LREC-COLING 2024

GPTEval: A Survey on Assessments of ChatGPT and GPT-4

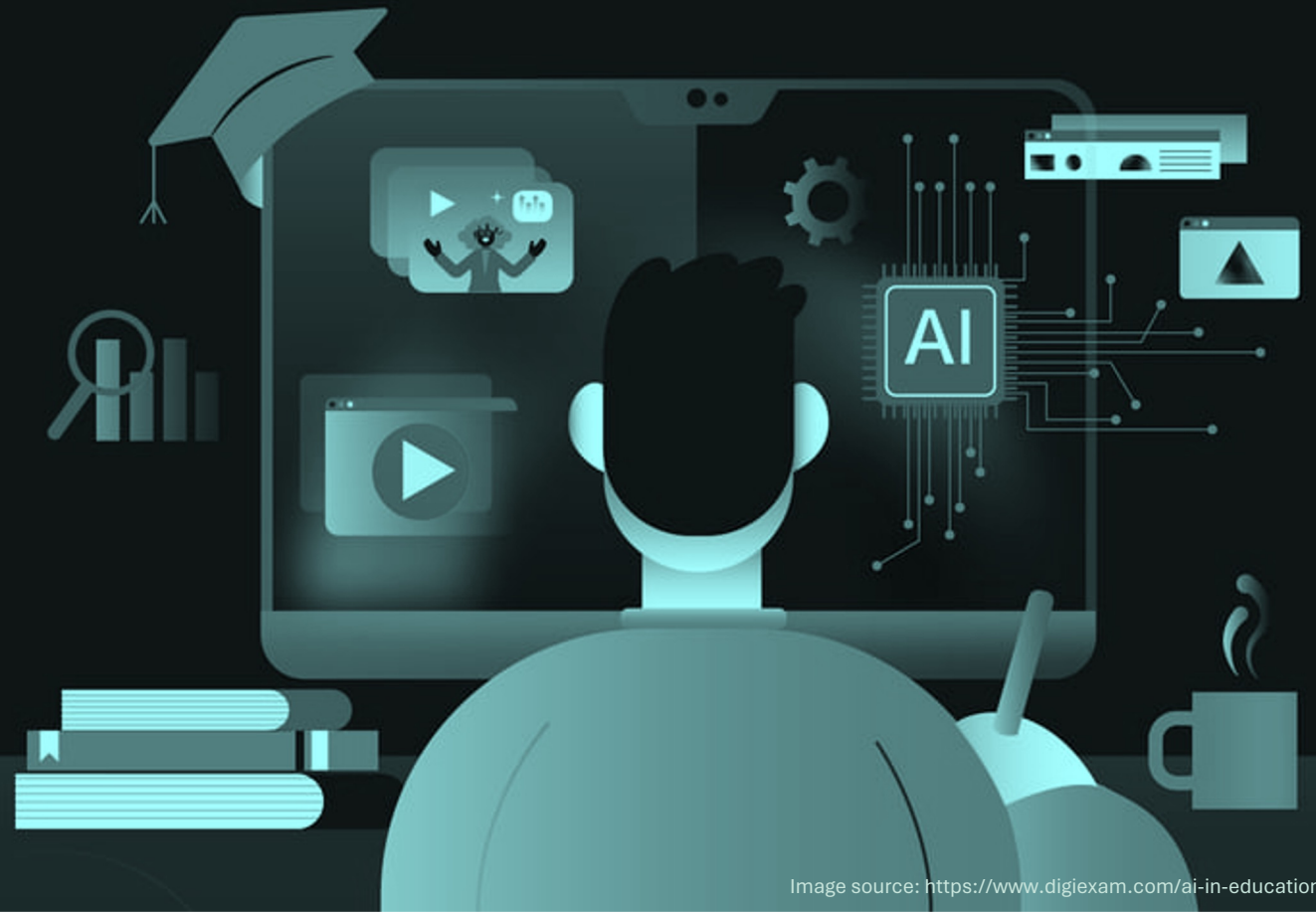


Image source: https://www.digiexam.com/ai-in-education

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria

rui.mao@ntu.edu.sg; g.chen@ccnu.edu.cn; xulang001@e.ntu.edu.sg; f.guerin@surrey.ac.uk; cambria@ntu.edu.sg

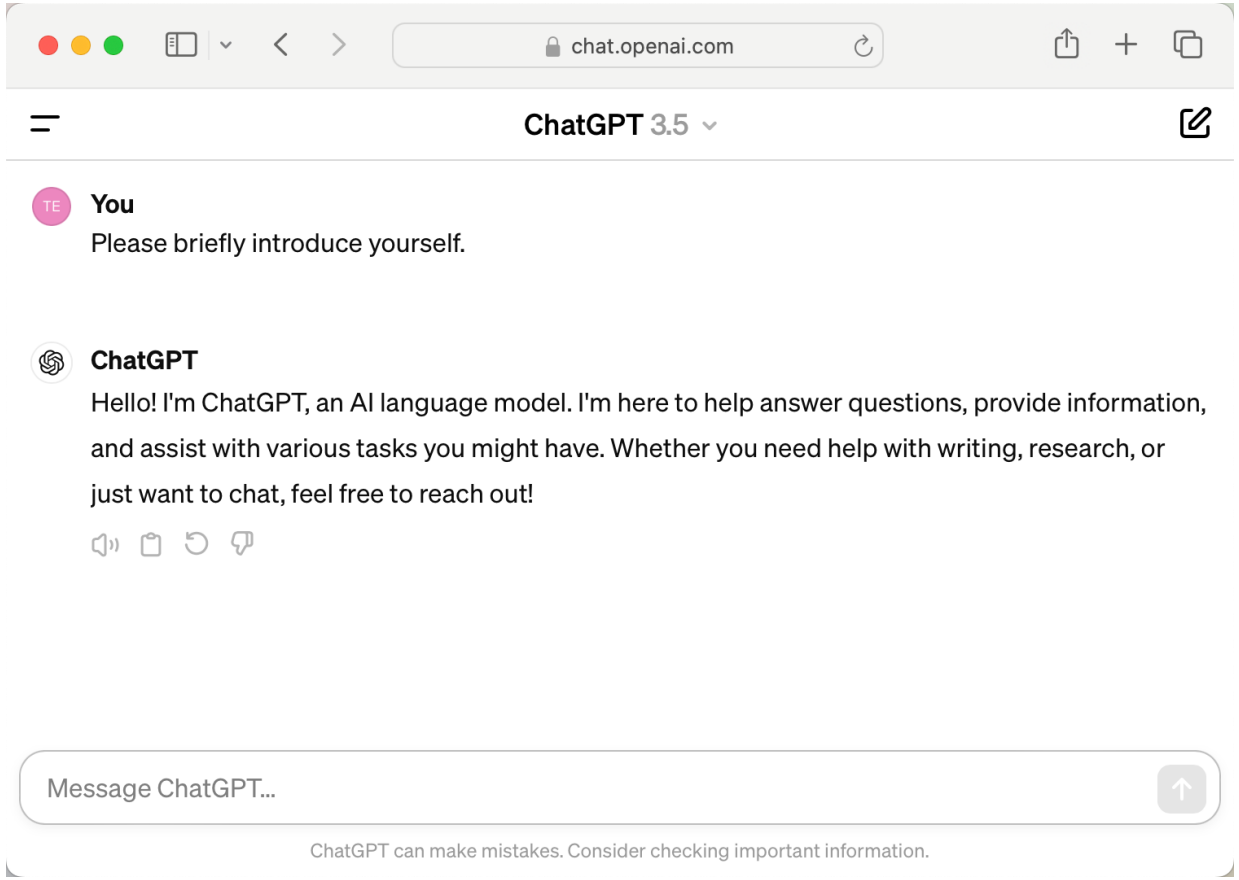
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

CENTRAL CHINA NORMAL UNIVERSITY

UNIVERSITY OF SURREY

ChatGPT and GPT-4

- Developed by OpenAI;
- Large language models;
- Language understanding and generation;
- Broad knowledge;
- Adapted to different use cases.



Motivation of This Survey

- Comprehensively understand the pros and cons of ChatGPT and GPT-4:
    - Language proficiency;
    - Scientific knowledge;
    - Ethical considerations.
- Analyze the limitation of current evaluation methods.

Coverage

Abbr.	Meaning	Abbr.	Meaning
ACI	Actual Causality Inferring	Natu.S	Natural Science
Asp.E	Aspect Extraction	NER	Named Entity Recognition
Asp.PD	Aspect Polarity Detection	NLI	Natural Language Inference
Caus.D	Cause Discovery	Offe.D	Offensiveness Detection
Caus.R	Causal Reasoning	OKEP	Ophthalmic Knowledge Assessment Program
CBQA	Complex Boolean Question Answering	Opin.E	Opinion Extraction
CCSE	Chinese Civil Service Examination	Overa.	Overall
CEG	Causal Explanation Generation	Pass.R	Qassage Re-ranking
CFA	Chinese-Featured Ability	Pers.D	Personality Detection
Chem.Exam	Chemistry Exam	Phys.Exam	Physics Exam
Com.R	Commonsense Reasoning	PoS	Part-of-Speech Tagging
Coun.R	Counterfactual Reasoning	Prag.	Pragmatic Processing Tasks
CS Exam	Computer Science Exam	Prof.A	Professional Ability
D2TE	Data-to-Text Generation Evaluation	QA	Question Answering
DGE	Dialogue Generation Evaluation	RE	Relation Extraction
Diagno.	Diagnosis	Reas.	Reasoning
Dial.S	Dialogue Systems	Sarc.D	Sarcasm Detection
DLR	Dialogue Logic Reasoning	Sem.	Semantic Processing Tasks
Edu.Ling.	Linguistic Quality in Education	Sent.A	Sentiment Analysis
Edu.Sci.	Scientific Accuracy in Education	Sent.R	Sentiment Ranking
EE	Event Extraction	SGE	Story Generation Evaluation
Emoj.P	Emoji Prediction	Soci.S	Social Science
Emot.Re	Emotion Recognition	SpamD	Spam Detection
Emot.Ra	Emotion Ranking	Stan.D	Stance Detection
Enga.A	Engagement Analysis	STEM	Science, Technology, Engineering, and Mathematics
Eval.	Evaluation	Subj.D	Subjectivity Detection
Form.S	Formal Science	Suic.D	Suicide Detection
Gene.	Generation Tasks	Summ.	Summarisation
Huma.	Humanity	Synt.	Syntactic Processing Tasks
Humo.R	Humour recognition	T-Lo.R	Task-Oriented Logical Reasoning
Info.R	Information Retrieval	T-Sy.R	Task-oriented Symbolic Reasoning
LeetCo.	LeetCode	T.Acc	Tweet Annotation Accuracy
LID	Language Identification	T.Agre	Tweet Annotation Agreement
Ling.A	Linguistic Acceptability	T2SQL	Text-to-SQL
Logi.R	Logical Reasoning	Toxi.D	Toxicity Detection
LSAT	Law School Admission Test	TSE	Text Summarisation Evaluation
Mach.T	Machine Translation	TUCE	Test of Understanding of College Economics
Med.Kno.	Medical Knowledge	USMLE	United States Medical Licensing Examination
Medi.S	Medical Science	WBA	Well-being Analysis
Misi.D	Misinformation Detection	WSD	Word Sense Disambiguation
MVQE	Medical and Vocational Qualification Examinations		

- 49 assessment papers;
- From Jan 2023 to Aug 2023;
- 81 evaluation tasks.

Findings: Language proficiency

- Classical NLP tasks:
  - Dialogue;
  - Generation;
  - Affective computing;
  - Information retrieval;
  - GPT as human annotator.
- Reasoning:
  - Logic reasoning;
  - Commonsense reasoning;
  - Causal reasoning;
  - Psychological reasoning;
  - Task-orientated reasoning;
  - Natural language inference.

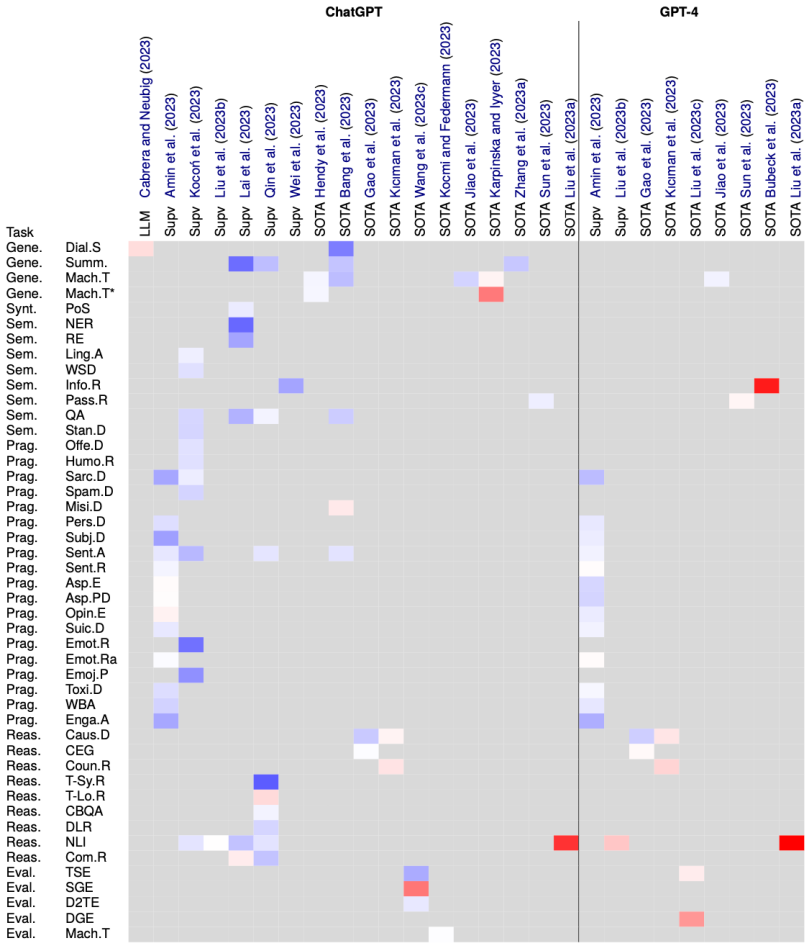


Figure 1. ChatGPT and GPT-4 performance on linguistic and reasoning tasks, compared to baseline models.

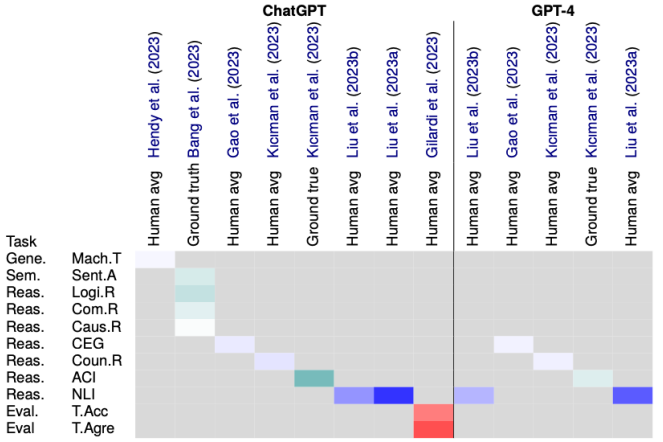
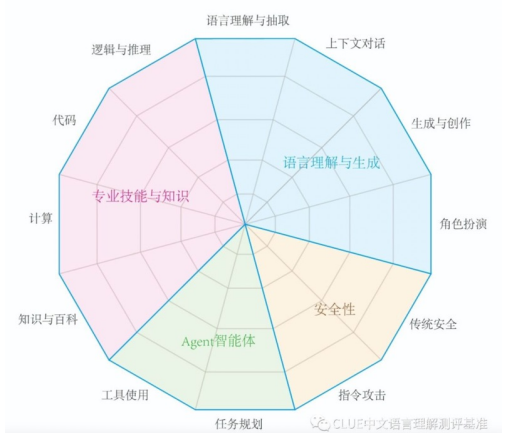


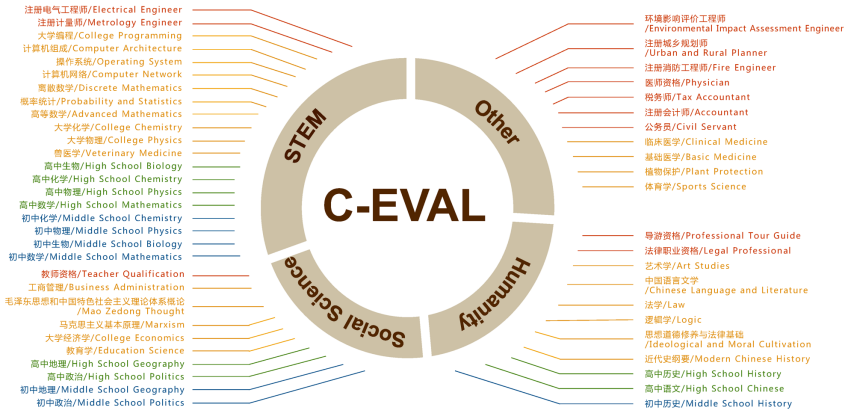
Figure 2. ChatGPT and GPT-4 performance on linguistic and reasoning tasks, compared to humans or ground truth.

Findings: Language proficiency

- Multilingualism:
  - Multilingual NLP tasks.
  - Chinese linguistic test;



SuperCLUE (Xu et al., 2023)



C-Eval (Huang et al., 2024)

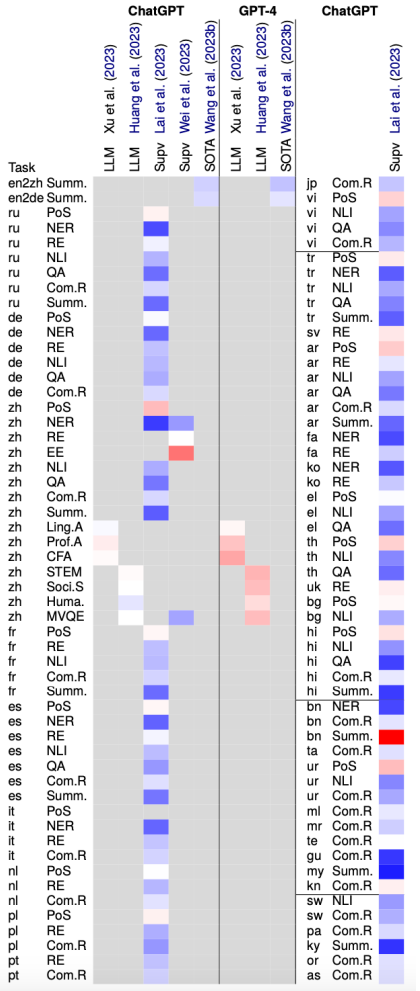


Figure 3. ChatGPT and GPT-4 performance on multi-lingual tasks, compared to baseline models.

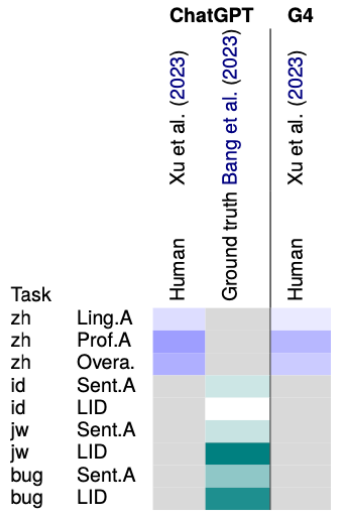


Figure 4. ChatGPT and GPT-4 performance on multi-lingual tasks, compared to humans or ground truth.

Xu, L., *et al*. (2023). SuperCLUE: A benchmark for foundation models in Chinese. (https://github.com/CLUEbenchmark/SuperCLUE)

Huang, Y., *et al*. (2024). C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, *36*.

Findings: Scientific knowledge

- Formal science:
  - Mathematics;
  - Computer science.
- Nature science:
  - Physics;
  - Chemistry;
  - Medicine.
- Social science:
  - Education;
  - Law;
  - Economics.

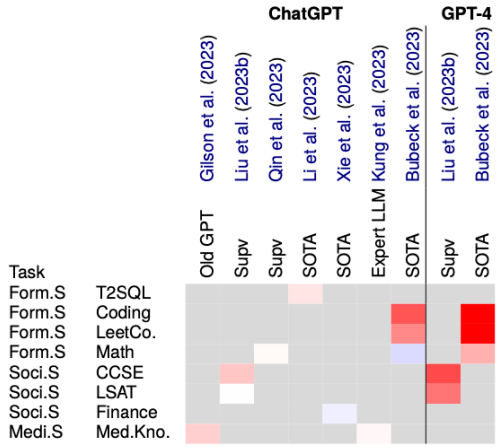


Figure 5. ChatGPT and GPT-4 performance on scientific knowledge, compared to baselines.

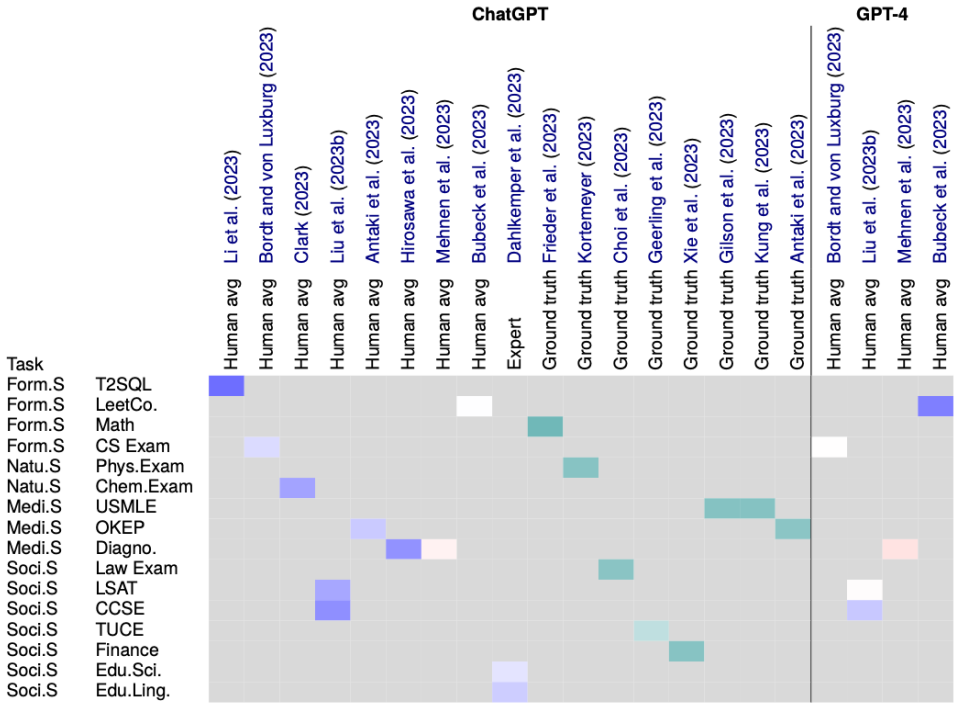


Figure 6. ChatGPT and GPT-4 performance on scientific knowledge, compared to humans or ground truth.

Findings: Ethical considerations

- Fairness
  - ChatGPT's performance in non-English languages was notably poorer (Seghier, 2023; Yong et al., 2023b).
  - It significantly reduced gender and race bias compared to previous versions (Zhuo et al., 2023).
- Robustness
  - ChatGPT's ability to handle noisy data, outliers, and SQL injection was found to be inadequate (Zhou et al., 2023; Ye et al., 2023; Wang et al., 2023; Peng et al., 2023).
- Reliability
  - Improvements in fact-based Q&A were not observed compared to earlier versions (Zhou et al., 2023). There are concerns about potential fabrications in scientific articles (Athaluri et al., 2023) and legal cases (Deroy et al., 2023).
- Toxicity
  - ChatGPT was found to be vulnerable to prompt injections through role-playing (Derner and Batistič, 2023; Zhou et al., 2023).

Mohamed L Seghier. (2023). ChatGPT: Not all languages are equal. *Nature*, 615(7951):216–216.

Yong, Z. X., et al. (2023). Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 43-63).

Zhuo, T. et al. (2023). Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867.*

Ye, W., et al. (2023). Assessing hidden risks of LLMs: an empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235.*

Wang, J., et al. (2023). On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Peng, X., et al. (2022). On the Security Vulnerabilities of Text-to-SQL Models. *arXiv preprint arXiv:2211.15363*.

Athaluri, S. A., et al. (2023). Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4).

Deroy, A., et al. (2023). How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?. *arXiv preprint arXiv:2306.01248*.

Derner, E., & Batistič, K. (2023). Beyond the safeguards: Exploring the security risks of chatgpt. *arXiv preprint arXiv:2305.08005*.

Discussion: Comparing GPTs vs. humans

- GPTs may not perform as well as experts or humans in NLP tasks with sufficient training data;
- However, they excel in scientific knowledge compared to earlier models;
- GPTs can outperform humans in specialist knowledge but struggles with easy tasks;
- GPTs' pre-training focused on “what is right”, neglecting “what is wrong”;
- Contrasting “thinking fast” with “thinking slow”;
- Altering the prompt wording can change the output.

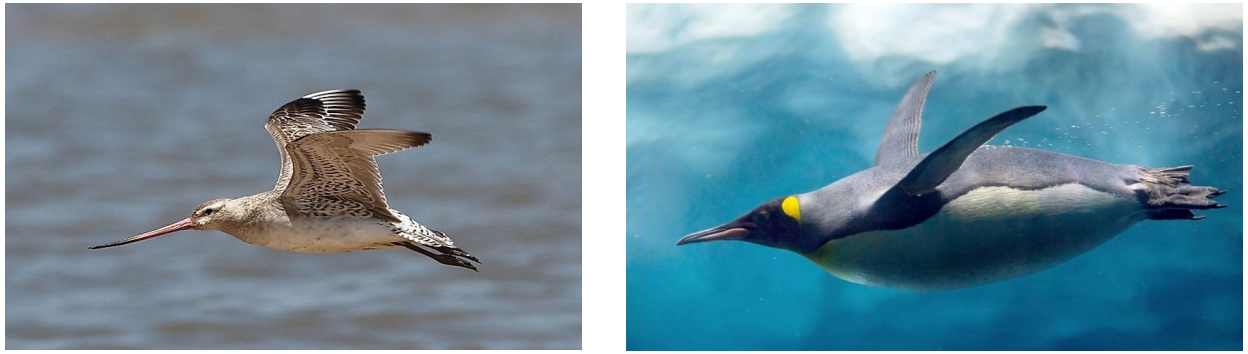


Figure 7. While ChatGPT can easily learn to predict “bird” when prompted with “if an animal has wings and can fly, it is likely a”, it struggles to learn from typical corpora that predicting “penguin” is incorrect, because of the absence of explicitly learning from negative samples.

Image source: https://www.worldatlas.com/articles/highest-flying-birds.html https://www.abc.net.au/education/curious-kids-do-penguins-fly-underwater/13943166

Discussion: GPT evaluation

- Different studies from different periods about the same task are not fully comparable;
- Data leakage may make the assessment unfair;
- The design of prompts highly influences the results;
- The factors that matter in previous NLP evaluations are still valid;
- Some evaluation tasks lack either objective criteria or large-scale benchmarks.

Discussion: Ethics

- Human perception about the reliability of ChatGPT's output can be misled by its seemingly scientific language style;
- RLHF may be misled by human-biased feedback, e.g., system gaming, positive reward cycles, and more;
- The concept preference of ChatGPT may exhibit potential cognitive biases.

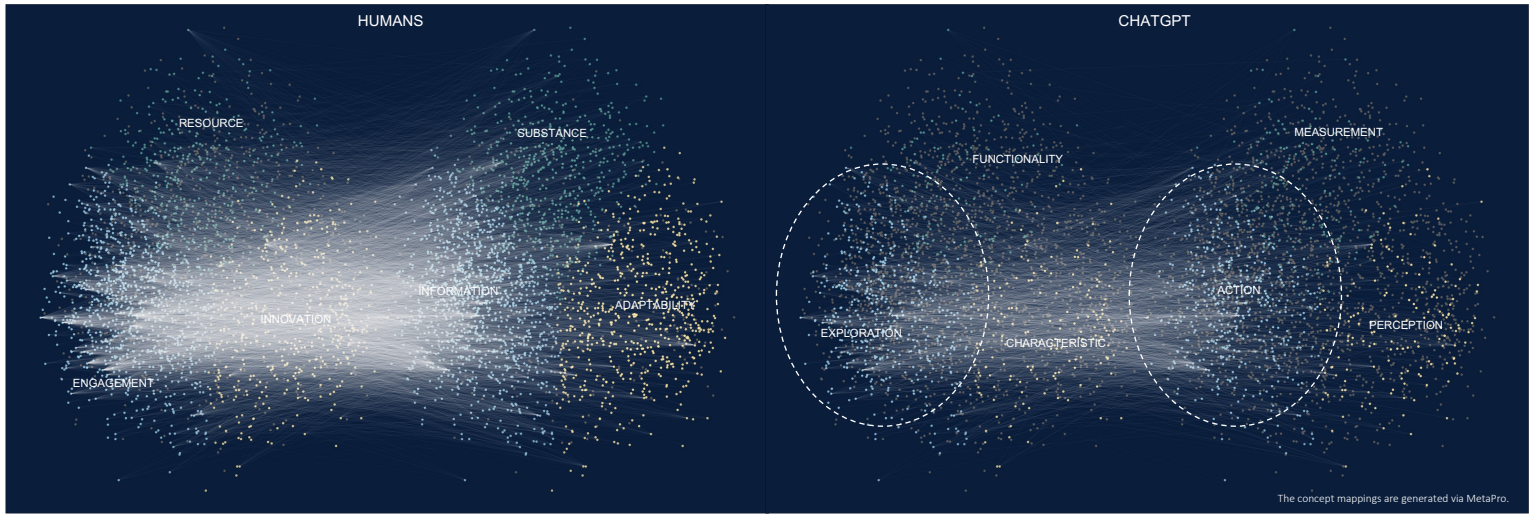


Figure 8. The concept mapping patterns between humans (left) and ChatGPT (right) from Mao et al. (2024). Each cluster on the left represents target concepts, while on the right, the cluster represents source concepts. Bright and grey dots denote activated and unactivated concepts, respectively. The capitalized terms represent key activated concepts within a cluster.

Mao, R., *et al*. (2024). A Comparative Analysis of Metaphorical Cognition in ChatGPT and Human Minds.

Recommendations

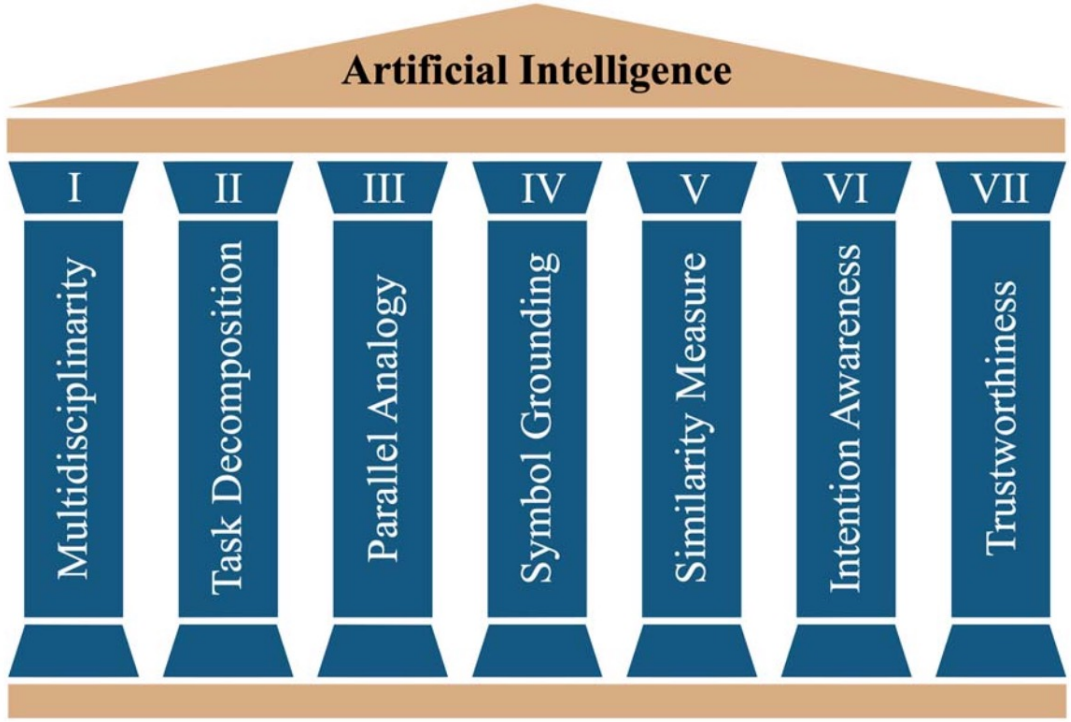


Figure 9. Seven pillars for the future of AI (Cambria et al., 2023).

- Task-agnostic evaluation is desirable;
- Fundamental research is still valuable;
- AI-generated content should be regulated;
- The future of LLMs may need advances in learning paradigms.

Cambria, E., et al. (2023). Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6), 62-69.