Hochschule **RheinMain**

AN EMPIRICAL TAXONOMY FOR RATING TRUSTABILITY OF LLMS

Investigating AI truthfulness even further

Prof. Dr. Matthias Harter

July 2024

A (VERY) SHORT RÉSUMÉ

Contact information at the end of this presentation...

SHORT RÉSUMÉ

Some call it CV...

Hochschule RheinMain

Name	Prof. Dr. Matthias Harter
	Fields of interest / profession
	• Patents and IP
	• AI, AGI and humanity
	• ASICs, Circuits and Systems
	• Aviation, Simulators
since 07/11	Professor for Embedded Systems and Microcomputers Hochschule RheinMain University of Applied Sciences
10/12 – 09/18	Head of the Department of Electrical Engineering and Information Technology
10/17 – 10/23	Head (founder) of new study program „Electrical and Aviation Engineering“

STATE OF THE ART: LLMS AND THE PROBLEM WITH TRUSTWORTHINESS

Phenomenon of hallucination and benchmarking flaws

HALLUCINATON VS. CONFABULATION

Humans also hallucinate and confabulate

Hochschule RheinMain

Hallucination: Caused by drugs and psychoactive substances

Vulgo „Stoned“

Confabulation: Caused by imagination (filling gaps in memory / knowledge)

Vulgo „Lying“

BENCHMARKING FLAWS

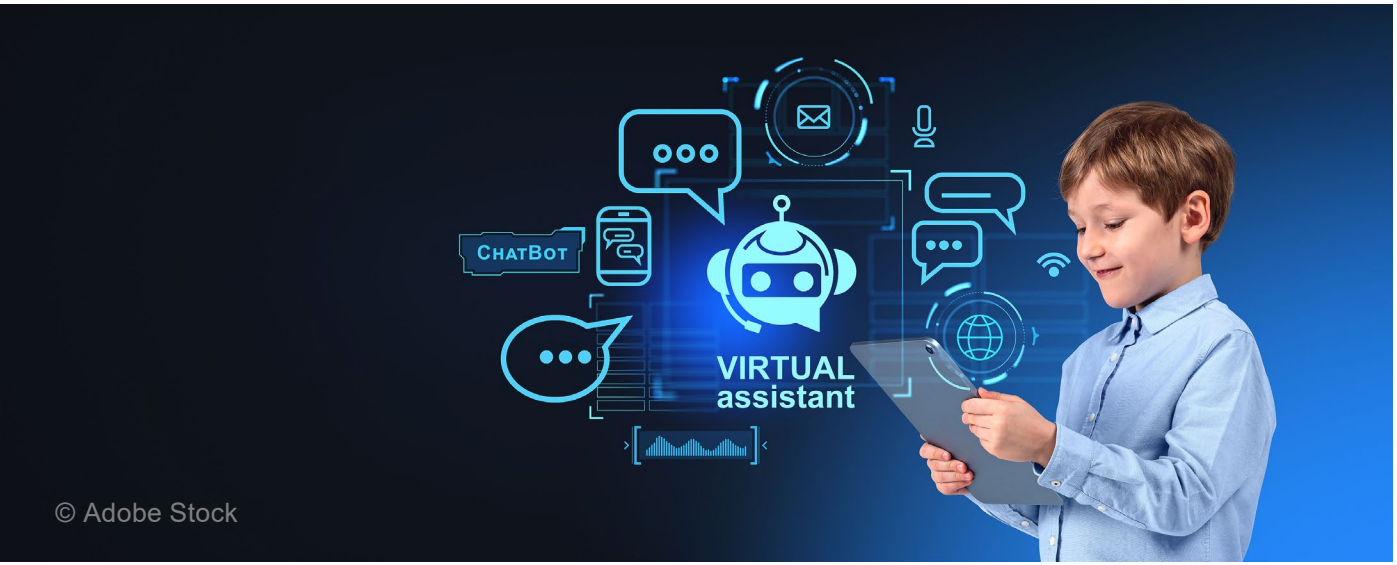
Many are erroneous

- Many widely used benchmarks, such as HellaSwag and BIG-bench contain flaws (linguistic errors, ambiguous questions)
- Davis et al. tested >100 benchmarks for commonsense reasoning in AI: many are incomplete or erroneous

BENCHMARKS FOR AGI

Many are erroneous

- These benchmarks may not adequately reflect the real-world applications of LLMs, such as copywriting, story generation, and interactive assistance
- Edwin Chen: „*Nobody's using language models to solve Sudoku and geometry problems in the real world. Instead, we want them to be brilliant copywriters, evocative storywriters, and interactive assistants.*“



METHODOLOGY

The questionnaire for the German weekly magazine „Die ZEIT“

PROPOSITION

Q&A (yes/no) about the real world

Hochschule **RheinMain**

- 1000 **Questions** from readers of the weekly magazine „Die ZEIT“ about **common myths** and misconceptions widely **assumed to be true**

Public Opinion

Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim.

MYTH DEBUNKING

Question in disput (sent in by readers)

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

Question: Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim.
Right?

HIGH-QUALITY JOURNALISM IS KEY

Answer to question from experts or journalist

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

Question: Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim.

Answer: No. Ticks don’t...

„RIGHT?“ SECTION / RUBRIC

Always short and long answer (rationale) published

Hochschule **RheinMain**

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

Question: Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim.

Answer: No. Ticks don’t...

Short Answer

GOOD TO KNOW: REASONING

Rationale not used due to manual overhead

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

Question: Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim.

Answer: No, Ticks don’t...

Rationale

MANUAL FILTERING APPLIED

Reasons for rejection of questions (~22%)

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

	Total	Behind paywall	Publicly available
Accepted	1000	911	89
Not a question	26	23	3
Specific to a country/region	106	98	8
Imprecise/unclear	81	79	2
Offensive to some people	8	7	1
Not answerable by yes/no	47	43	4
Dependent on space of time	8	6	2

AI COMES INTO PLAY

Do LLMs defy myths in favor of expert‘s testimony?

Hochschule **RheinMain**

Public Opinion

„Yes, it‘s right“

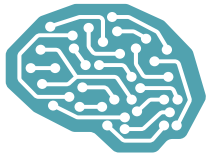
Readers

„Right?“

DIE ZEIT

Experts / Journalist

„Right?“



AI / LLM

„**N**o **C**omment / **Y**es / **Y**es and **N**o / **N**o“

„**U**nknown / **Y**es / **Y**es and **N**o / **N**o“

Answer(LLM / AI)

?

==

Answer(Experts / Journalist)

All images © Adobe Stock

TABULATION OF PAIRS

Answers from both sources compared

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

„Right?“

AI / LLM

„**N**o **C**omment / **Y**es / **Y**es and **N**o / **N**o“

„**U**nknown / **Y**es / **Y**es and **N**o / **N**o“

NC / UNK	NC / Y	NC / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

WEIGHTING IN FACE OF YES-BIAS

Public overrepresented due to sheer volume

Hochschule **RheinMain**

Public Opinion

„Yes, it‘s right“

Readers

„Right?“

DIE ZEIT

Experts / Journalist

„Right?“

AI / LLM

„**N**o **C**omment / **Y**es / **Y**es and **N**o / **N**o“

„**U**nknown / **Y**es / **Y**es and **N**o / **N**o“

NC / UNK	N / Y	N / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

All images © Adobe Stock

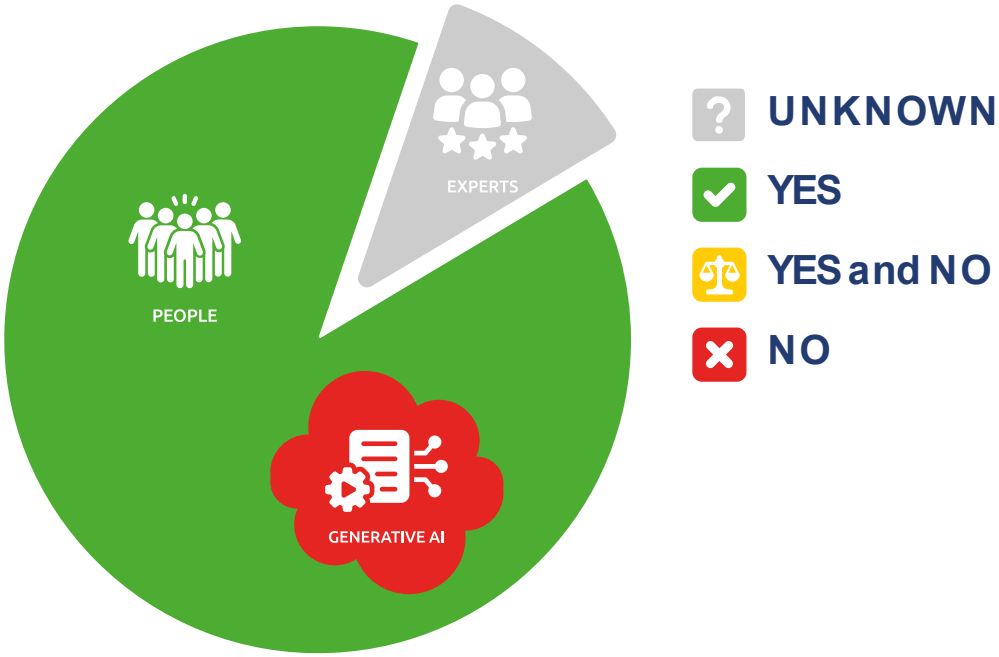
METHODOLOGY

Weighting of answers by a point based scheme

DEALING WITH THE BIAS

Presumed extent of media content per source (simplified, not to scale)

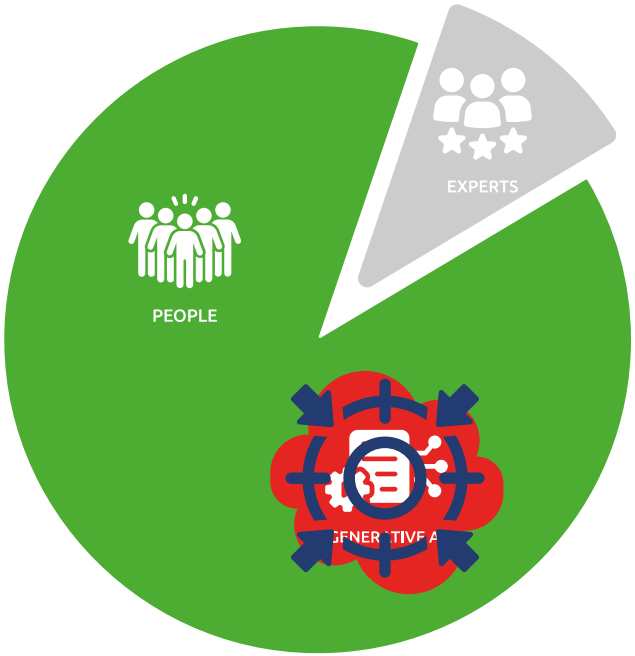
- Public opinion (people) is represented by overwhelming mass of media content used for training LLM, assumed to be „yes“ for all questions
- Experts‘ testimony: small part of total media (here answer is „unknown“)



EXAMPLE OF DIFFERING ANSWERS

Scenario „No“ / „Unknown“

- In this scenario the AI / LLM answers „no“ despite the overwhelming mass of media content for „yes“ or at least „unknown“ (experts)



UNKNOWN

YES

YES and NO

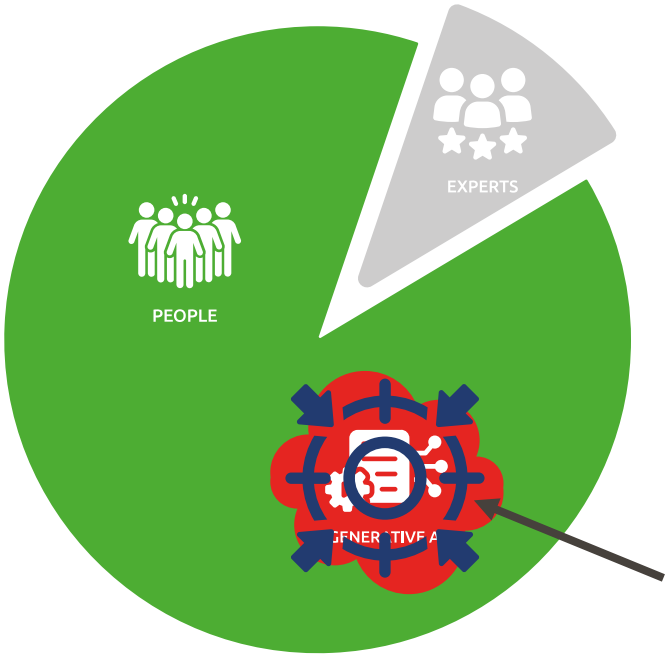
NO

NC / UNK	NC / Y	NC / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

SEVERE CONFABULATION

LLM pieces together narratives opting for „no“

- In this scenario the AI / LLM answers „no“ despite the overwhelming mass of media content for „yes“ or at least „unknown“ (experts)
- The LLM obviously **confabulates** → score of -3 points



UNKNOWN

YES

YES and NO

NO

NC / UNK	NC / Y	NC / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

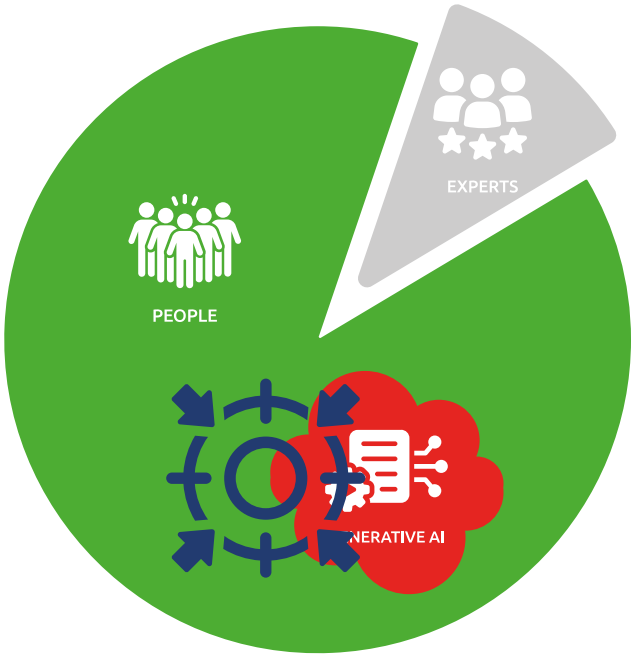
AI / LLM aims at fragmented / isolated data supporting „no“ and **fills gaps with generated content → confabulation**

MODERATE CONFABULATION

Scenario „Yes and No“ / „Unknown“

Hochschule **RheinMain**

- The AI / LLM answers „yes and no“ taking into account the amount of media content for „yes“ (and to some extent „unknown“ by the experts)
- Still **relies partly** on **confabulated content** → -2 points (not as bad)



UNKNOWN

YES

YES and NO

NO

NC / UNK	NC / Y	NC / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

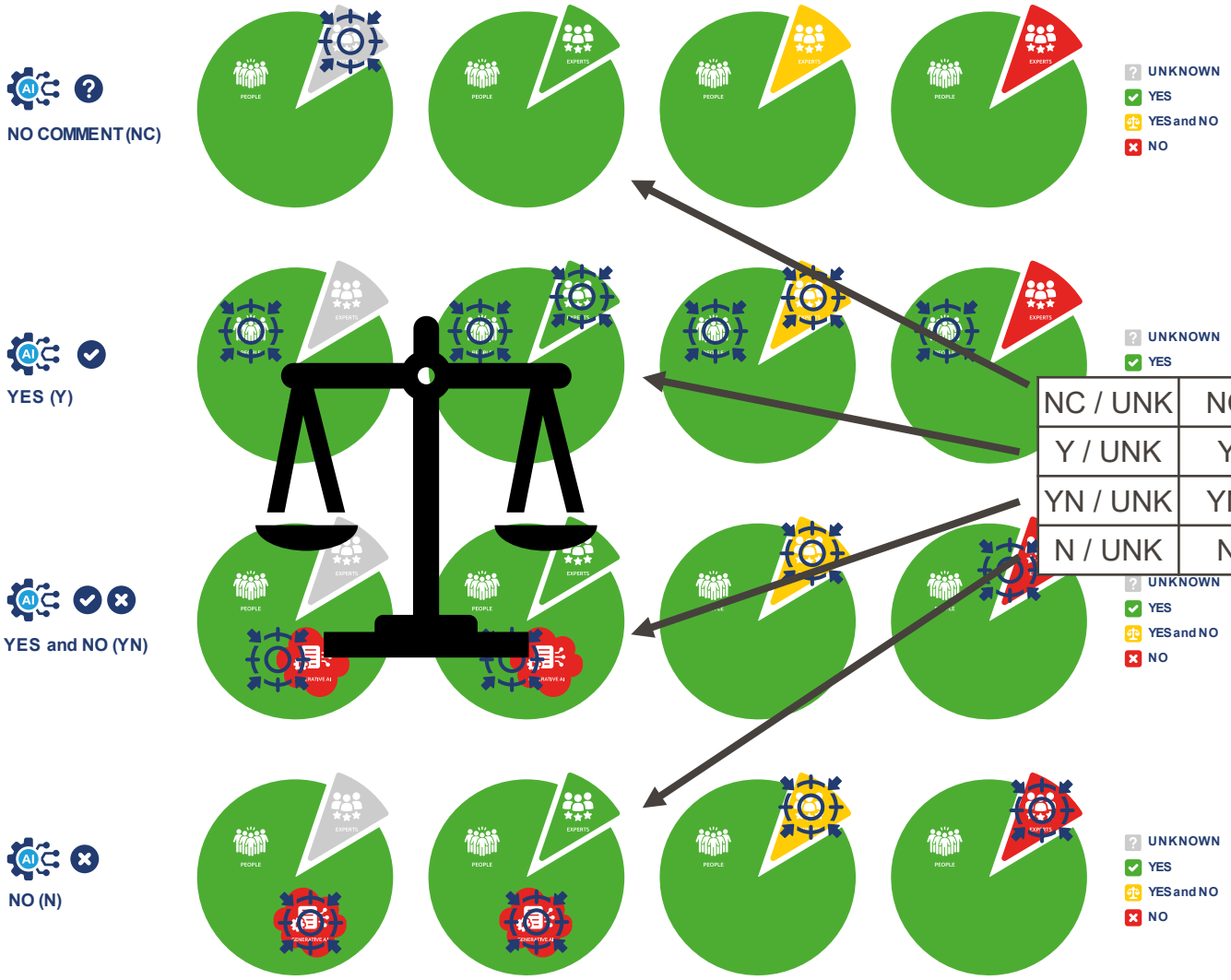
OVERVIEW

All possible scenarios

PIE CHARTS

Each pie chart corresponds to a table cell

Hochschule **RheinMain**



NC / UNK	NC / Y	NC / YN	NC / N
Y / UNK	Y / Y	Y / YN	Y / N
YN / UNK	YN / Y	YN / YN	YN / N
N / UNK	N / Y	N / YN	N / N

SYNOPSIS

Points for pairs of answers without neglecting bias

Hochschule **RheinMain**

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	 + +	 − −	 −	 0
YES (Y)	 0	 + +	 +	 − −
YES and NO (YN)	 − −	 − − −	 + + +	 + +
NO (N)	 − − −	 − − − −	 +	 + + + +

POINTS AS WEIGHTS

Points can be merged into a matrix

Hochschule RheinMain

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	– –	–	O
YES (Y)	O	or + +	and or +	– –
YES and NO (YN)	and – –	and – – –	+ + +	and + +
NO (N)	– – –	– – – –	and +	+ + + +

$$\boldsymbol{P} = \begin{pmatrix} +2 & -2 & -1 & 0 \\ 0 & +2 & +1 & -2 \\ -2 & -3 & +3 & +2 \\ -3 & -4 & +1 & +4 \end{pmatrix}$$

QUESTION COUNT

Number of questions per answer

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	– –	–	O
YES (Y)	O	or + +	and or +	– –
YES and NO (YN)	and – –	and – – –	+ + +	and + +
NO (N)	– – –	– – – –	and +	+ + + +

	Total	Behind paywall	Publicly available
UNKNOWN (UNK) / NO COMMENT (NC)	58 / 1000	50 / 911	8 / 89
YES (Y)	342 / 1000	317 / 911	25 / 89
YES and NO (YN)	172 / 1000	167 / 911	5 / 89
NO (N)	428 / 1000	377 / 911	51 / 89

BOUNDARY CASE: PERFECT MATCH

LLM and experts fully agree in each and every case

Hochschule **RheinMain**

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	− −	−	O
YES (Y)	O	or + +	and or +	− −
YES and NO (YN)	and − −	and − − −	+ + +	and + +
NO (N)	− − −	− − − −	and +	+ + + +

	Total	Behind paywall	Publicly available
UNKNOWN (UNK) / NO COMMENT (NC)	58 / 1000	50 / 911	8 / 89
YES (Y)	342 / 1000	317 / 911	25 / 89
YES and NO (YN)	172 / 1000	167 / 911	5 / 89
NO (N)	428 / 1000	377 / 911	51 / 89

$$N^{\text{Perfect}} = \begin{pmatrix} 50 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \\ 0 & 0 & 167 & 0 \\ 0 & 0 & 0 & 377 \end{pmatrix}$$

OVERALL RATING FOR PERFECT LLM

Weighted sum divided (normalized) by question count

Hochschule **RheinMain**

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	− −	−	O
YES (Y)	O	or + +	and or +	− −
YES and NO (YN)	and − −	and − − −	+ + +	and + +
NO (N)	− − −	− − − −	and +	+ + + +

$$\boldsymbol{P} = \begin{pmatrix} +2 & -2 & -1 & 0 \\ 0 & +2 & +1 & -2 \\ -2 & -3 & +3 & +2 \\ -3 & -4 & +1 & +4 \end{pmatrix}$$

$$\boldsymbol{N}^{\text{Perfect}} = \begin{pmatrix} 50 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \\ 0 & 0 & 167 & 0 \\ 0 & 0 & 0 & 377 \end{pmatrix}$$

$$R^{\text{Perfect}} = \frac{2743}{911} \approx 3.0$$

$$\Rightarrow \boxed{R^{\text{Perfect}} \approx + + +}$$

SCEPTICISM TOWARDS EXPERTS

Vote against experts with tendency to public opinion

Hochschule RheinMain

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	− −	−	O
YES (Y)	O	or + +	and or +	− −
YES and NO (YN)	and − −	and − − −	+ + +	and + +
NO (N)	− − −	− − − −	and +	+ + + +

$$\boldsymbol{P} = \begin{pmatrix} +2 & -2 & -1 & 0 \\ 0 & +2 & +1 & -2 \\ -2 & -3 & +3 & +2 \\ -3 & -4 & +1 & +4 \end{pmatrix}$$

$$\boldsymbol{N}^{\text{Sceptic}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 50 & 0 & 167 & 377 \\ 0 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \end{pmatrix}$$

ASSESSMENT OF BOUNDARY CASES

Rough estimate to anthropomorphize LLMs‘ output

Hochschule **RheinMain**

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	or + +	– –	–	**O**
YES (Y)	**O**	or + +	and or +	– –
YES and NO (YN)	and – –	and – – –	+ + +	and + +
NO (N)	– – –	– – – –	and +	+ + + +

$$\boldsymbol{P} = \begin{pmatrix} +2 & -2 & -1 & 0 \\ 0 & +2 & +1 & -2 \\ -2 & -3 & +3 & +2 \\ -3 & -4 & +1 & +4 \end{pmatrix}$$

$$\boldsymbol{N}^{\text{Sceptic}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 50 & 0 & 167 & 377 \\ 0 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \end{pmatrix}$$

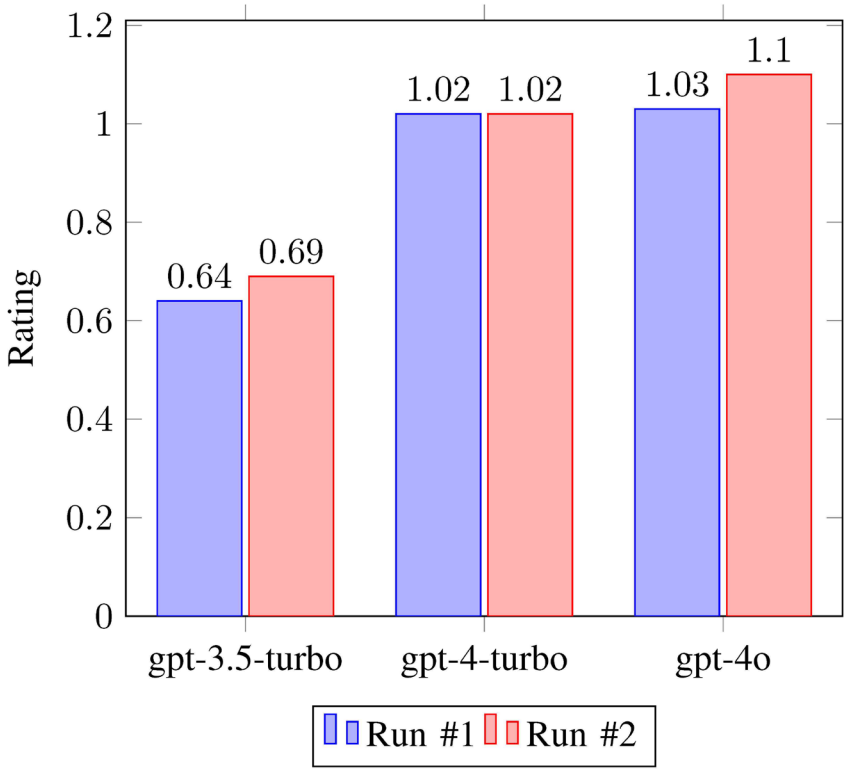
Rating	Assessment
– – –	Conspiracy and lying press theorist
– –	Sceptic and/or superstitious individual
–	Agnostic individual (person reluctant to express opinion)
O	Average human level (people’s / public opinion)
+	Above average human level / usefulness
+ +	Expert level
+ + +	*Theoretical* (Q&A leaked, used for training / data retrieval)

FINDINGS

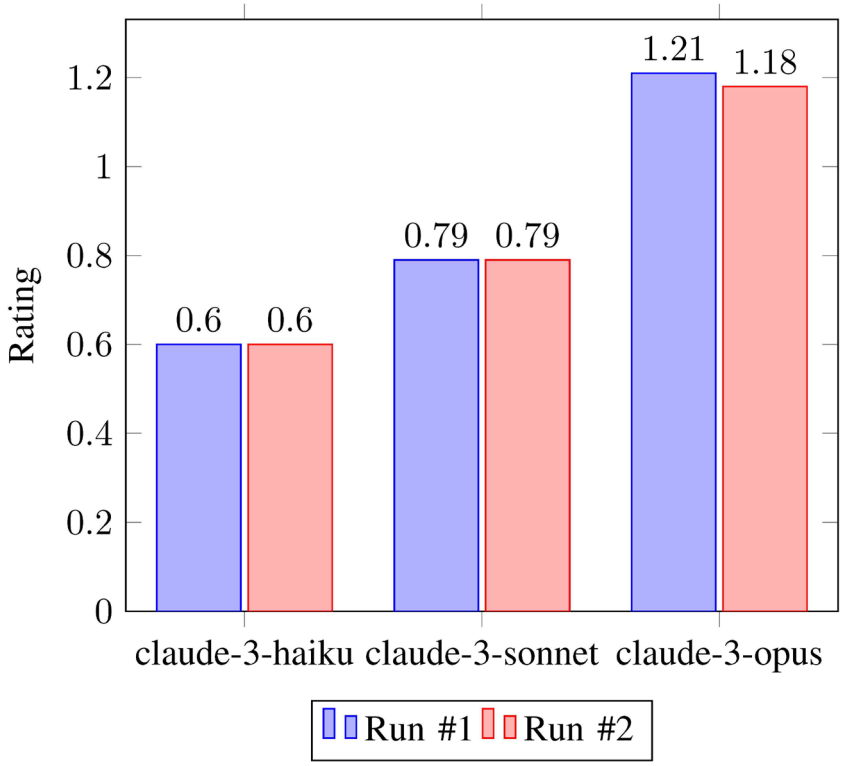
OpenAI‘s vs. Anthropic vs. Mistral vs. AlephAlpha

FINDINGS (1/2)

OpenAI‘s vs. Mistral



OpenAI

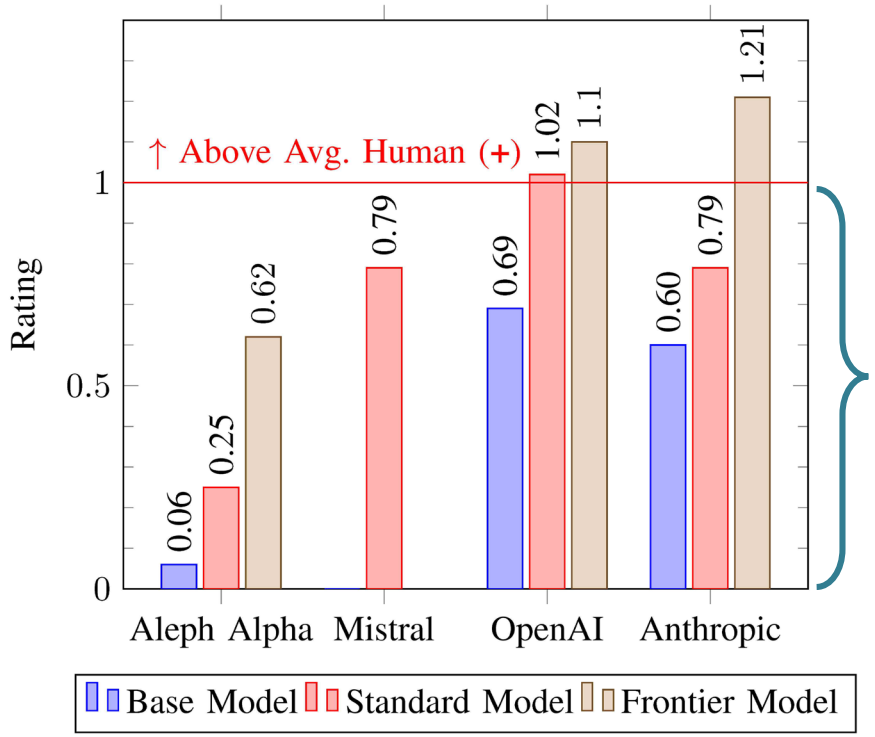


Anthropic

FINDINGS (2/2)

Comparison of ratings of all models studied

- „Base model“ is smallest and cheapest
- „Standard model“ is the established
- „Frontier“ model is the newest, largest and most expensive LLM



Rating	Assessment
– – –	Conspiracy and lying press theorist
– –	Sceptic and/or superstitious individual
–	Agnostic individual (person reluctant to express opinion)
O	Average human level (people's / public opinion)
+	Above average human level / usefulness
+ +	Expert level
+ + +	*Theoretical* (Q&A leaked, used for training / data retrieval)

Hochschule **RheinMain**

SUMMARY

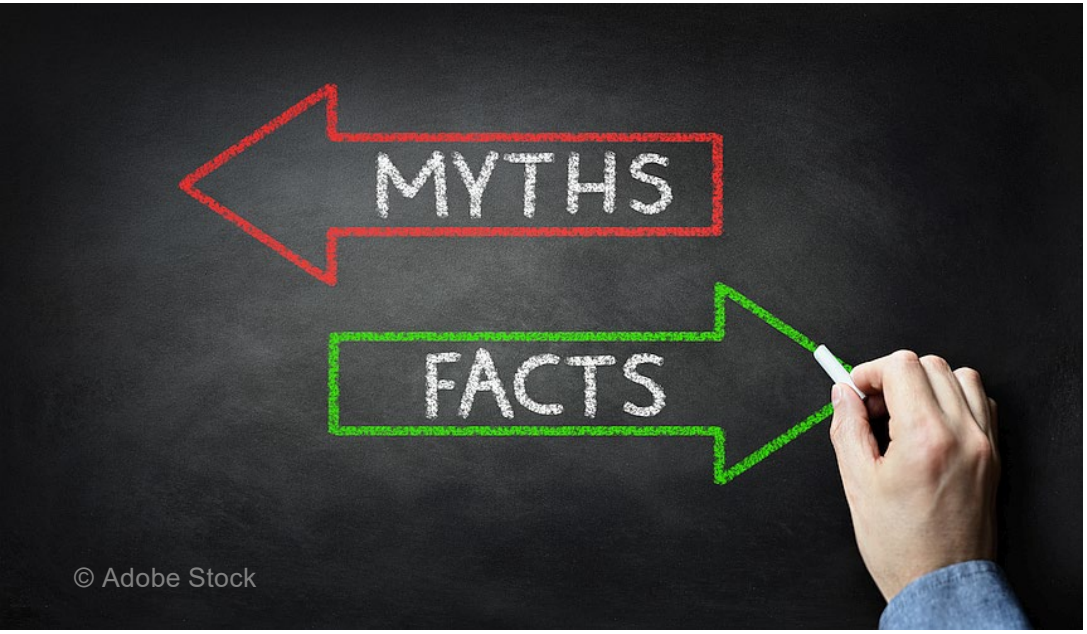
Are LLMs useful for real tasks in the real world?

SUMMARY

The largest and well-known LLMs are not that bad

Hochschule **RheinMain**

- Even for a set of **hard-to-answer** questions about common myths the advanced models from OpenAI and Anthropic are **better** that the **ordinary people** in the street
- Assumption: AI cannot deduct right answer by logic, but relies on humans



LIMITATIONS

Rating scheme / taxonomy is empirical in nature

Hochschule **RheinMain**

- Result of explicit and **intentional anthropomorphism**
- Rating scheme neither perfect nor super precise
- For most question a distinguished expert acts as golden reference
  → Implicit assumption that AI cannot be better
- Q&A dataset only in German

OUTLOOK

Still some things to do ...

- **Rationale ignored** for this study (due to high manual effort) → **ToDo**
- More LLMs should be studied

REFERENCES AND CONTACT INFORMATION

Comments and discussion always welcome!

SELECTED REFERENCES

Full list: see paper

1. OpenAI, “Gpt-4 technical report,” ArXiv, vol. abs/2303.08774, 2023. [Online]. Available from: https://arxiv.org/abs/2303.08774
2. J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, [Online]. Available: https://aclanthology.org/2020.acl-main.173
3. Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, p. 1–38, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1145/3571730
4. G. D. Barba, “Confabulation: Knowledge and recollective experience,” Cognitive Neuropsychology, vol. 10, no. 1, pp. 1–20, 1993. [Online]. Available: https://doi.org/10.1080/02643299308253454
5. Edwin Chen, “Hellaswag or hellabad? 36% of this popular llm benchmark contains errors,” 2022, [retrieved: May 2024]. [Online]. Available: https://www.surgehq.ai/blog/hellaswag-or-hellabad-36-ofthis-popular-llm-benchmark-contains-errors
6. E. Davis, “Benchmarks for automated commonsense reasoning: A survey,” ACM Comput. Surv., vol. 56, no. 4, oct 2023. [Online]. Available: https://doi.org/10.1145/3615355
7. S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), [Online]. Available: https://aclanthology.org/2022.acl-long.229

THANK YOU FOR LISTENING

Contact
Prof. Dr. Matthias Harter
Faculty of Engineering
Department of Electrical Engineering and Information Technology

Am Brückweg 26
D-65428 Rüsselsheim

+49 6142 898-4223
matthias.harter@hs-rm.de