LREC-COLING 2024

Halwasa: Quantify and Analyze Hallucinations in Large Language Models: Arabic as a Case Study

Hamdy Mubarak, Hend Al-Khalifa, Khaloud Alkhalefah

Qatar Computing Research Institute (QCRI), Qatar
King Saud University, KSA
Imam Mohammad Ibn Saud Islamic University, KSA

May 2024, Torino, Italy

Overview

1. Introduction
2. Related Work
3. Data Collection
4. Analysis
5. Conclusion

01

Introduction

Introduction

- Large Language Models (LLMs) show remarkable proficiency in text generation

- **Hallucination**: LLMs generate text that is factually incorrect, nonsensical, or misleading

- Hallucination is one of the major risks associated with LLMs
- Addressing hallucinations:

✓ enhance the reliability and applicability of LLMs

✓ hold potential implications for a wide array of applications, e.g., IR, MT…

- First Arabic study:
  - evaluate factuality and reliability of LLMs when generating Arabic text
  - pave path to mitigate hallucinations

Risks associated with large language models

IBM

Although LLM outputs sound fluent and authoritative, there can be risks that include offering information based on “hallucinations” as well as problems with bias, consent or security. Education on these risks is one answer to these issues of data and AI.

- Hallucinations, or falsehoods, can result from the LLM being trained on incomplete, contradictory, or inaccurate data or from predicting the next accurate word based on context without understanding

Research Questions

- Experiment two LLMs: **GPT 3.5** (aka **ChatGPT**) and **GPT-4**

RQ1: To what extent can we rely on the factual output of LLMs in the Arabic language?

RQ2: What are the types of factual errors and linguistic errors?

02

Related Work

Related Work

- English:
  - Survey: Large Foundation Models tend to produce inaccurate content (Rawte et al., 2023)
  - Survey: Metrics, mitigation methods (Ji et al., 2023)
  - Categorize hallucinations: degree, orientation, and type (Rawte et al., 2023)
  - Med-HALT: evaluate hallucination in medical domain (Pal et al., 2023)
  - Chain-of-Verification method: correct mistakes and reduce hallucination (Dhuliawala et al., 2023)
  - etc.

03

Data Collection

Arabic Background

- Arabic has 3 **varieties**:
  - Modern Standard Arabic (MSA): news papers, books, ad formal speeches
  - Classical Arabic (CA): historical books and literature
  - Dialectal Arabic (DA): daily communications and social media

- Typically nouns and adjectives have gender markers such as Taa Marbouta letter "ة"
  Ex: كاتب (kAtib, male write) vs كاتبة (kAtiba, female writer)

- Nouns, verbs, adjectives, and some articles can be singular, dual, or plural
- Nouns and adjectives can be preceded by a definite article as a prefix
  ex: "ال+بيت" (the+home: the home)

- Agreement on gender, number, and definiteness between:
  - the nouns and their describing adjectives,
  - the verbs and their subjects
  - …

Data Collection

- Choose random 1000 words from **SAMER readability corpus** (Al Khalil et al., 2020)
- For each word: stem, POS, translation, and readability score (1: beginner to 5: specialist)

- Ask **ChatGPT** and **GPT-4** to generate 5 **factual sentences that can be verified**
- Total: 5,000 sentences from ChatGPT and 5,000 from GPT-4

Annotation

- **Annotation**: 50 university students, each student judges 200 sentences
- **Quality Control**: 50 test questions are inserted in the range of each student
- **Labels:**
  - Factual, Correct
  - Linguistic Error, Corrected Sentences
  - Reference link used for verification
  - **Guidelines:** Do not use AI models, verify all facts, be tolerant in some cases (ex: country population), …

Sentence	F	C	E	CS	Ref
يبلغ طول نهر النيل حوالي ٦٦٥٠ كيلومتراً (The Nile River is about 6,650 kilometers long)	1	1	0	-	ar.wikipedia.org/wiki/Nile
هو الجمال يعتبر أسرع حيوان على اليابسة بسرعة ١٢٠ كيلومتراً في الساعة (*It's camels considered the fastest animal on land at a speed of 120 km/hour)	1	0	1	الفهد يعتبر.. (Cheetah is ..)	natgeotv.com/... (National Geographic)
الصبر يساعد على تحقيق الأهداف/أخرج الطالب كتابه من حقيبته (Patience helps achieve goals/The student took his book out of his bag)	0	0	0		

Table 1: Annotation examples. F: Factual, C: Correct, E: Linguistic Error, CS: Corrected Sentence, Ref: Ref. Link

04

Analysis

Statistics

Model	Accuracy%
Reference: Human Annotation	**87**
Random Guess	50
ChatGPT	54
GPT-4	**66**

Table 2: Annotation quality for correctness of the test questions

Model	Total	Factual	Factual/Total	Correct	Correct/Factual	Correct/Total	Ling. Error	Ling. Error/Total
ChatGPT	5,000	3,317	66%	2,506	**76%**	50%	387	**07.7%**
GPT-4	5,000	4,163	**83%**	3,072	74%	**61%**	586	11.7%

Table 3: Model Accuracy

Hallucination Example: سافر المهندس العالمي أنور السادات في أول رحلة عربية إلى الفضاء عام 1985
(International engineer Anwar Sadat traveled on the first Arab flight into space in 1985)

ChatGPT and GPT-4 Errors

Model	str	vocab	agr	prep	dial
ChatGPT	39	36	16	9	0
GPT-4	37	36	21	2	4

Table 4: Analysis of Linguistic Errors for 100 Sample Errors

str: structure error
vocab: wrong vocabulary or odd synonym
agr: wrong agreement between verbs and their subjects
prep: preposition error
dial: using dialectal words

Reasons of Factual Errors in GPT-4

Error Type	Example	%	Comment
Fact	في المملكة العربية السعودية يوجد *أكبر احتياطي نفط في العالم	22	ثاني أكبر
(ex: ordering)	(In Saudi Arabia there is the *largest oil reserve in the world)		second largest
Human Error	ازدهرت الفلسفة والطب والرياضيات في العالم الإسلامي *خلا العصور الوسطى	20	Confusion
	(Philosophy, medicine, and mathematics flourished in the Islamic world *durin the Middle Ages)		
Date	تأسس مقر الإنتربول الدولي في العام * ١٩٤٩ بمدينة ليون الفرنسية	20	في ١٩٢٣
	(Interpol's international headquarters was established in *1949 in Lyon, France)		in 1923
Name	*دل يعتبر عاصمة دولة الهند وتأسست في القرن السادس عشر	10	دلهي
	(*Del is considered the capital of India and was founded in the sixteenth century)		Delhi
Area	مساحة الهرم الأكبر في مصر * ٢٠ فدان	8	١٣ فدان
(Quantities)	(The area of the Great Pyramid in Egypt is *20 acres)		13 acres
Fractions	مساحة الصين * ٩,٥ مليون كيلومتر مربع	6	٩,٦
	(China's area is estimated at about *9.5 million square kilometers)		9.6

Table 5: Common Reasons of Factual Errors in GPT-4

Verification Websites

Website	Description	%
ar.wikipedia.org	**Arabic Wikipedia**	**55.6**
aljazeera.net	Qatari Aljazeera Media Network	3.4
mawdoo3.com	Jordanian content publisher	2.3
bbc.com	British broadcaster	1.5
un.org	The United Nations	1.5

Table 6: Verification Websites and their Usage

- Annotators use 800+ diverse websites as references
- The remaining websites are used less than 1% of the cases

Factuality vs Readability

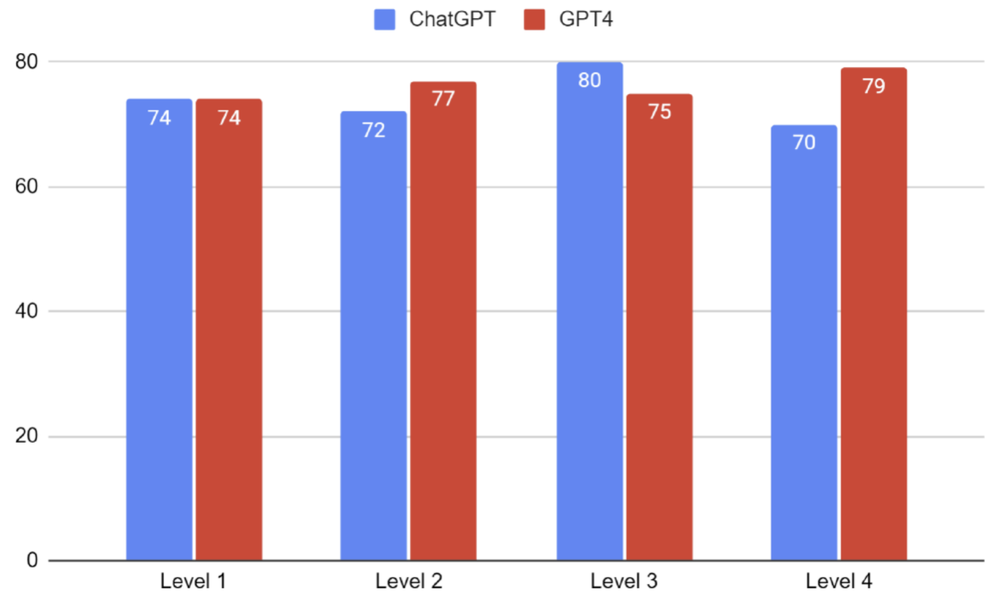

Figure 1: Percentage of Factual/Correct Sentences for Different Readability Levels

- There is no clear relation between factuality and readability
- This needs more investigation

05

Conclusion

Conclusion

- ChatGPT and GPT-4 generate incorrect information
- GPT-4 generates less hallucination than ChatGPT

- We analyzed:
  - error types
  - Reasons for hallucination

- Download link: https://alt.qcri.org/resources/ArabicLLMsHallucination.zip
- Shared Task: https://sites.google.com/view/arabic-llms-hallucination