EvalGPT: A Visual Analytic Framework for Enhancing Trust in Large Language Models

Xu Yang Clemson University Clemson, U.S.A xyang2@g.clemson.edu

Yiheng Liang Northwestern Polytechnical University Xi'an, China yhliang15@gmail.com

Le Liu[∗] Northwestern Polytechnical University Xi'an, China leliu304@gmail.com

Lianwei Wu Northwestern Polytechnical University Xi'an, China wlw@nwpu.edu.cn

Xiaoru Yuan Peking University Beijing, China xiaoru.yuan@pku.edu.cn

ABSTRACT

The use of large language models (LLMs) in education is significantly increasing due to their ability to generate quick and easy answers to questions. However, their potential for producing misinformation must be noticed. Educators face the challenge of discerning the reliability of answers generated by these models. To address this challenge, we present EvalGPT, a visualization approach for evaluating, revealing, and explaining the uncertainty of content generated by LLMs. EvalGPT first simulates how students ask these questions via various prompts and quantitatively measures the uncertainty of their LLM-generated answers. Second, a group-boxplots-based view is developed to convey the uncertainty and enable interactive recognition of anomalies in generated answers. Third, a text analysis view is provided to support the detailed explanation of these anomalies. For our demonstration, we sourced questions from a widely used visualization textbook and utilized EvalGPT to assess the answers provided by ChatGPT.

VLDB Workshop Reference Format:

Xu Yang, Yiheng Liang, Le Liu, Lianwei Wu, and Xiaoru Yuan. EvalGPT: A Visual Analytic Framework for Enhancing Trust in Large Language Models. VLDB 2024 Workshop: Relational Models.

VLDB Workshop Artifact Availability:

The source code, data, and/or other artifacts have been made available at [URL_TO_YOUR_ARTIFACTS.](URL_TO_YOUR_ARTIFACTS)

1 INTRODUCTION

With the rapid development in the field of artificial intelligence, the application of large language models (LLMs) is gradually permeating various domains, causing a significant impact on the field of education [\[2,](#page-5-0) [15,](#page-6-0) [27\]](#page-6-1). LLMs quickly and interactively generate answers to questions, offering an extremely easy and fast way

Proceedings of the VLDB Endowment. ISSN 2150-8097.

for knowledge summarization, text summarization, essay writing, and other practical applications. Consequently, they have found widespread use and have been changing the way of learning. For instance, students may ask an LLM for definitions or explanations of vital concepts instead of reading a textbook. However, due to the well-known hallucination problems, the contents generated by LLMs are not a hundred percent accurate, potentially leading to bias and misconceptions [\[24\]](#page-6-2). Even worse, it is hard to explain the reasons and mechanisms behind inducing this inaccuracy because of the LLMs' black box problem. Therefore, educators face the significant challenge of preventing students from obtaining incorrect or biased knowledge and information when learning by using LLMs.

This challenge can be mitigated by responsible adoption of LLMs through comprehensive awareness of their uncertainty. To do so, we introduce the EvalGPT, a visual analytic framework for enhancing trust in LLMs in education. This framework begins with generating questions and answers using LLMs to mimic how students utilize them to learn knowledge. Then, we developed a metric to quantify the uncertainty of the LLM contents. Subsequently, we developed a visualization featuring a grouped-box-plot view for portraying content uncertainty, a filtering view to extract abnormal contents, a detailed view for displaying the text of extractions, and a semantic correspondence view for reasoning the abnormals. This paper runs a case study using ChatGPT as an example of LLM and a visualization textbook to source questions and their ground truth answers for demonstration purposes.

In summary, this research makes the following contributions:

- We presented an educational-specific visual analytic framework for both quantitatively and qualitatively investigating the uncertainty of contents generated by LLMs.
- We conduct a case study to demonstrate our framework's usefulness in the visualization field by adopting ChatGPT as an example of LLMs.
- We publish a data set of visualization questions and their ground truth answers and answers generated by ChatGPT, supporting further research in this direction.

[∗]Corresponding author.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit<https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org.](mailto:info@vldb.org) Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

2 RELATED WORK

2.1 LLMs for Education

LLMs have permeated various aspects of human life, with one of the most prominent examples being the GPT series developed by OpenAI (e.g., GPT-3[\[3\]](#page-5-1), ChatGPT[\[18\]](#page-6-3), GPT-4[\[4\]](#page-5-2), GPT-4o[\[18\]](#page-6-3)). A growing body of research has started investigating the application of LLMs in education and, more importantly, evaluating their potential impacts on this field. For instance, Fütterer et al. [\[10\]](#page-6-4) analyzed Twitter data since ChatGPT was released and summarized an overview of global perceptions and reactions to ChatGPT regarding education. Choi et al. [\[7\]](#page-6-5) utilized ChatGPT to generate answers on four actual Law School exams, demonstrating ChatGPT's ability to pass these exams with an average performance. Studies investigating the Graduate Record Examination (GRE) [\[19\]](#page-6-6), the Uniform Bar Exam [\[13\]](#page-6-7), and the United States Medical Licensing Examination practice [\[17\]](#page-6-8) have reported similar results. Chen et al. [\[6\]](#page-5-3) utilized GPT-4 to complete class assignments featuring various visualization tasks. The results indicate that GPT-4 scored 80%, and there is a 30% chance that teaching fellows could not distinguish between GPTand human-generated content. The authors concluded there is a need to redesign visualization education by incorporating LLMs.

2.2 Uncertainty Visualization of LLMs

Comprehensive evaluations of the uncertainty of LLMs are crucial for the responsible application of these models and have gained increasing attention. For example, Ajith et al. [\[1\]](#page-5-4) developed the InstructEval evaluation suite, covering multiple LLMs and tasks to compare the impact of different instructions on the generated results. Strobel et al. [\[21\]](#page-6-9) developed PromptIDE, allowing users to try various prompts and visualize their performance. Hämäläinen et al.[\[11\]](#page-6-10) tested whether GPT-generated content could be distinguished from human-generated answers. Wang et al.[\[23\]](#page-6-11) introduced a mechanism to evaluate the content generated by LLMs for crowdsourcing tasks, and they concluded potentially maliciously may lead to untrustable results that are hard to distinguish. These studies indicate an urgent need to investigate the uncertainty of LLMs-generated content.

Visualization techniques have been proven to be effective in assessing complex uncertainty [\[5,](#page-5-5) [8,](#page-6-12) [22\]](#page-6-13). Recently, Zhao et al.[\[26\]](#page-6-14) evaluated the effectiveness of visualizing uncertainty in AI modelgenerated results through comparative experiments. They emphasized the significance of cognitive accessibility of the visualization technique in enhancing people's trust and promoting appropriate reliance on the model.

3 DESIGN FRAMEWORK

This research developed a visual analytic framework called Eval-GPT to quantitatively and qualitatively evaluate the uncertainty of LLM-generated content. It encompasses three main components: question-answer pairs generation, uncertainty quantification, and visualization. For demonstration purposes, this paper used Chat-GPT as an example of LLMs and used the book titled Visualization Analysis and Design, which is a widely-adopted visualization textbook [\[16\]](#page-6-15), to extract questions and their ground truth answers.

3.1 Generate Question-and-Answer Pair Ensembles

Our framework begins by generating an ensemble of questionand-answer pairs to assess the model uncertainty. As a demonstration, we identified questions from a visualization textbook [\[16\]](#page-6-15). We picked this textbook not only because it is widely adopted but also because each subsection is well organized to explain a specific question or concept, as exemplified in Figure [2.](#page-3-0) This makes it convenient for us to form a list of questions by simply extracting the title of the subsections. We extracted all subsections of this textbook, ending up with 84 domain-specific questions that students may encounter while learning in a visualization class. The following are a few examples:

- In the field of data visualization, why have a human in the loop?
- In the field of data visualization, why use interactivity?
- In the field of data visualization, what are data types?

The standard answers to these questions, generated from the subtitles, are just the contents of their corresponding subsections. To make the answers more concise and clear, we entrusted ChatGPT with simplifying the contents without losing critical information. Four visualization experts meticulously evaluated the output, ensuring its accuracy and reliability.

To mimic the situation in which students ask the same question in different ways, we further tasked ChatGPT with generating variations of the presentation of individual questions. An example of an original question and its two variations are demonstrated as follows:

- Original Question: In the field of data visualization, why have a human in the loop?
- Variation 1: What is the importance of human involvement in the field of data visualization?
- Variation 2: How does human presence impact the field of data visualization?

Subsequently, we asked ChatGPT to answer the variations of these questions, and the generated answers were further analyzed to evaluate their accuracy compared to the standard answers, revealing the uncertainty of the LLM.

3.2 Uncertainty Quantification

To quantitatively evaluate the uncertainty of an LLM, we first developed a metric to measure one answer's similarity:

$$
dis_answer_i = similarity(answer_0, answer_i)
$$
 (1)

where answer₀ denotes the standard textbook answer, answer_i represents the ChatGPT-generated answer to the *i*th variant of the corresponding question, and dis_answer_i means the similarity between the two answers.

The $similarity()$ function is the average of four components. First, it uses the BERT model's last_hidden_state output as the text's embedding. It calculates the cosine similarity between these embeddings to measure the semantic similarity between the original text and its variations. BERT[\[9\]](#page-6-16), a pre-trained NLP model based on the Transformer architecture, converts input words or sentences into continuous vector representations. These embeddings capture semantic information and can be used for semantic similarity

Figure 1: The visualization developed in this research. It consists of 4 components: a group-box-plots view to present the uncertainty distribution (top left), a filtering view to extract abnormals (top right), a detailed view to display the questionand-answer data of selected abnormals (bottom left), and an edge-bundling view for explaining the selected abnormals (bottom-right).

calculations. The definition of cosine similarity is shown in Eq. [2:](#page-2-0)

$$
Cosine_sim(A, B) = \frac{A \cdot B}{\|A\| \|B\|},
$$
\n(2)

where, A and B are embedding vectors, \cdot represents the dot product of vectors, $||A||$ and $||B||$ represent the norms of the vectors A and B, respectively.

The second component calculates the cosine similarity using the BERT model's pooler_output output as the text's embedding. In addition to the first components, utilizing $pooler_output$ output provides an aggregated representation of the entire sequence for subsequent task processing.

The third component calculates the Jaccard similarity [\[12\]](#page-6-17) between two keyword sets extracted from two answers, respectively. Jaccard similarity is defined in Eq. [3,](#page-2-1)

Jaccard_sim(s1, s2) =
$$
\frac{|s1 \cap s2|}{|s1 \cup s2|}
$$
, (3)

where s1 and s2 are two keyword sets. The Jaccard similarity is from 0 to 1, where 0 indicates that the two sets have no common elements, and 1 indicates that the two sets are identical. Incorporating this method considers the themes of the contents.

To extract keywords, we utilized the KeyBERT technique introduced by Sharma and Li[\[20\]](#page-6-18). It first uses BERT to extract document embeddings to obtain document-level vector representations. Subsequently, word vectors are extracted for N-gram words/phrases, and then cosine similarity is used to find the words most similar to

the document. Finally, the most similar words can be identified as the words that best describe the entire document.

The fourth component calculates the Levenshtein similarity between two keyword sets extracted from two answers [\[14\]](#page-6-19), considering grammatical and structural features. Levenshtein similarity measures the edit distance between two strings, indicating how many insertions, deletions, or replacements are needed to transform one string into another, as shown in Eq. [4,](#page-2-2)

$$
LS_sim(s1, s2) = 1 - \frac{edit_distance(s1, s2)}{max(len, (s1), len(s2))}
$$
(4)

where $edit_distance(s1, s2)$ means the edit distance between sets s1 and s2, where *edit_distance* refers to the minimum number of editing operations required to transform one string into another. Permissible editing operations include insertion, deletion, and substitution. $len(s)$ means the length of set s. Levenshtein similarity ranges from 0 to 1, where 0 indicates that the two strings are completely dissimilar, and 1 indicates that the two are identical.

Finally, we average these four components to measure the difference between the original text and its GPT generations. This comprehensive calculation method, incorporating semantics and keywords, can more comprehensively evaluate text similarity.

To evaluate an LLM's uncertainty reliably, we must generate question-and-pair ensembles of sufficient size to cover the distribution of all possible predictions from the model. To determine this size, we first randomly selected 10 from the 84 original questions. For each of these 10 questions, we subsequently generated 100

Contents

Figure 2: An example demonstrates that this research extracts questions from the contents of the Visualization Analysis and Design book [\[16\]](#page-6-15).

question variations and their corresponding answers using Chat-GPT. We then randomly select n (ranging from 1 to 100) similarity calculation values from the 100 variations of each question. These n values are averaged, resulting in all results for n ranging from 1 to 100. These results are sorted in ascending order, theoretically identifying the convergence point where the mean no longer significantly changes beyond n . Since this method does not consider the inevitable differences in content length and complexity among answers to different questions, we repeat the random selection and mean calculation process for all 10 questions, obtaining 10 means. These 10 means are averaged to obtain a final value, effectively covering the differences in samples with varying lengths and complexities. Plotting these values against n , Figure [3](#page-3-1) illustrates the result of the samplings.

From this graph, it is evident that the result converges when the variation quantity is around 10. Thus, we ultimately determine the quantity of generated variations to be 10.

In addition, we invited four visualization domain experts to assess the 10 generated variations from two folds qualitatively: if the generated question is semantically similar to the original question and if these generated questions can realistically simulate students' questions in a visualization class. The results are shown in Figure [4.](#page-3-2) The table shows that out of 40 responses regarding semantic similarity, 39 was "Yes" and 1 was "No." Regarding whether the questions could simulate the scenario of students' questioning, 37 out of 40 responses were "Yes" and 3 were "No." Therefore, the experts

Figure 3: The results showing the average similarities of answers of 100 variations of 10 randomly selected questions.

unanimously agree that this design scheme reflects the semantic information of the original questions and effectively simulates the scenario of students asking questions after class.

Questions Experts	Q ₁		Q ₂		Q ₃		Q ₄		Q5		lQ6		Q 7		Q8		Q ₉		Q10	
Expert1						v		v	\mathbf{v}					Ÿ		v		v	\mathbf{v}	
Expert2		$\mathbf v$		N	v	$\mathbf v$		٧	v	v	v		N	Ÿ	v	۷	$\mathbf v$	٧	v	
Expert3					$\mathbf v$	v		٧	$\mathbf v$		v		Ÿ	N	\vee	v	\checkmark	Ÿ	v	
Expert4		v			v	$\overline{\mathsf{v}}$	lv	N	\mathbf{v}	$\mathbf v$	v	Ÿ	v	Y	$\mathbf v$	v		v	$\mathbf v$	

Figure 4: The result of experts qualitative assessment. Y means Yes and N means No. The first columns of each question (green) represent the assessment result of the semantic similarity and the second columns of each question (yellow) represent whether this design can simulate the scenario of students asking questions.

3.3 Visualization Design

As demonstrated in Figure [1,](#page-2-3) the visualization developed in this research consists of four components: a group-box-plots view to present the uncertainty distribution (top left), a filtering view to extract abnormals (top right), a details view to display the question and answer data of selected abnormals (bottom left), and an edgebundling view to explain the selected abnormals (bottom right).

3.3.1 Group-box-plots-based View. Our visualization design first contains a group-box-plots view to display the uncertainty distributions of the previously introduced ensemble, as demonstrated by the top left view in Figure [1.](#page-2-3) Specifically, the x-axis represents individual questions extracted from the textbook, while the y-axis represents the similarity values calculated by the method introduced in section [3.2,](#page-1-0) which is referred to as score in the visualization. Each column contains two boxplots. The yellow one depicts the score of questions, which portrays how close the question variations generated by ChatGPT can represent the original question. The blue one shows the core of the answers, which conveys the certainty of GPT-generated answers compared to the standard answer.

3.3.2 Filtering View. The Filtering View (top right in Figure [1\)](#page-2-3) is designed to interactively identify abnormals from the uncertainty distribution displayed by the group-box-plot view. Specifically, it utilizes the DBM/OVS metrics proposed by Wild et al.[\[25\]](#page-6-20) to compare the question-box-plot and answer-box-plot. DBM/OVS refers to dividing the difference between the medians of the two box plots (Difference Between Medians) by the overall visible spread (Overall Visible Spread) to obtain a percentage. This percentage represents the degree of difference between the two box plots: the larger the value, the greater the difference, and vice versa. The filtering view provides a bar chart of the distribution of DBM/OVS so that the user can quickly determine the abnormals in a specific range of DB-M/OVS. The group box plots within this range will be highlighted in the group box plot view.

3.3.3 Details View. After identifying the abnormals, users can directly click on the question of in the group-box-plot view, and all its relevant information, including the scores of the variations of questions, scores of the variations of answers, and the specific text content will be displayed in the details view (bottom left in Figure [1\)](#page-2-3). The keywords corresponding to the text will be highlighted in yellow. This details view provides direct access to the row data.

3.3.4 Edge-bundling View. In addition to highlighting keywords in the original text, we utilized a Hierarchical Edge Bundling plot to illustrate the semantic correspondences represented by keyword similarity. Specifically, the keywords rendered in purple indicate those extracted from the original questions, and the keywords rendered with orangeish represent those extracted from the GPTgenerated variations. Semantically similar keywords are linked together. Users can click the "details" button to zoom in on the plot for a more detailed examination.

4 CASE STUDY

Filtered by DBM/OVS, noticing the significant differences between the box plots of Question_5_3 and its answers, this case study will analyze Question_5_3 in detail. The question's dispersion is significantly greater than that of the answers.

Further analysis reveals no outliers in either box plot. The average score of the question box plot is about 0.83, while that of the answer box plot is about 0.52. Generated variations show moderate correctness but low uncertainty in answers. This study explores why high dispersion in the question does not lead to high dispersion in the answers.

Firstly, we examine the generated question variations. The original question is "In the field of visualization, how to use Marks and Channels?" High-scoring variations include "How can Marks and Channels be employed effectively in visualization?" while lowscoring ones are like "What are some key considerations when employing Marks and Channels in visualization?" Different wordings cause semantic differences, resulting in varied question dispersion. However, high question dispersion does not necessarily increase answer dispersion. Observing the keyword results of the answers provides insights.

All variations connect to Var0, with Var8 having the smallest number of connections, which is 3, indicating low uncertainty in generated answers. Figure [5](#page-5-6) shows keywords "encoding," "marks," and "visual" are frequently connected, while specific terms like "attributes" and "categorical" have fewer connections. This suggests that ChatGPT-generated answers cover common terms but rarely specific ones.

This case shows that high question dispersion does not necessarily lead to high answer dispersion. Textbook answers contain common terms, so generated answers also include these, leading to similar score values but failing to cover specific or rare terms. Thus, answer box plot dispersion is low, and correctness is also low because only common terms are covered.

If the question box plot had high dispersion and high correctness (mean), answer uncertainty (dispersion) would likely be high. A balanced mix of common and specific terms in the textbook answers would require ChatGPT to generate high-quality answers based on questions, increasing answer uncertainty.

In summary, using ChatGPT for educational assistance requires considering each question's context and understanding result variations. Teachers' guidance remains crucial, with ChatGPT as an auxiliary tool.

5 DISCUSSION AND FUTURE WORK

This study has several limitations that need further exploration. Firstly, our research findings, while insightful, are limited by the use of a single textbook as the data source. To broaden the applicability of our research, it is crucial to gather a more diverse dataset spanning multiple fields, a direction that holds the potential to establish a stronger foundation for subsequent research.

Secondly, our current uncertainty measurement metrics, while effective, are based on the average results of four algorithms. To further enhance the precision and reliability of our uncertainty assessments, it is imperative to collaborate with experts in natural language processing to develop more comprehensive measurement metrics.

Furthermore, our study only analyzed keywords extracted from the text content. Future research could introduce additional semantic analysis methods to gain a more comprehensive understanding and analysis of the text content.

Lastly, in terms of visualization design, we should focus on two aspects: firstly, exploring better ways to combine text information visualization with uncertainty visualization to make the final visualization more intuitive and easier to understand, and secondly, conducting comparative experiments to evaluate and validate our proposed design against other existing uncertainty visualization methods to assess their effectiveness and advantages in different application scenarios.

By addressing these points, we aim to enhance this study's scientific rigor and practical value, contributing to the development of the research community in extensive language model uncertainty analysis and visualization.

6 CONCLUSION

This research focuses on evaluating the responses generated by LLMs in educational scenarios. We used ChatGPT as an example and systematically assessed how the model responded to different question formats on the same topic. We developed a visualization tool to help educators make informed decisions.

Figure 5: The connection status of Var0 keywords for Question_5_3 shows that the majority of connections are concentrated on the keywords "encoding," "marks," and "visual"

To begin, we created a process to extract questions and answers from textbooks and then generated various question variants using an LLM. We constructed a specific dataset from one textbook, which can support further research in related fields. Additionally, we developed metrics based on natural language processing algorithms to measure the uncertainty of LLM-generated responses. Furthermore, we developed a visual analysis system that allows educators to intuitively observe and analyze the characteristics and uncertainties of LLM-generated content. Finally, we demonstrated the effectiveness of the visualization through a case study.

Future research includes exploring this framework in various application scenarios and further optimizing methods for measuring and visualizing the uncertainty of LLM-generated content.

ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62102323, Innovation Capability Support Program of Shaanxi(Program No. 2023KJXX-142)

REFERENCES

- [1] Anirudh Ajith, Mengzhou Xia, Ameet Deshpande, and Karthik R Narasimhan. 2023. Instructeval: Systematic evaluation of instruction selection methods. In R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models.
- [2] Zeljana Basic, Ana Banovac, Ivana Kruzic, and Ivan Jerkovic. 2023. Better by you, better than me, chatgpt3 as writing assistance in students essays. arXiv preprint arXiv:2302.04536 (2023).
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.
- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023).
- [5] Hua Cai and Yingzi Lin. 2010. Tuning trust using cognitive cues for better humanmachine collaboration. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 2437–2441.
- [6] Zhutian Chen, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg, and Hanspeter Pfister. 2023. Beyond generating code: Evaluating gpt on a data visualization course. In 2023 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVis). IEEE, 16–21.
- [7] Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. ChatGPT goes to law school. J. Legal Educ. 71 (2021), 387.
- [8] Dan Conway, Fang Chen, Kun Yu, Jianlong Zhou, and Richard Morris. 2016. Misplaced trust: A bias in human-machine trust attribution–In contradiction to learning theory. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 3035–3041.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
- [10] Tim Fütterer, Christian Fischer, Anastasiia Alekseeva, Xiaobin Chen, Tamara Tate, Mark Warschauer, and Peter Gerjets. 2023. ChatGPT in education: global reactions to AI innovations. Scientific reports 13, 1 (2023), 15310.
- [11] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. $1 - 19$.
- [12] Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. New phytologist 11, 2 (1912), 37–50.
- [13] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. Philosophical Transactions of the Royal Society A 382, 2270 (2024), 20230254.
- [14] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, Vol. 10. Soviet Union, 707–710.
- [15] Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnák, and Filip Janus. 2023. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?. In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1. 47–53.
- Tamara Munzner. 2014. Visualization analysis and design. CRC press.
- [17] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023).
- [18] OpenAI. 2024. ChatGPT: A Large Language Model. [https://www.openai.com/](https://www.openai.com/chatgpt) [chatgpt.](https://www.openai.com/chatgpt) Accessed: 2024-06-17.
- [19] OpenAI. 2024. GPT-4. [https://openai.com/research/gpt-4.](https://openai.com/research/gpt-4) Accessed: 2024-06-17.
- [20] Prafull Sharma and Yingbo Li. 2019. Self-supervised contextual keyword and keyphrase retrieval with self-labelling. (2019).
- [21] Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. IEEE transactions on visualization and computer graphics 29, 1 (2022), 1146–1156.
- [22] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns 1, 4 (2020).
- [23] Chaofan Wang, Samuel Kernan Freire, Mo Zhang, Jing Wei, Jorge Goncalves, Vassilis Kostakos, Zhanna Sarsenbayeva, Christina Schneegass, Alessandro Bozzon, and Evangelos Niforatos. 2023. Safeguarding Crowdsourcing Surveys from ChatGPT with Prompt Injection. arXiv preprint arXiv:2306.08833 (2023).
- [24] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022).
- [25] CJ Wild, Maxine Pfannkuch, Matthew Regan, and Nicholas J Horton. 2011. Towards more accessible conceptions of statistical inference. Journal of the Royal Statistical Society Series A: Statistics in Society 174, 2 (2011), 247–295.
- [26] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. 2023. Evaluating the impact of uncertainty visualization on model reliance. IEEE Transactions on Visualization and Computer Graphics (2023).
- [27] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223 (2023).