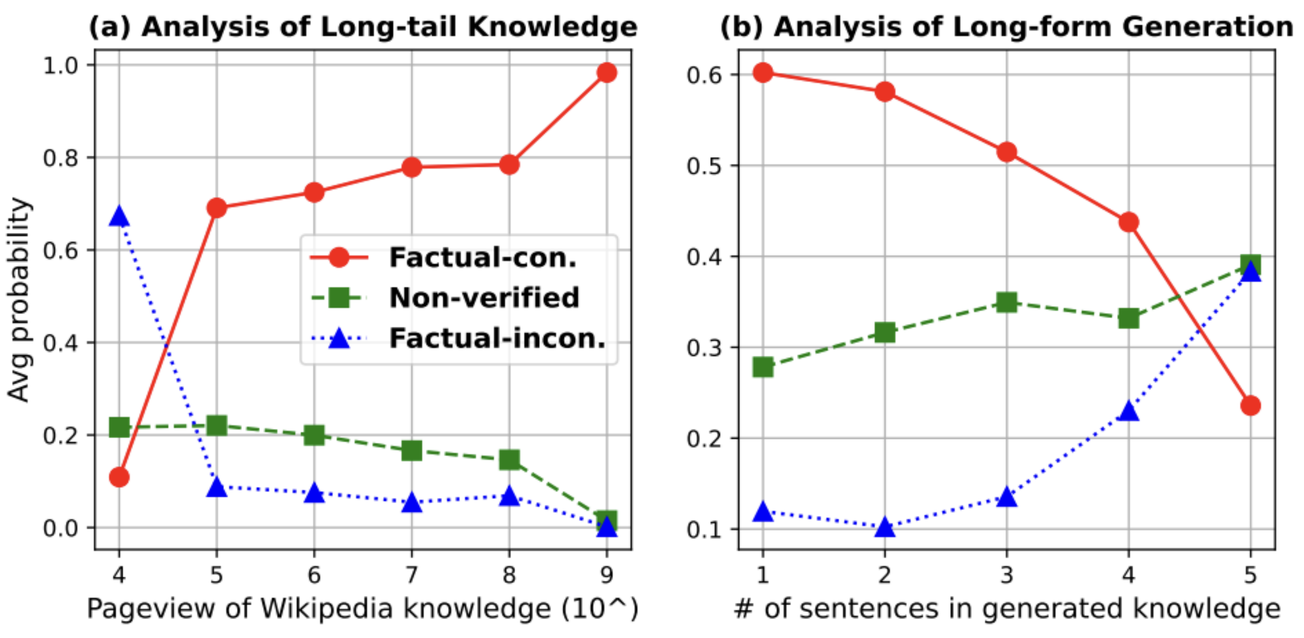
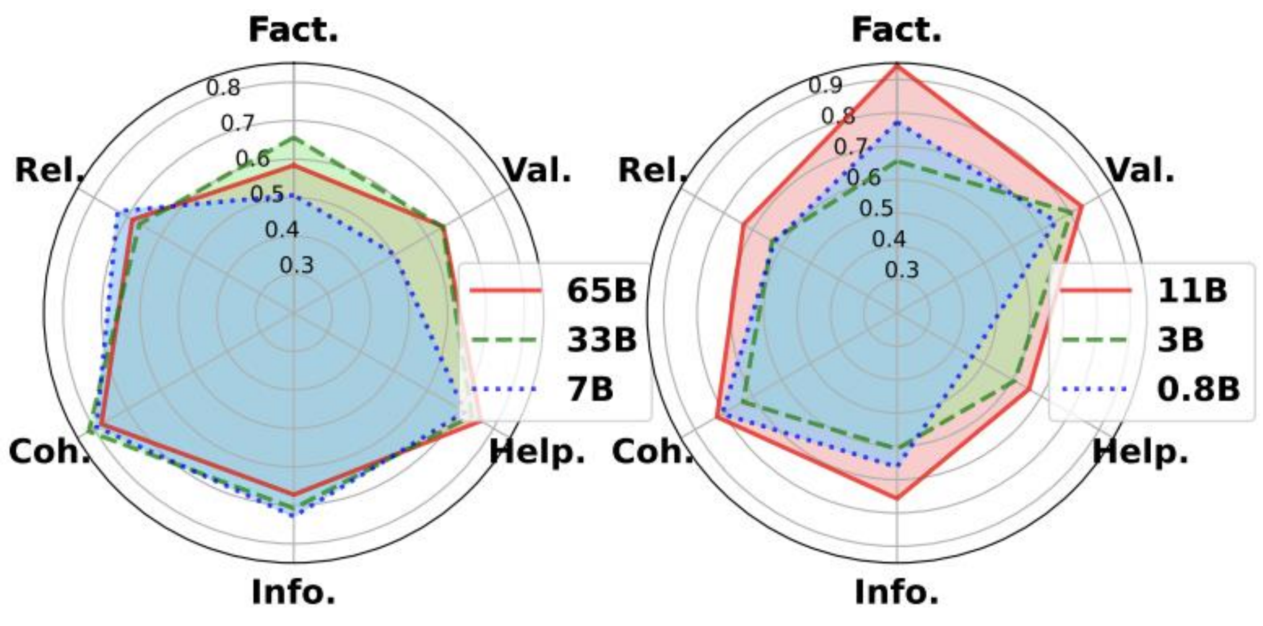


Figure 2: The impact of knowledge frequency and length on the factuality of the generated knowledge.



Model scale of FLAN-T5 Model scale of FLAN-T5

❖ Main Experimental Results

➢ Open-Domain Question Answering Evaluation

- Open-source LlaMA-65B outperforms ChatGPT in factuality while lagging in relevance.
- Generated knowledge generally surpasses retrieved ones, except for factuality and informativeness.
- Though less factual, generated knowledge can enhance downstream task validity better.

Model	Setting	Factuality Fact-cons.	 Non-verif.	 Fact-incon.	Relevance	Coherence Coh-sent.	 Coh-para.	Inform.	Helpful.	Validity
DPR	Supervised	**97.78%**	2.23%	0.00%	0.7514	0.0301	0.7194	**0.8965**	0.1236	36.86%
FLAN-T5	Zero-shot	58.40%	27.80%	13.80%	0.6848	**0.1249**	0.7776	0.6727	0.0000	32.47%
LLaMA		94.20%	4.80%	1.00%	0.7316	0.1183	0.8240	0.7572	0.2191	42.00%
ChatGPT		83.63%	13.6%	2.77%	0.8491	0.0909	**0.9033**	0.7330	0.1461	**43.35%**
FLAN-T5	Few-shot	20.75%	62.40%	25.40%	0.6787	0.0416	0.8110	0.6899	0.0000	34.65%
LLaMA		89.00%	9.20%	1.80%	0.6966	0.0776	0.8550	0.8545	**0.2528**	40.49%
ChatGPT		86.07%	10.97%	2.96%	**0.9205**	0.0653	0.8837	0.7700	0.1966	42.36%

Table 2: Automatic evaluation results of different LLMs in the Natural Question test set. Underlined and **Bold** results denote the best results among each setting and among all settings, respectively.

➢ Knowledge-grounded Dialogue Evaluation

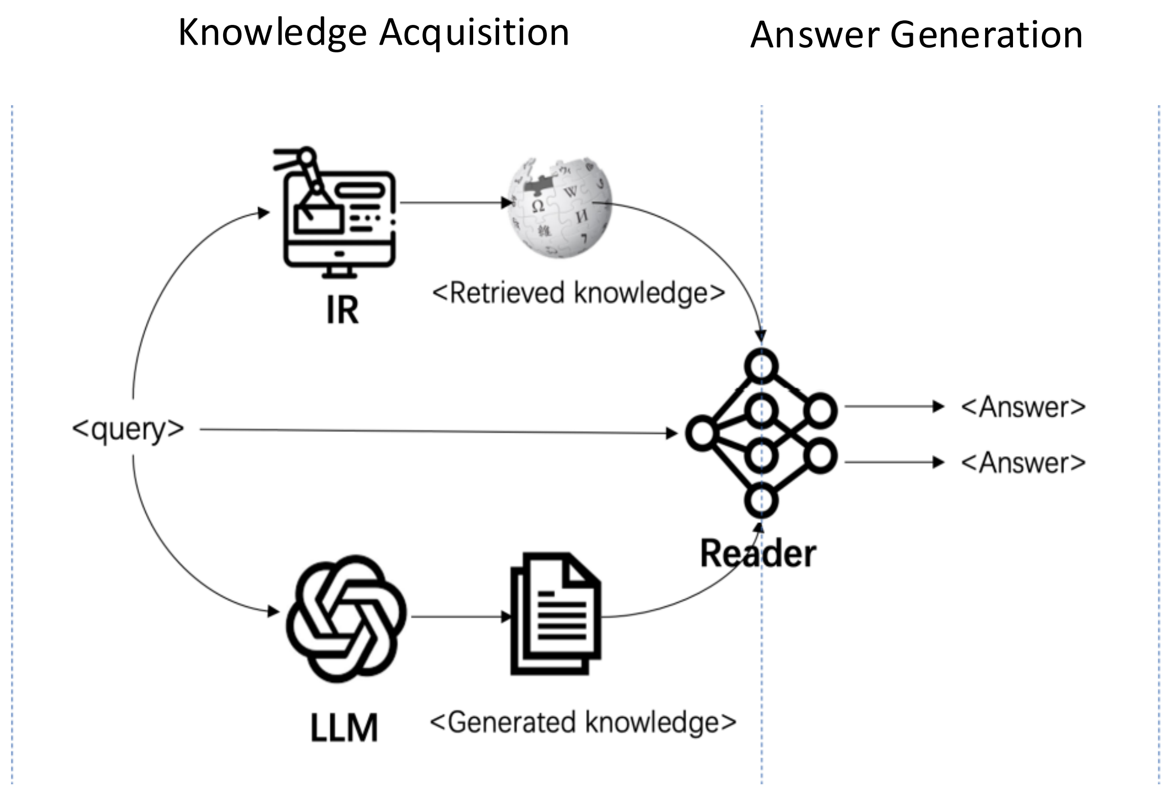
- DPR struggles to retrieve relevant, useful knowledge for knowledge-grounded dialogues.
- Few-shot in-context learning often compromises the factuality of LLM-generated knowledge.
- FLAN-T5 underperforms as a knowledge generator due to its low factuality and limited usefulness in downstream tasks.

Model	Setting	Factuality Fact-cons.	 Non-verif.	 Fact-incon.	Relevance	Coherence Coh-sent.	 Coh-para.	Inform.	Helpful.	Validity
DPR	Supervised	**91.96%**	5.18%	2.87%	0.0907	0.0223	0.6569	**0.9357**	0.0000	61.52%
FLAN-T5	Zero-shot	77.90%	17.28%	4.82%	0.3776	0.1203	0.8331	0.7239	0.0904	56.97%
LLaMA		89.46%	8.89%	1.65%	0.5041	0.0548	0.8389	0.7889	**0.1178**	63.50%
ChatGPT		88.51%	10.38%	1.11%	**0.5283**	0.1028	**0.9250**	0.7448	0.1023	59.76%
FLAN-T5	Few-shot	76.50%	17.20%	6.30%	0.4463	**0.1523**	0.7988	0.6983	0.0934	57.18%
LLaMA		85.07%	12.05%	2.88%	0.3930	0.1088	0.7947	0.7855	0.1132	**63.79%**
ChatGPT		85.75%	12.01%	2.24%	0.4618	0.0979	0.8632	0.7922	0.1164	60.27%

Table 3: Automatic evaluation results of different LLMs in the Wizard of Wikipedia test set.

❖ CONNER Framework

➢ Task Formulation



➢ Design Principles

- Employed theme analysis to uncover error patterns, informing our evaluation perspectives, and ensuring comprehensive coverage of the error spectrum.
- Developed unsupervised metrics, eliminating the need for reference knowledge, and enhancing the applicability in real-world scenarios.
- Implemented intuitive score standardized within the [0, 1] range, facilitating easier comparison and interpretation.

➢ Intrinsic Evaluation: knowledge's internal properties

- ***Factuality***: Whether the information in the knowledge can be verified by external evidence.
- ***Relevance***: Whether the knowledge is relevant to the user query.
- ***Coherence***: Whether the knowledge is coherent at the sentence and paragraph levels.
- ***Informativeness***: Whether the knowledge is novel or unexpected against the model's existing knowledge.

➢ Intrinsic Evaluation: knowledge's downstream impact

•***Helpfulness***: Whether the knowledge can improve the downstream tasks.

•***Validity***: Whether the results of downstream tasks using the knowledge are factually accurate.

❖ Further Analysis

➢ Correlation Analysis: What is key for downstream tasks?

- Factuality issues within LLM-generated knowledge indeed impair downstream task performance.
- For retrieval models, the performance of downstream tasks isn't ensured by high factuality alone. Relevance and coherence of retrieved knowledge also crucial roles..

Model	Extrinsic	Instrinsic Fact.	 Rel.	 Coh-sent.	 Coh-para.	 Info.
DPR	helpful.	0.10	**0.24**†	0.07	-0.03	-0.14†
	validity	0.04	**0.19**†	0.04	-0.06	-0.09
LLMs	helpful.	**0.14**	-0.05	0.10	-0.09	-0.05
	validity	**0.15**†	-0.02	0.07	-0.03	-0.03

Table 5: The Somers' correlation between intrinsic and extrinsic metrics on NQ. Scores with $p\text{-}value < 0.05$

Model	Extrinsic	Instrinsic Fact.	 Rel.	 Coh-sent.	 Coh-para.	 Info.
DPR	helpful.	0.01	**0.27**†	0.10†	-0.03	-0.14†
	validity	-0.01	-0.06	**0.13**†	-0.12†	-0.13†
LLMs	helpful.	0.06	0.05	**0.10**	0.00	-0.16
	validity	**0.24**†	0.09	0.05	-0.02	-0.07

Table 6: The Somers' correlation between intrinsic and extrinsic metrics on WoW.

➢ Case Study: Relevance vs Factuality

- Factual errors in non-critical information minimally impact downstream tasks.
- However, deriving correct answers from irrelevant retrieved knowledge is highly unlikely.

Model	Query	Acquired Knowledge	Answer
LLaMA	Who wrote the song the glory of love?	"The Glory of Love" is a song written by Billy Hill, recorded by American singer Benny Goodman in 1936. The song has been covered by many other artists, including a version by American singer Peggy Lee with Bing Crosby and the Johnny Mann Singers in 1959.	Billy Hill ✓
FLAN-T5	Who wrote the song the glory of love?	The Glory of Love is a song written by the American composer and lyricist Irving Berlin.	Irving Berlin ✗
DPR	Rain sleet or snow that contains a high concentration of acids is called?	... Ice pellets are a form of precipitation consisting of small, translucent balls of ice. This form of precipitation is also referred to as ""sleet"" by the United States National Weather Service. (In British English ""sleet"" refers to a mixture of rain and snow)...	icy rain ✗

Table 4: Factuality of acquired knowledge may not influence the validity of the answer. Red words represent factual errors in critical information, while blue words represent factual errors in non-critical information.