GeneTuring tests GPT models in genomics

Wenpin Hou[1,†] **and Zhicheng Ji**[2,†]

[1]Department of Biostatistics, The Mailman School of Public Health, Columbia University, New York City, NY, USA
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
[†]Corresponding author. E-mail: wh2526@cumc.columbia.edu; zhicheng.ji@duke.edu

ABSTRACT

Generative Pre-trained Transformers (GPT) are powerful language models that have great potential to transform biomedical research. However, they are known to suffer from artificial hallucinations and provide false answers that are seemingly correct in some situations. We developed GeneTuring, a comprehensive QA database with 600 questions in genomics, and manually scored 10,800 answers returned by six GPT models, including GPT-3, ChatGPT, and New Bing. New Bing has the best overall performance and significantly reduces the level of AI hallucination compared to other models, thanks to its ability to recognize its incapacity in answering questions. We argue that improving incapacity awareness is equally important as improving model accuracy to address AI hallucination.

Main

Generative Pre-trained Transformers (GPT), such as GPT-2[1], BioGPT[2], BioMedLM[3], GPT-3[4], ChatGPT[5], and New Bing[6] (Figure 1a), are powerful language models that use transformers, a type of deep learning model, to learn from massive amounts of training data and produce texts similar to human speech. While GPT models have shown enormous potential in various tasks such as question answering, music and poet composition, programming, and translation, there is a deep concern about artificial intelligence (AI) hallucinations[7–10], where GPT models give confident answers that are against the truth. One likely cause of AI hallucination is that GPT models tend to randomly choose an answer when there is a lack of information in the training data. AI hallucination is particularly detrimental to the rigor of biomedical research. For example, the wrong answers given by the GPT models may generate false hypotheses that result in a huge waste of resources in validation experiments or may mislead the judgement of clinical practitioners and result in unforeseen consequences in patient healthcare. However, the exact degree of AI hallucination is little understood for GPT models, and it is largely unknown if GPT models can be readily deployed for biomedical research on a large scale.

In this study, we aim to quantitatively test the performance of GPT models in genomics. We choose genomics as a subfield of biomedical research for several reasons. First, it is possible to design evaluation tasks that may or may not find support in the training data. For example, aligning a randomly chosen DNA sequence to the human genome is unlikely to be supported by the training data. This is important to understand the behavior of GPT models and AI hallucination when the GPT models are given tasks that fall within or outside the scope of the training data and cannot be done by other Q&A databases such as PubMedQA[11] that are likely to be fully covered by the training data. Second, an accurate GPT model may substantially benefit genomics research by reducing the time spent for locating and retrieving information from a reliable source, which is often time-consuming for interdisciplinary researchers with limited knowledge of genomics. Our evaluation results provide a guideline for whether the current GPT models can generate reliable answers for researchers. Third, designing evaluation tasks with straightforward questions and unambiguous gold standards is relatively easy in genomics.

To this end, we developed GeneTuring, a comprehensive question-and-answer (Q&A) database, to evaluate the performance of GPT models in genomics (Methods). GeneTuring contains twelve modules with a total of 600 question-answer pairs, and these modules can be grouped into four categories. These twelve modules represent tasks that are commonly seen in genomics research. Figure 1b shows the categories and names of the twelve modules, example pairs of questions and true answers, example answers from ChatGPT, and scores of the example answers. To accommodate models based on GPT-2, the same questions were converted to tasks of sentence completion (Methods).

We used GeneTuring to test six GPT models: GPT-2[1], BioGPT[2], BioMedLM[3], GPT-3[4], ChatGPT[5], and New Bing[6] (Figure 1a). GPT-2, GPT-3, and ChatGPT are all GPT models developed by the company openAI. New Bing is a GPT model jointly developed by the company Microsoft and openAI and is based on GPT architectures. GPT-2 is an earlier GPT model while GPT-3, ChatGPT, and New Bing represent the latest GPT models. BioGPT and BioMedLM are both GPT models relying on GPT-2 architecture but are trained on biomedical literature instead of documents from general sources. For GPT-3, ChatGPT,

and New Bing, each question was repeatedly given three times, and the three responses were recorded. For GPT-2, BioGPT, and BioMedLM, the question was given once and the top three answers returned by the model were recorded. For each of the 10,800 responses, we then manually evaluated if the model correctly understood the question, or if the model acknowledged that it lacked the capacity to answer the question (we call this "incapacity awareness"), or if the model provided a relevant answer to the question. If a relevant answer was provided, a numeric score between 0 and 1 was given according to a scoring mechanism (Methods). All Q&A pairs used in the evaluation, responses from the GPT models, and the scores can be found in Supplementary Table 1.

We first ask if the GPT models correctly understood the questions (Figure 2a). GPT-3 models (GPT-3, ChatGPT, and New Bing) provide relevant answers to almost all questions and show superior ability in understanding questions. In comparison, GPT-2 models (GPT-2, BioGPT, and BioMedLM) fail to understand questions in many tasks. The inferior performance of GPT-2 models is potentially caused by the sentence completion mechanism or smaller model capacities. Interestingly, BioGPT and BioMedLM, which used the same GPT-2 architecture but were trained on specialized biomedical corpus, significantly improve their performance of question understanding over the original GPT-2 model.

We next examined how likely GPT models generate false answers, given that the question was correctly understood. Figure 2b shows the proportion of questions that receive zero scores for the answers, or AI hallucination. New Bing significantly reduces AI hallucination compared to other models. One important reason is that New Bing is able to recognize and report its incapacity of answering the question and does not provide a definite answer (Figure 2c-d). For example, New Bing recognizes that it lacks the ability to answer all sequence analysis questions, thus do not generate any false answer. This ability to know the boundary of its capacity, or incapacity awareness, is especially important in addressing AI hallucination since there will inevitably be questions that fall outside the knowledge base provided by the training data and are unlikely for the GPT models to answer correctly. Admitting the limitation of capacity is better than generating random and false answers which are misleading and could cause further damage. In comparison, GPT models other than New Bing show certain degrees of incapacity awareness but still have a significant level of AI hallucination in many tasks.

Finally, we report the overall score, which is the average score of all answers within each module for each GPT model (Figure 2e). A zero score is given when the question is not correctly understood or when the model recognizes its incapacity. New Bing has the best overall performance and performs well in all tasks unrelated to SNPs or sequence analysis. A potential reason that New Bing does not excel in all tasks is that the SNP and DNA sequence information may not be covered in the training data. Using specialized training corpus (BioGPT and BioMedLM) and using a more advanced architecture (GPT-3, ChatGPT, and New Bing) both improve the performance of the basic GPT-2 model, but a more advanced architecture clearly results in a larger improvement. Thus, it is reasonable to expect that a model trained using the latest GPT architecture and with a specialized biomedical corpus will lead to an even better performance in the GeneTuring tests.

In summary, we show that the latest development of GPT models has significantly reduced but hasn't fully eliminated AI hallucination. New Bing, which has the best overall performance, still generates a small proportion of AI hallucination. Thus, researchers should be cautious when applying New Bing in their own research. To address the issue of AI hallucination, the improvement of incapacity awareness is equally important to the improvement of model accuracy and capacity.

Methods

GeneTuring compilation and scoring criteria

GeneTuring contains 12 modules. Each module contains 50 pairs of questions and answers designed for models built on GPT-3 architecture. Geneturing also converts each pair of question and answer into a sentence completion task with the same meaning to accommodate models built on GPT-2 architectures. The details of how the 12 modules were compiled and scored numerically are discussed as follows. Note that only relevant and definite answers from GPT models were scored numerically. The outputs of GPT models were not scored numerically if they didn't directly answer the questions or acknowledge their limitations in answering the questions.

Gene name extraction

This module tests the ability of GPT models to extract gene and gene product names from a given sentence. We downloaded the test set of BioCreative II challenge Task 1A: Gene Mention Tagging (BC2GM) from the biocreative website[12]. The test set contains 5000 pairs of sentences and corresponding names of genes and gene products mentioned in the sentence. 50 pairs were randomly selected from the 5000 pairs. To create question-answering tasks, a question was created by pasting "What are the gene and protein names in the sentence: " before the BC2GM sentence and a question mark after the BC2GM sentence. To create sentence completion tasks, a question was created by pasting "The gene and protein names in the sentence " before the BC2GM sentence and " is" after the BC2GM sentence. The set of gene and gene product names provided by the original BC2GM task was used as the gold standard answer. Each sentence may contain zero, one, or multiple gene and gene product names.

A Jaccard index was used to evaluate the performance of an answer given by a GPT model. Denote $\mathbb{A}$ as the set of gene and gene product names answered by a GPT model and $\mathbb{B}$ as the gold standard set of gene and gene product names. The Jaccard index is calculated as $\frac{|\mathbb{A}\cap\mathbb{B}|}{|\mathbb{A}\cup\mathbb{B}|}$, where $|.|$ indicates the cardinality of a set. In case both the gold standard and the GPT model report no gene or gene product name ($\mathbb{A} = \emptyset$ and $\mathbb{B} = \emptyset$), the Jaccard index is set to be 1.

Gene alias

This module tests the ability of GPT models to find the official gene symbol of a given gene alias. The information of official gene names and their aliases for human protein-coding genes was downloaded from NCBI website[13]. 50 genes with at least one alias were randomly selected. For each gene, one alias was randomly selected to produce the question. To create question-answering tasks, a question was created by pasting "What is the official gene symbol of " before the alias and a question mark after the alias. To create sentence completion tasks, a question was created by pasting "The official gene symbol of gene " before the alias and " is" after the alias. The official gene symbol was used as the gold standard.

Similar to the previous section, a Jaccard index was used to compare the official gene symbol given by a GPT model with the gold standard and provide a score. The Jaccard index is used since GPT models may give multiple official gene symbols as an answer to some questions.

Gene name conversion

This module tests the ability of GPT models to convert the Ensembl gene names into gene symbols. The gene annotation gtf file for human GRCh38 genome was downloaded from GENCODE website[14]. 50 protein-coding genes were randomly selected. The Ensembl gene name and gene symbol of each gene were recorded. To create question answering tasks, a question was created by pasting "Convert " before the Ensembl gene name and " to official gene symbol." after the Ensembl gene name. To create sentence completion tasks, a question was created by pasting "The official gene symbol of " before the Ensembl gene name and " is" after the Ensembl gene name. The gene symbol was used as the gold standard.

A score of 1 is given if the gene symbol given by the GPT model agrees with the gold standard and a score of 0 is given if they disagree.

Gene location

This module tests the ability of GPT models to find which chromosome a gene is located on. The gene annotation gtf file for human GRCh38 genome was downloaded from GENCODE website[14]. Genes with multiple locations on the genome were removed. 50 genes were randomly selected. For each gene, its gene symbol and the name of the chromosome it is located on were recorded. To create question answering tasks, a question was created by pasting "Which chromosome is " before the gene symbol and " gene located on human genome?" after the gene symbol. To create sentence completion tasks, a question was created by pasting " gene is located on human genome chromosome" after the gene symbol. The chromosome name was used as the gold standard.

A score of 1 is given if the chromosome name given by the GPT model agrees with the gold standard and a score of 0 is given if they disagree.

SNP location

This module tests the ability of GPT models to find which chromosome a single nucleotide polymorphisms (SNP) is located on. The SNP locations were downloaded from NCBI website[13]. 50 SNPs were randomly selected. For each SNP, its SNP symbol and the name of the chromosome it is located on were recorded. To create question answering tasks, a question was created by pasting "Which chromosome does SNP " before the SNP symbol and " locate on human genome?" after the SNP symbol. To create sentence completion tasks, a question was created by pasting "SNP " before the SNP symbol and " is located on human genome chromosome" after the SNP symbol. The chromosome name was used as the gold standard.

A score of 1 is given if the chromosome name given by the GPT model agrees with the gold standard and a score of 0 is given if they disagree.

Gene SNP association

This module tests the ability of GPT models to find which gene a SNP is associated with. The SNP-gene association information was downloaded from NCBI website[13]. SNPs that are associated with zero or more than one genes were removed. 50 SNPs were randomly selected. For each SNP, its SNP symbol and the gene symbol it is associated with were recorded. To create question answering tasks, a question was created by pasting "Which gene is SNP " before the SNP symbol and " associated with?" after the SNP symbol. To create sentence completion tasks, a question was created by pasting "The name of the gene associated with SNP " before the SNP symbol and " is" after the SNP symbol. The gene symbol was used as the gold standard.

A score of 1 is given if the gene symbol given by the GPT model agrees with the gold standard and a score of 0 is given if they disagree.

Protein-coding genes

This module tests the ability of GPT models to identify if a gene codes protein. Gene information was downloaded from NCBI website[13]. 50 human genes that are protein-coding genes, non-coding RNA (ncRNA) genes or pseudogenes were randomly selected. For each gene, its gene symbol and gene type were recorded. To create question answering tasks, a question was created by pasting "Is " before the gene symbol and " a protein-coding gene?" after the gene symbol. To create sentence completion tasks, a question was created by pasting "Regarding if the gene codes a protein, " before the gene symbol and " is" after the gene symbol. A binary value of if the gene is a protein-coding gene was used as the gold standard.

A score of 1 is given if the GPT model agrees with the gold standard in deciding if the gene is a protein-coding gene, and a score of 0 is given if they disagree.

Gene disease association

This module tests the ability of GPT models to identify genes associated with a specific disease. Gene-disease association information was downloaded from OMIM website[15]. 50 diseases were randomly selected and all genes associated with each disease were recorded. To create question answering tasks, a question was created by pasting "What are genes related to " before the disease name and a question mark after the disease name. To create sentence completion tasks, a question was created by pasting "The name of the gene related to " before the disease name and " is" after the disease name. The set of genes associated with the disease was used as the gold standard.

The answer given by a GPT model is scored as the proportion of gold standard genes mentioned by the answer.

Gene ontology

This module tests the ability of GPT models to identify gene ontology (GO) terms enriched in a set of genes. GO information for biological process was downloaded from MSigDB[16]. 50 GO terms were randomly selected and all genes associated with the GO term were recorded. To create question answering tasks, a question was created by pasting "What is the enriched gene ontology term associated with " before the list of genes and a question mark after the list of genes. To create sentence completion tasks, a question was created by pasting "The enriched gene ontology term associated with " before the list of genes and " is" after the list of genes. The name of the GO term was used as the gold standard.

A score of 1 is given if one of the GO terms answered by a GPT model fully matches with the gold standard. A score of 0.5 is given if one of the GO terms answered by a GPT model partially matches with the gold standard. A score of 0 is given if none of the GO terms answered by a GPT model fully or partially matches with the gold standard.

TF regulation

This module tests the ability of GPT models to judge if a transcription factor activates or represses the expression of a gene. The information of transcription factor regulating genes in human was downloaded from Trrust database[17]. 50 pairs of transcription factors and genes that have an activation or repression relationship were randomly selected. To create question answering tasks, a question was created by pasting "Does transcription factor ", the name of transcription factor," activate or repress gene ", the name of the gene, and "?" together. To create sentence completion tasks, a question was created by pasting "The regulatory relationship between transcription factor ", the name of transcription factor," and gene ", the name of the gene, and " is" together. A binary value of if the transcription factor activates the gene was used as the gold standard.

A score of 1 is given if the GPT model agrees with the gold standard in deciding if the relationship between the transcription factor and the gene is activation or repression, and a score of 0 is given if they disagree.

DNA sequence aligment to human genome

This module tests the ability of GPT models to identify on which chromosome a DNA sequence aligns to the human genome. The DNA sequence information of the human genome was downloaded from the Bioconductor[18] package BSgenome.Hsapiens.UCSC.hg38. Only autosomes and sex chromosomes were retained. Regions with "N" in DNA sequence were removed. To select one genomic region, we randomly selected a chromosome, a starting position within the range of the selected chromosome, and a length between 100 to 150 base pairs. The DNA sequence was acquired for the selected chromosome, starting position and length. The process was repeated 50 times to generate 50 DNA sequences and their chromosome names. To create question answering tasks, a question was created by pasting "Align the DNA sequence to the human genome:" before the DNA sequence. To create sentence completion tasks, a question was created by pasting "The DNA sequence " before the DNA sequence and " is on the human genome chromosome" after the DNA sequence. The chromosome name was used as the gold standard.

A score of 1 is given if the chromosome name given by the GPT model agrees with the gold standard and a score of 0 is given if they disagree.

DNA sequence aligment to multiple species

This module tests the ability of GPT models to identify the species from which a DNA sequence come. The DNA sequence information of the human, mouse, rat, chicken, zebrafish, worm, and yeast genomes was downloaded from the Bioconductor[18] packages BSgenome.Hsapiens.UCSC.hg38, BSgenome.Mmusculus.UCSC.mm10, BSgenome.Rnorvegicus.UCSC.rn5, BSgenome.Ggallus.UCSC.galGal6, BSgenome.Drerio.UCSC.danRer11, BSgenome.Celegans.UCSC.ce11, and BSgenome.Scerevisiae.UCSC. Only autosomes and sex chromosomes were retained. Regions with "N" in DNA sequence were removed. To select one genomic region, we randomly selected one of the seven species, a chromosome within the selected species, a starting position within the range of the selected chromosome, and a length between 100 to 150 base pairs. The DNA sequence was acquired for the selected species, chromosome, starting position and length. The process was repeated 50 times to generate 50 DNA sequences and their species names. To create question answering tasks, a question was created by pasting "Which organism does the DNA sequence come from:" before the DNA sequence. To create sentence completion tasks, a question was created by pasting "The name of the species where DNA sequence " before the DNA sequence and " comes from is" after the DNA sequence. The species name was used as the gold standard.

A score of 1 is given if the species name given by the GPT model matches exactly with the gold standard. A score of 0.5 is given if the GPT model provides a correct superset of the gold standard but does not match exactly with the gold standard. A score of 0 is given if the GPT model answer is neither a superset nor matches with the gold standard.

GPT models

GPT-2, BioGPT, and BioMedLM models were implemented using the interface provided by huggingface[19]. GPT-3 with "text-davinci-003" model engine was implemented using the standard pipeline provided on openai.com. ChatGPT January 31 version was directly accessed via https://chat.openai.com/. New Bing was directly accessed via a Microsoft Edge Dev web browser.

Acknowledgments

Z.J. was supported by the National Institutes of Health under Award Number 1U54AG075936-01.

Author contributions

All authors conceived the study, conducted the analysis, and wrote the manuscript.

Competing interests

All authors declare no competing interests.

a

	GPT-2	BioGPT	BioMedLM	GPT-3	ChatGPT	New Bing
Architecture	GPT-2	GPT-2	GPT-2	GPT-3	GPT-3.5	Prometheus
Initial release	Feb 2019	Oct 2022	Dec 2022	June 2022	Nov 2022	Feb 2023
Parameter number	1.5 billion	347 million	2.7 billion	175 billion	117-345 million *	~100 trillion **
Training data	General corpus	Biomedical corpus	Biomedical corpus	General corpus	General corpus	General corpus

* Answer given by ChatGPT ** Answer given by New Bing

b

Category	Module	Example question	Example true answer	Example ChatGPT answer	Example score
Nomenclature	Gene name extraction	What are the gene names in the sentence: "This deletion disrupts the PU.1 Ets domain."?	PU.1, Ets domain	The gene name in the sentence is "PU.1".	0.5
	Gene alias	What is the official gene symbol of SNAT6?	SLC38A6	The official gene symbol for SNAT6 is SLC38A6 (solute carrier family 38, member 6).	1
	Gene name conversion	Convert ENSG00000149476 to official gene symbol.	TKFC	The official gene symbol for ENSG00000149476 is BRCA2 (breast cancer 2, early onset).	0
Genomic location	Gene location	Which chromosome is RGS16 gene located on human genome?	chr1	The RGS16 gene is located on human chromosome 1.	1
	SNP location	Which chromosome does SNP rs1430464868 locate on human genome?	chr13	SNP rs1430464868 is located on chromosome 9 in the human genome.	0
	Gene SNP association	Which gene is SNP rs983419152 associated with?	LINC02055	The SNP rs983419152 is associated with the gene CHFR (Checkpoint with Forkhead and Ring Finger domains).	0
Functional analysis	Protein-coding genes	Is UCKL1-AS1 a protein-coding gene?	FALSE	UCKL1-AS1 is a non-coding RNA gene, specifically an antisense long non-coding RNA (lncRNA)…...	1
	Gene disease association	What are genes related to Pseudohypoparathyroidism Ic?	GNAS	Pseudohypoparathyroidism type 1c (PHP1c) is a genetic disorder caused by mutations in the GNAS gene……	1
	Gene ontology	What is the gene ontology term associated with PARP3, APLF, TFIP11, HMGB1, RAD51, XRCC1?	regulation of dna ligation	These genes are associated with DNA repair or DNA metabolic process in the Gene Ontology database.	0.5
	TF regulation	Does transcription factor RELA activate or repress gene IL10?	Activation	The transcription factor RELA (p65 subunit of NF-kappaB) has been shown to activate the expression of the IL10 (interleukin 10) gene.	1
Sequence alignment	DNA sequence aligment to human genome	Align the DNA sequence to the human genome:TGGGCTCAAGTGATCATA……	chr7	As a language model AI, I do not have the capability to align a DNA sequence to the human genome……	NA
	DNA sequence aligment to multiple species	Which organism does the DNA sequence come from:CGTACACCATTGGTGC……	yeast	The organism from which the DNA sequence comes cannot be determined based solely on the DNA sequence……	NA

Figure 1. a, A table showing the six GPT models' names, architectures, initial release dates, numbers of parameters, and types of training data. **b,** A table showing the categories and names of the twelve modules, example pairs of questions and true answers, example answers from ChatGPT, and scores of the example answers.

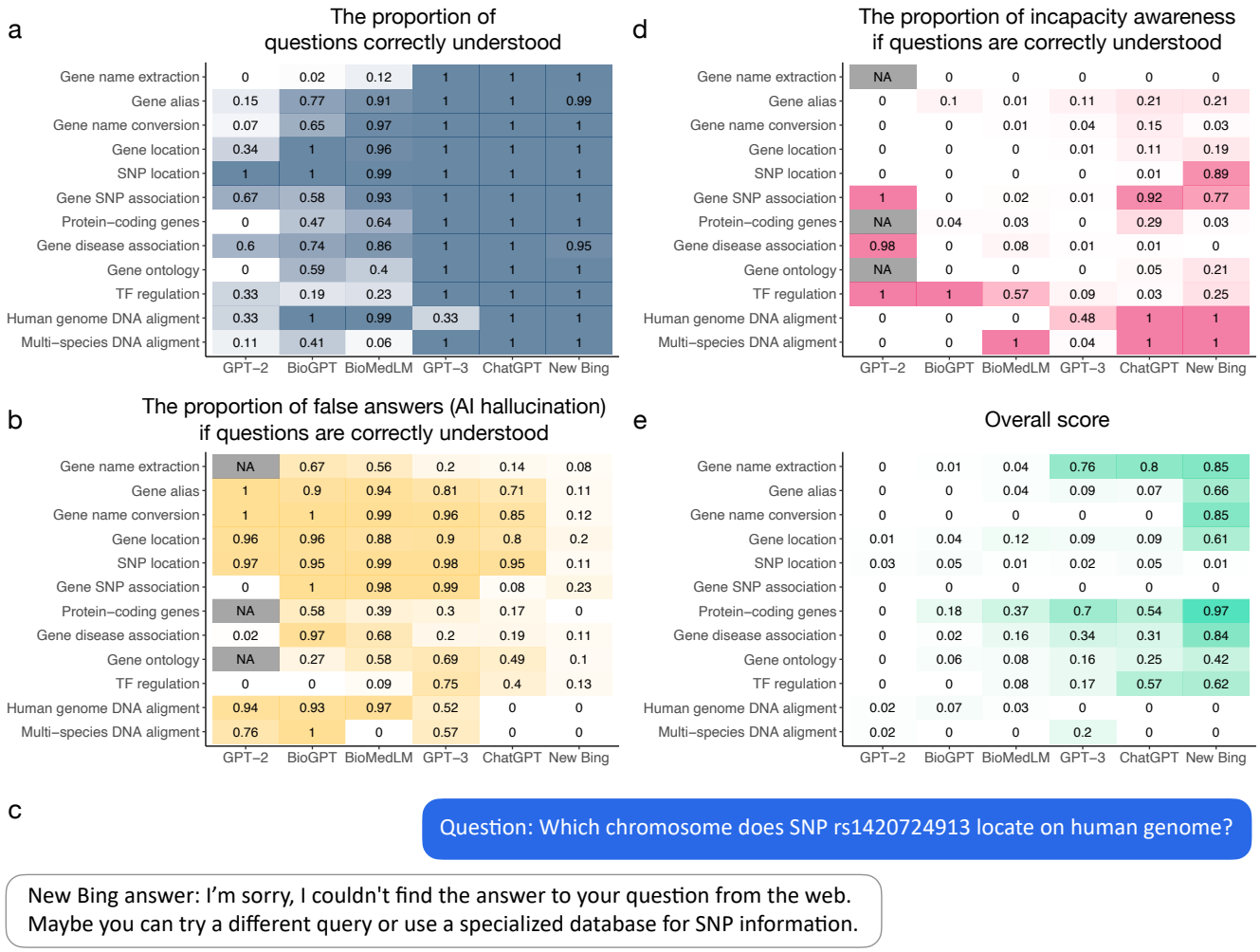

Figure 2. a, The proportion of questions that are correctly understood by each method for each module. **b,** The proportion of answers with zero scores (AI hallucinations), calculated only for questions that are correctly understood. **c,** An example New Bing answer where it recognizes its incapacity in answering the question. **d,** The proportion of questions where the model acknowledges its incapacity in answering the question (incapacity awareness), calculated only for questions that are correctly understood. **e,** The overall score, which is the average of scores across all answers in that module.

References

1. Radford, A. *et al.* Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
2. Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinforma.* **23** (2022).
3. Venigalla, A., Frankle, J. & Carbin, M. Biomedlm: a domain-specific large language model for biomedical text. *https://www.mosaicml.com/blog/introducing-pubmed-gpt* .
4. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
5. van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. Chatgpt: five priorities for research. *Nature* **614**, 224–226 (2023).
6. The new bing. https://www.bing.com/new. Accessed: 2023-03-11.
7. Alkaissi, H. & McFarlane, S. I. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus* **15** (2023).
8. Bang, Y. *et al.* A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
9. Sallam, M. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv* 2023–02 (2023).
10. Eysenbach, G. *et al.* The role of chatgpt, generative language models, and artificial intelligence in medical education: A conversation with chatgpt and a call for papers. *JMIR Med. Educ.* **9**, e46885 (2023).
11. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* (2019).
12. Smith, L. *et al.* Overview of biocreative ii gene mention recognition. *Genome biology* **9**, 1–19 (2008).
13. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **35**, D5–D12 (2007).
14. Frankish, A. *et al.* Gencode reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766–D773 (2019).
15. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. Omim. org: leveraging knowledge across phenotype–gene relationships. *Nucleic acids research* **47**, D1038–D1043 (2019).
16. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1**, 417–425 (2015).
17. Han, H. *et al.* Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research* **46**, D380–D386 (2018).
18. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with bioconductor. *Nat. methods* **12**, 115–121 (2015).
19. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45 (2020).