AI-Psychometrics

Was LLMs im Durchschnitt zu sagen haben

Charit

Psychometrics

Messen des Unsichtbaren in **Menschen**

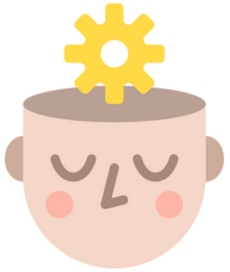
Messverfahren und statistische Methoden um psychologische Merkmale wie **Intelligenz & Fähigkeiten** oder **Persönlichkeit & Einstellungen** zu messen und zu analysieren.

Ermöglicht das Interpretieren und Vorhersagen von Bewegungen in der Gesellschaft und Individuen.

Intelligenz & Fähigkeiten

Persönlichkeit & Einstellungen

Psychometrics



Intelligenz & Fähigkeiten

Wechsler Adult Intelligence Scale
Stanford-Binet Intelligence Scale
Minnesota-Manual-Dexterity-Test
...

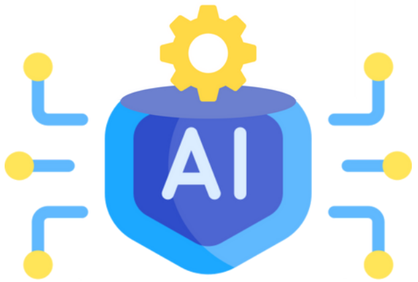
Persönlichkeit & Einstellungen

Big Five Personality Inventory
MMPI
DISC-Assessment
...

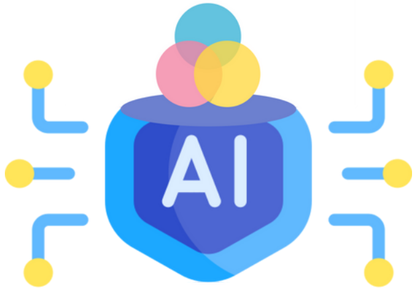
AI-Psychometrics

Messen des Unsichtbaren in **LLMs**

Viele KI-Verfahren nutzen **messbare Trainingsdatenkörper**, wodurch ihr Verhalten erklärt werden kann. LLMs agieren nondeterministisch, ihr Wesen muss indirekt gemessen werden, ähnlich wie beim Menschen.



Intelligenz & Fähigkeiten



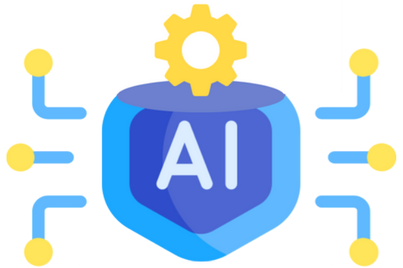
Persönlichkeit & Einstellungen

AI-Psychometrics

Ähnliche **Fähigkeitstests** wie beim Menschen

Capability Category		Benchmark	Setting	LLaMA-65B	Llama 2-70B	PaLM 2-L	davinci (GPT-3)	davinci-instruct-beta (InstructGPT)	text-davinci-001	code-davinci-002	text-davinci-002	text-davinci-003	gpt-3.5-turbo-0301	gpt-3.5-turbo-0613	gpt-4-0314	gpt-4-0613
Knowledge	Question Answering	Natural Questions	1-shot	27.7	27.0	(37.5)	17.8	7.1	23.5	29.2	28.2	38.1	39.6	38.8	48.4	48.6
		WebQuestions	1-shot	42.2	38.2	(28.2)	37.3	11.1	42.1	43.3	45.8	55.4	53.0	53.4	60.3	58.6
		TriviaQA	1-shot	73.4	74.0*	(86.1)	61.5	51.6	68.0	82.6	78.6	82.5	83.2	84.9	92.3	92.1
	Multi-subject Test	MMLU	5-shot	60.1*	67.8*	(78.3)	34.3	39.9	46.7	69.1	62.1	63.7	66.6	67.4	83.7	81.3
		AGIEval-EN	few-shot	38.0	44.0	–	22.0	25.1	31.0	48.4	43.6	44.3	43.3	44.5	57.1	56.7
		ARC-e	1-shot	87.2	93.4	(89.7)	57.2	60.6	74.7	92.8	90.1	91.5	94.1	92.7	98.9	98.6
		ARC-c	1-shot	71.8	79.6	(69.2)	35.9	40.9	53.2	81.7	75.7	79.5	82.9	81.7	94.9	94.6
Reasoning	Commonsense Reasoning	LAMBADA	1-shot	30.9	30.4	(86.9)	53.6	13.8	51.1	84.9	66.0	56.2	67.8	68.2	78.6	87.8
		HellaSwag	1-shot	47.8	68.4	(86.8)	22.8	18.9	34.6	56.4	64.9	60.4	78.9	79.4	92.4	91.9
		WinoGrande	1-shot	54.6	69.8	(83.0)	48.0	49.6	54.6	67.6	65.5	70.6	65.8	55.3	86.7	87.1
	Comprehensive Reasoning	BBH	3-shot CoT	58.2	65.0	(78.1)	39.1	38.1	38.6	71.6	66.0	69.0	63.8	68.1	84.9	84.6
Comprehension	Reading Comprehension	RACE-m	1-shot	77.0	87.6	(77.0)	37.0	43.0	54.4	87.7	84.5	86.3	86.0	84.1	93.5	94.0
		RACE-h	1-shot	73.0	85.1	(62.3)	35.0	33.5	44.3	82.3	80.5	79.5	81.4	81.2	91.8	90.8
		DROP	3-shot, F1	56.4	67.6	(85.0)	16.5	21.4	33.1	10.7	47.5	56.3	39.1	53.7	78.7	87.2
Math	Mathematical Reasoning	GSM8K	8-shot CoT	53.6	56.4	(80.7)	12.1	10.8	15.6	60.2	47.3	59.4	78.2	76.3	92.1	92.1
		MATH	4-shot CoT	2.6	3.7	(34.3)	0.0	0.0	0.0	10.2	8.5	15.6	32.0	15.0	38.6	34.9
Coding	Coding Problems	HumanEval	0-shot, pass@1	10.7	12.7	–	0.0	0.1	0.6	24.2	29.3	57.6	53.9	80.0	66.3	66.4
		MBPP	3-shot, pass@1	44.8	58.0	–	4.6	7.6	11.9	67.3	70.2	77.0	82.3	98.0	85.5	85.7
Multilingual	Multi-subject Test	AGIEval-ZH	few-shot	31.7	37.9	–	23.6	23.9	28.0	41.4	38.6	39.3	41.9	38.4	56.5	56.7
		C-Eval	5-shot	10.7	38.0	–	5.5	1.6	20.7	50.3	44.5	49.7	51.8	48.5	69.2	69.1
	Mathematical Reasoning	MGSM	8-shot CoT	3.6	4.0	(72.2)	2.4	5.1	7.4	7.9	22.9	33.7	53.5	53.7	82.2	68.7
	Question Answering	TyDi QA	1-shot, F1	12.1	18.8	(40.3)	5.7	3.7	9.3	14.3	12.5	16.3	21.2	25.1	31.3	31.2
Safety	Truthfulness	TruthfulQA	1-shot	51.0	59.4	–	21.4	5.4	21.7	54.2	47.8	52.2	57.4	61.4	79.5	79.7
	Toxicity	RealToxicityPrompts ↓	0-shot	14.8	15.0	–	15.6	16.1	14.1	15.0	15.0	9.6	8.0	7.7	7.9	7.9

Benchmark of capability tests in various LLMs (Bastian, 2023)



Intelligenz & Fähigkeiten

AI-Psychometrics

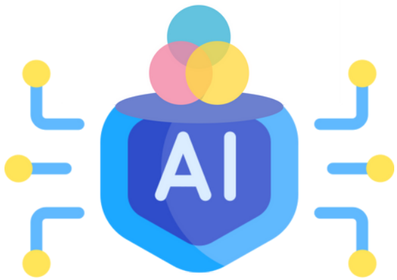
Menschliche **Persönlichkeitstests** oft ungeeignet

Persönlichkeitstests wie die **Big Five** sind für LLMs nicht geeignet, da die Fragen oft einen **persönlichen sozialen Kontext** voraussetzen.

Ergebnisse für GPT3.5 im Big Five Personality Test

Factor	Factor label	Raw score	Score percentile
I	Extroversion		54
II	Emotional stability		78
III	Agreeableness		17
IV	Conscientiousness		67
V	Intellect/Imagination		18

Apte, 2023
Big Five Personality test. https://openpsychometrics.org/tests/IPIP-BFFM/



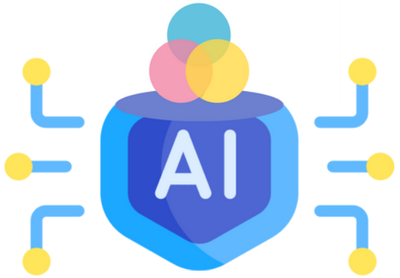
Persönlichkeit

Charit

AI-Psychometrics

Einstellungen & Meinungen von LLMs messen

Zum Messen von Meinungen und Einstellungen von Menschen lassen sich Stichproben aus der Gesellschaft ziehen. Zum Messen von Einstellungen einzelner LLMs lassen sich Stichproben an Instanzen ziehen.



Einstellungen

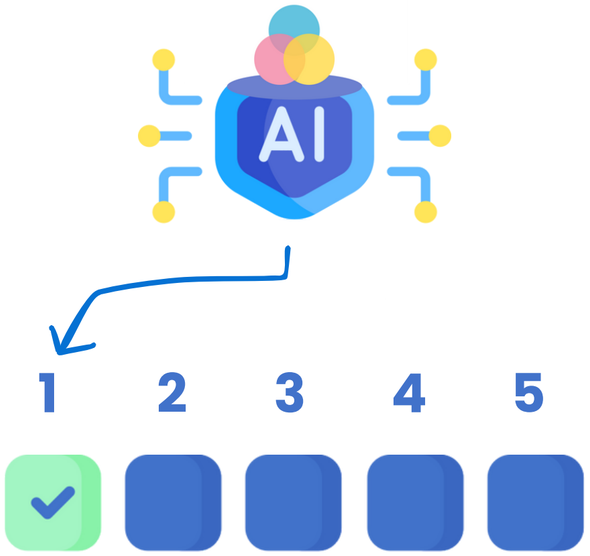
Relevanz

LLMs werden zunehmend als diskrete **Entscheidungsträger** eingesetzt

LLMs sind besonders attraktiv für das Treffen kleiner diskreter und kontextbezogener Entscheidungen wie:

Scoring-Tasks oder **Klassifizierungs-Aufgaben.**

Es ist somit relevant zu verstehen, wie das Entscheidungsbild von LLMs mit Veränderung der Aufgabenstellung variiert und sich zum Menschen unterscheidet.



Charit

Anwendung

Bewertung von Produktideen

*prompt and items in Appendix

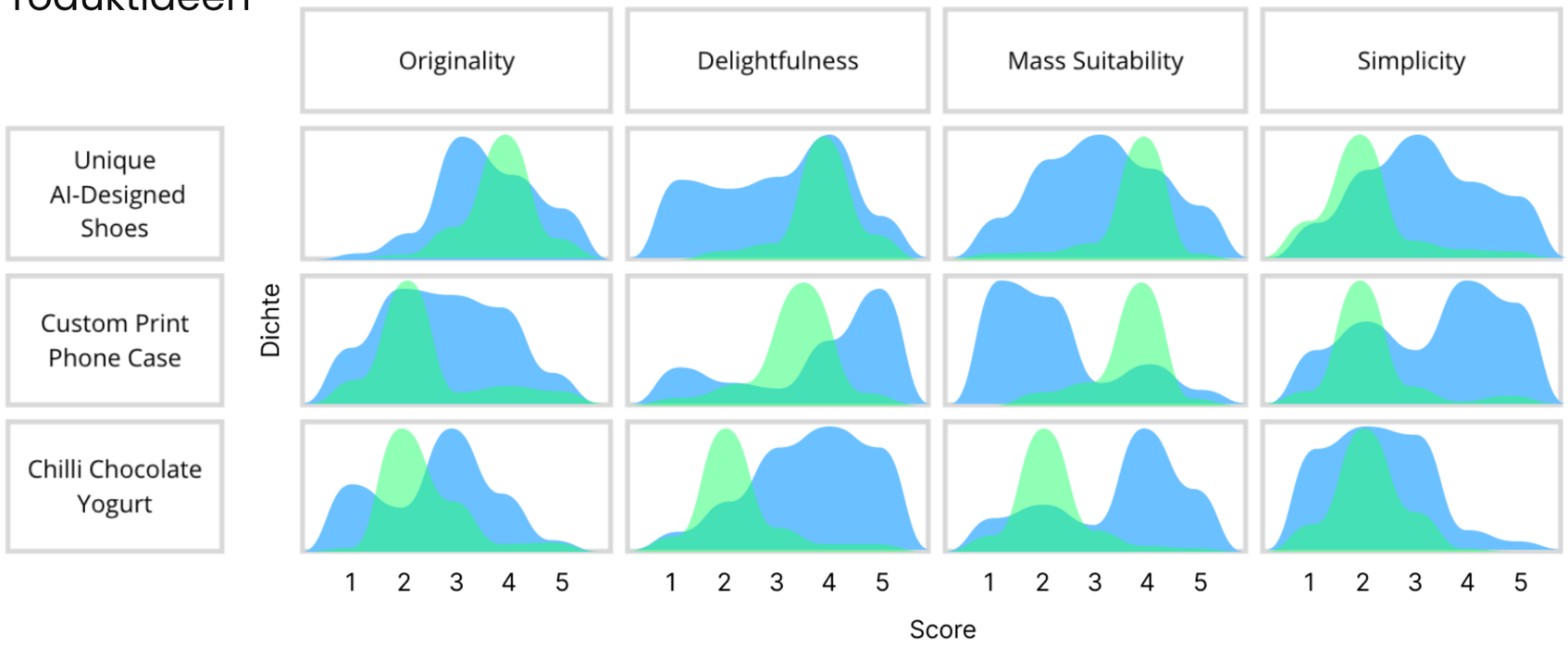
Exploration

N.Mensch = 135

N.KI = 135

Mensch

LLM



Modell = Vanilla gpt-4-0125-preview

Skala = 1 (very low) - 5 (very high)

Implikationen

Die Meinungen von LLMs können stark von der menschlichen **Wirklichkeit** abweichen

- Gezieltes und induktives Prompting von Nöten (klare Bewertungskriterien anstatt vage Konzepte) bei Klassifizierung und Scoring.
- Klare Kriteriendefinition durch Experten im jeweiligen Bereich.
- Kein blindes Vertrauen in deduktive Bewertungsergebnisse von LLMs
- Validieren durch menschliche Stichprobe bei subjektiven Themen von Nöten
- Variation in LLM-Antworten sagen uns etwas über den Non-Determinismus des LLM

Charit

Kontakt

Charit Insights
charit-app.com

Represented by:

Johannes Jamroszczyk
johannes.jamroszczyk@charit-app.com

Sandra Heuer
sandra.heuer@charit-app.com

Charit

Quellen

Apte, J. (2023, February 7). I gave ChatGPT the big five personality test. Here is what happened next. . .. *Medium*. https://medium.com/@jayant91089/i-gave-chatgpt-the-big-five-personality-test-here-is-what-happened-next-9b5ed01bef15

Bastian, M. (2023, October 2). GPT-4 “crushes” other LLMs according to new benchmark suite. *THE DECODER*. https://the-decoder.com/gpt-4-crushes-other-llms-according-to-new-benchmark-suite/

Big Five Personality test. (n.d.). https://openpsychometrics.org/tests/IPIP-BFFM/

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2022). AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Sage Journals*. https://doi.org/10.31234/osf.io/jv5dt

Charit

Appendix A

Zero Shot Task - Prompt Body

Rate the following <concept> based on its {category} on a
Scale from 1 (Very Low) to 5 (Very High).
<Concept>: {concept}

Accepted output: Single Integer [1,2,3,4,5]
Unaccepted output: [Context, Reasoning, Text, Code Block]

scoring iterations per combination : 135

modell = Vanilla gpt-4-0125-preview
temperature=0.7
top_p=0.2
top_k=2

Permutated Containers

categories : ["Originality", "Delightfulness", "Mass suitability", "Simplicity"]

concepts : ["Unique AI-Designed Shoes", "Custom Print Phone Case", "Chilli Chocolate Yogurt"]

Charit

Appendix B

Questionnaire - Human

1A-C: Rate the following concept based on its **originality** on a Scale from 1 (Very Low) to 5 (Very High).
Concept: ["Unique AI-Designed Shoes", "Custom Print Phone Case", "Chilli Chocolate Yogurt"]**Repeated for all concepts*

2A-C: Rate the following concept based on its **delightfulness** on a Scale from 1 (Very Low) to 5 (Very High).
Concept: ["Unique AI-Designed Shoes", "Custom Print Phone Case", "Chilli Chocolate Yogurt"]*Repeated for all concepts

3A-C: Rate the following concept based on its **mass suitability** on a Scale from 1 (Very Low) to 5 (Very High).
Concept: ["Unique AI-Designed Shoes", "Custom Print Phone Case", "Chilli Chocolate Yogurt"]*Repeated for all concepts

4A-C: Rate the following concept based on its **simplicity** on a Scale from 1 (Very Low) to 5 (Very High).
Concept: ["Unique AI-Designed Shoes", "Custom Print Phone Case", "Chilli Chocolate Yogurt"]*Repeated for all concepts

Answer: Forced Choice on an anchored Likert scale. [Very Low 1 | 2 | 3 | 4 | 5 Very High]