Faulty scientific questions identification

Farzad Azarmi
fka5196@psu.edu

1 Abstract

This study investigates the application of the BERT-base model with a cyclic learning rate scheduler for evaluating text generation models and classifying faulty and paradoxial scientific questions. The cyclic learning approach dynamically adapts the learning rate, improving optimization and generalization, and achieving robust performance metrics, including a validation accuracy of 83.33% and a test accuracy of 74.94% at epoch 8. Classification analysis revealed high accuracy in structured domains like "Economics," while "Technology" and "Miscellaneous" categories faced challenges due to ambiguity and heterogeneity. Further insights from the confusion matrix and Q-Q plot highlighted cross-domain misclassifications, underscoring the difficulties of handling questions that blur disciplinary boundaries or exhibit subtle paradoxes. These findings emphasize the need for tailored feature engineering, refined data curation, and enhanced model architectures to address category-specific challenges. The study provides valuable methodological and practical insights for improving the evaluation and classification of complex and ambiguous textual data.

2 Introduction

Large Language Models (LLMs) have become the standard for a wide range of machine learning tasks, including text generation, summarization, and even code generation. As key components of various Natural Language Processing (NLP) applications, LLMs are increasingly integrated into the fabric of everyday life. (Abdali et al., 2024) Despite their impressive performance, recent research highlights the vulnerabilities of LLMs, particularly their susceptibility to adversarial attacks. These weaknesses can take various forms, such as prompt injections, jailbreaking attacks, and other similar exploits. The fundamental paradox of Large Language Models (LLMs) is that many tasks requiring intelligence come naturally to humans but are challenging to implement successfully in these models. This is evident in basic human activities like nuanced language understanding, conversational context, and emotional intelligence. However, it also extends to more specialized tasks, such as creative reasoning and ethical judgment. While LLMs excel at processing vast amounts of data, identifying intricate patterns, and generating complex text, they often struggle with tasks that require genuine comprehension or adaptive thinking (Davis, 2016).

Developing AI for scientific purposes requires a fundamental understanding of the physical and natural world as experienced in everyday human life, along with an awareness of how scientific concepts and laws relate to these experiences. To gauge progress toward this goal, standardized benchmarks would be highly beneficial. Additionally, setting specific challenges could inspire bold and groundbreaking research efforts. Accordingly, providing LLMs that are capable of accurately answering scientific questions is a critical task. The goal of this research is to find methods by which an LLM can be improved in answering scientific questions.

Existing comprehensive benchmarks for LLMs include tasks related to scientific data, but they have notable limitations. First, they primarily emphasize memorization, often overlooking more advanced reasoning and analytical abilities. Second, they fail to account for the evaluation of diverse multimodal inputs, such as charts, molecular structures, and tables, which are essential for understanding scientific literature (Hendrycks et al., 2020).

Answering scientific questions correctly in a large language model (LLM) enhances its reliability, utility, and ethical responsibility, benefiting both users and developers. Accurate responses provide users with accessible, trustworthy information, reducing errors and fostering learning across diverse fields. For developers, maintaining high ac-

curacy boosts the model's credibility, broadens its applications in specialized domains like medicine and academia, and demonstrates effective training on quality data. Socially, correct scientific answers democratize knowledge, bridge educational gaps, and support innovation by aiding researchers and interdisciplinary collaboration. Ultimately, delivering accurate scientific information aligns with the goal of making LLMs impactful, trustworthy tools for solving real-world problems.

Accordingly, The goal of this research is to find methods by which an LLM can be improved in answering scientific questions.

3 Related Work

Multiple studies evaluated AI performance in scientific question answering. For example, Ernest Davis (Davis, 2016) explored the design of hand-constructed multiple-choice tests tailored to evaluate artificial intelligence (AI) by focusing on problems that are easy for humans but challenging for machines. The author outlines strategies for constructing questions at the levels of a fourth-grade child and a high-school student. At the fourth-grade level, questions involving concepts such as time, causality, the human body, sets of objects, and scenarios requiring inductive reasoning or combining facts are identified as effective for distinguishing human and AI capabilities. At the high-school level, questions that connect formal scientific principles to real-world observations or laboratory experiments are highlighted as particularly challenging for AI while remaining intuitive for human students. The study emphasizes that these types of questions are more effective benchmarks for AI evaluation than standardized tests like the SATs or Regents exams, as standardized tests are designed to challenge humans and often fail to capture areas where humans excel but AI struggles. This framework offers a novel perspective on creating benchmarks that better measure AI's limitations relative to human reasoning.

Recent advancements in Large Language Models (LLMs) have transformed scientific literature analysis, yet existing benchmarks fall short in evaluating their higher-level capabilities, particularly in tasks requiring reasoning beyond memorization and handling multimodal data. To address this limitation, **SciAssess**, a specialized benchmark, has been developed to comprehensively evaluate LLMs in scientific literature analysis. SciAssess assesses LLMs across three key dimensions: Memorization (L1), Comprehension (L2), and Analysis & Reasoning (L3), incorporating tasks from diverse scientific domains such as biology, chemistry, materials science, and medicine. The benchmark ensures reliability through stringent quality control measures, including accuracy verification, anonymization, and adherence to copyright standards. By evaluating 11 LLMs, SciAssess identifies their strengths and areas for improvement, providing valuable insights to guide the development of LLM applications in scientific research and analysis. This benchmark offers a more targeted and nuanced approach to assessing LLM performance in scientific contexts. (Cai et al., 2024).

Recent advancements in LLMs like GPT offer potential for automating question generation, yet evaluating their adherence to best practices is essential for ensuring quality. This study analyzes the quality of LLM-generated MCQs in computer science and medicine. Using GPT-based services and a Moodle plugin, MCQs were generated and compared against established item-writing guidelines. While LLMs improved efficiency, challenges included broad or ambiguous items and implausible distractors, emphasizing the need for human oversight to ensure instructional alignment. The paper also suggests solutions for developers to enhance automated question generation toold (Grévisse et al., 2024).

4 Project Plan

In this research, we utilized a BERT-based model with cyclic learning to classify scientific questions into 15 distinct categories. These categories included 'Economics,' 'Physics,' 'Technology,' 'Biology,' 'Mathematics,' 'Astronomy,' 'Chemistry,' 'Psychology,' 'Medicine,' 'Ecology,' 'Environment,' 'Geology,' 'Philosophy of Science,' 'Miscellaneous,' and 'Engineering.' Our primary goal was to evaluate classification metrics such as accuracy, F1 score, precision, and recall, aiming to identify which categories require additional training to achieve top performance in a large language model (LLM).

We employed this approach to gain deeper insights into the performance and gaps of the model in handling diverse scientific domains. Additionally, we analyzed the results using QQ plots and confusion matrices to better understand the distribution and patterns of misclassifications. This al-

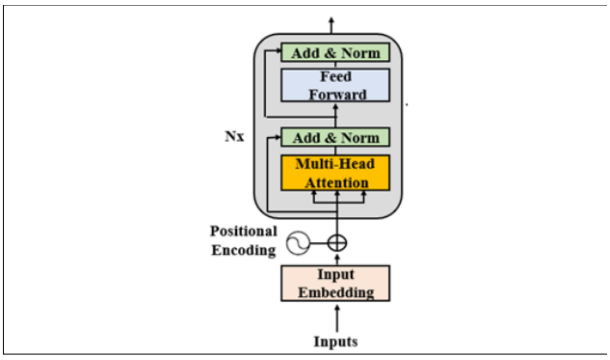

Figure 1: BERT-base architecture

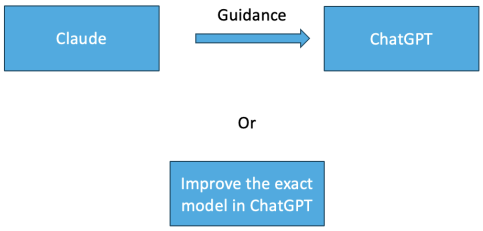

Figure 2: Initial idea

lowed us to pinpoint specific areas where improvements in model training or dataset representation might be necessary. In the subsequent sections, we provide a detailed account of the dataset, the characteristics of the cyclic learning methodology, and the evaluation metrics used to assess the model's performance across the 15 categories.

5 Data

The dataset utilized in this research comprises 4,217 scientifically flawed questions, categorized into 2,961 for training, 421 for validation, and 835 for testing. These questions were carefully curated to include instances of fallacious reasoning or paradoxical logic, extracted from interactions with advanced language models such as ChatGPT and Claude. This unique dataset is specifically designed to analyze and improve model performance in identifying and addressing incorrect scientific reasoning.

This project aims to enhance AI model performance on scientific questions by combining model collaboration and specialization. Observing that Claude identifies more incorrect answers (59/75) than ChatGPT (12/75) highlights their differing strengths. The approach involves two key strategies: (1) Supervisory collaboration, where Claude evaluates and refines ChatGPT's responses through iterative feedback loops; and (2) Classification-based routing, where questions are categorized by discipline or complexity and directed to the most suitable model. This framework optimizes accuracy by leveraging each model's strengths, improves adaptability with dynamic feedback, and scales with additional models or fine-tuning for specific domains, ensuring precise and specialized answers across disciplines.

The training dataset is employed to enable the model to learn patterns and identify key features associated with scientifically incorrect logic. The validation dataset is used to fine-tune hyperparameters and monitor the model's performance during training, helping to mitigate overfitting and ensure robust generalization. The test dataset serves as a benchmark for assessing the model's effectiveness on unseen data, providing an unbiased evaluation of its capabilities.

By employing this dataset, we systematically trained and tested our model to classify and address these flawed scientific questions, enhancing its ability to detect and handle logical inconsistencies in text. This approach facilitates a deeper understanding of the challenges posed by scientific inaccuracies and contributes to the development of more robust systems for reasoning and error detection in natural language processing.

Table3 shows the size of the train, test, and validation datasets respectively.

	Size
Train dataset	2,961
Validation dataset	421
Test dataset	835

Table 1: Table showing the sizes of the train, validation, and test datasets.

6 Models

We fine-tuned a BERT-base model to categorize 4,217 scientifically paradoxical questions, identifying logical inconsistencies from ChatGPT and Claude outputs. The model demonstrated its potential in detecting flawed reasoning, advancing natural language understanding and error detection.

BERT-base model

The BERT-base model (Bidirectional Encoder Representations from Transformers) is a

transformer-based language model developed by Google. It uses a deep bidirectional architecture to understand context in text by analyzing words in relation to all other words in a sentence, rather than sequentially. This approach allows BERT to grasp the nuances of meaning and context more effectively than traditional models.

BERT-base consists of 12 layers (transformer blocks), 768 hidden units, and 12 attention heads, with 110 million parameters in total. It is pretrained on large text corpora using two key tasks: Masked Language Modeling (MLM), where some words in a sentence are masked, and the model predicts them, and Next Sentence Prediction (NSP), which determines if two sentences logically follow each other. These pretraining tasks equip BERT with a robust understanding of language structure and semantics.

BERT-base is widely used for fine-tuning on specific tasks such as text classification, question answering, and sentiment analysis, achieving state-of-the-art performance in many natural language processing (NLP) benchmarks. Its ability to process and understand context-rich language makes it a versatile and powerful tool for NLP applications (Aum and Choe, 2021; Islam and Zhang, 2024; Khadhraoui et al., 2022) .

Cyclic learning

We used cyclic learning in this research for faster and more accurate optimization. Cyclic learning, or cyclical learning rates (CLR), is a training technique where the learning rate oscillates between a predefined minimum and maximum instead of decaying over time. This dynamic adjustment allows the model to explore the loss landscape effectively, helping it escape local minima and converge more precisely. The learning rate changes following specific patterns, such as triangular, triangular2 (with reduced amplitude after each cycle), or exponential decay. Each cycle spans a set number of iterations or epochs, and the rate is updated at every step during training.

The primary benefits of cyclic learning include faster convergence, reduced risk of overfitting, and less need for manual tuning of the learning rate. By alternating between high and low learning rates, it balances exploration and precision in optimization. Cyclic learning is widely used in machine learning tasks like image classification and text analysis, where finding an optimal learning rate is challenging (Smith, 2017).

7 Experiments

We utilized the **BERT-base** model to classify outputs from eight text generation datasets, focusing on specific features to evaluate whether these outputs varied significantly across classifiers. The model was trained using the *AdamW* optimizer with a learning rate of 2×10^{-5}, ensuring efficient weight updates during training. It was trained for 20 epochs with a batch size of 32, striking a balance between training speed and performance. To mitigate overfitting, a weight decay of 0.01 was applied, discouraging overly complex models.

7.1 Classification

This configuration enabled a thorough evaluation of the text generation datasets, revealing how distinct features influence the variability and quality of generated text across contexts. Table 2 provides the classification model and the corresponding parameter values.

	Parameters
Model	BERT-base
Optimization Method	AdamW
Learning Rate	2×10^{-5}
Number of Epochs	20
Batch Size	64
Weight Decay	0.01

Table 2: Summary of training hyperparameters and optimization method used in the experiment. The table includes details on the optimization method, learning rate, number of epochs, batch size, and weight decay.

7.2 Cyclic learning parameters

In this implementation, a cyclic learning rate scheduler is integrated to dynamically vary the learning rate between a minimum and maximum during training. Unlike a static learning rate, which remains fixed, the cyclic scheduler oscillates the learning rate in a predefined pattern. This variation helps the optimizer escape local minima and explore the loss landscape more effectively, which can lead to better model performance and faster convergence. The scheduler also includes a warmup phase, where the learning rate gradually increases during the initial portion of training,

stabilizing optimization and improving training stability.

The cyclic learning parameters are carefully chosen for balanced learning. The base learning rate (10^{-6}) ensures the learning rate never drops too low, while the maximum learning rate (2×10^{-4}) allows for aggressive updates during certain phases. The total training steps, calculated based on the dataset size and epochs, define the overall schedule, while the warmup steps (10% of the total training steps) enable smoother ramp-up at the start. This dynamic approach eliminates the need to manually fine-tune a static learning rate and adapts to different training phases, promoting both exploration and fine-tuning of the model parameters for improved generalization.

7.3 Results and Discusion

In this study, we employed the BERT-base model to evaluate the performance of eight distinct text generation models. Training began with a loss of 2.67 during the first epoch and steadily decreased, indicating effective learning. On the validation dataset, the loss showed a consistent decline until epoch 8, after which it started to rise, suggesting the onset of overfitting as the model began specializing too heavily on the training data. To balance accuracy and generalization, we selected epoch 8 as the optimal point for testing. At this stage, the model achieved a validation accuracy of 83.33% and a test accuracy of 74.94%, demonstrating robust performance while minimizing the effects of overfitting.

To further enhance training efficiency and improve performance metrics, a cyclic learning rate scheduler was integrated into the training process. Unlike static learning rates, which remain fixed, the cyclic scheduler dynamically oscillates the learning rate between a predefined minimum and maximum. This variation allows the optimizer to escape local minima and explore the loss landscape more effectively. Additionally, a warmup phase was incorporated, gradually increasing the learning rate during the early stages of training to stabilize optimization and improve convergence. For this study, the base learning rate was set to 10^{-6}, ensuring the learning rate never dropped too low, while the maximum learning rate of 2×10^{-4} enabled more aggressive updates during specific training phases. Warmup steps, covering 10% of the total training steps, further smoothed the transition to optimal learning.

This cyclic learning approach contributed significantly to the model's performance metrics by dynamically adapting the learning rate to the evolving training process. By avoiding the pitfalls of static learning rates, the scheduler supported both efficient exploration and fine-tuning of the model parameters. The integration of this method also enhanced the BERT-based evaluation, providing deeper insights into the distinct strategies used by the text generation models. These insights enable researchers to identify specific performance patterns, discern the relative strengths and weaknesses of each approach, and make informed decisions about the most suitable methods for different applications. Figures 2, 3, and 4 illustrate the trends in training and validation losses, as well as validation accuracy, across different epochs, highlighting the effectiveness of combining cyclic learning rates with the BERT-based evaluation framework.

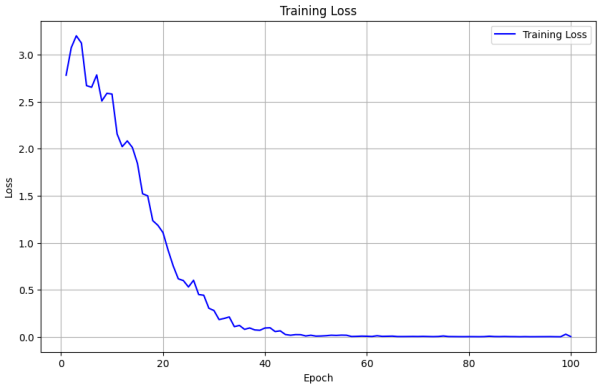

Figure 3: Training loss progression over 20 epochs during model training.

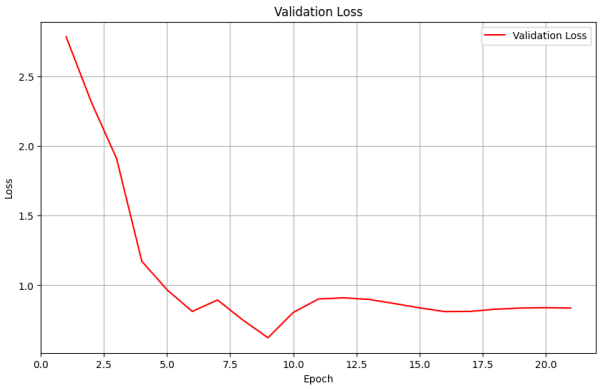

Figure 4: Validation loss over 20 epochs, reflecting model performance on the validation set.

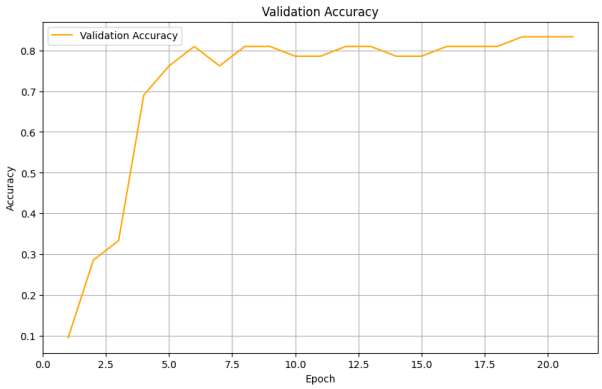

Figure 5: Evaluation accuracy plotted over 20 epochs, demonstrating the model's accuracy during evaluation.

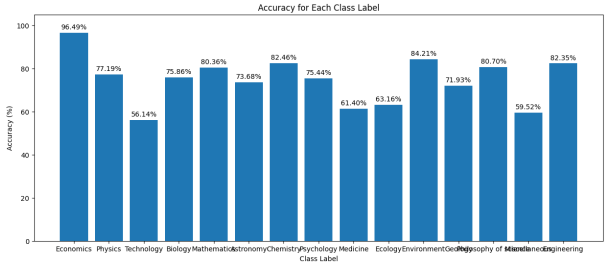

Figure 6: Accuracy of the trained model for each class.

Bert-base model	performance metric
Accuracy	74.94%
F1 score	74.84%
Precision	75.97%
Recall	74.94%

Table 3: Table showing the overall accuracy of the trained model.

Figure 8 presents the classification accuracy for each model separately. The performance of the model varied significantly across different categories, reflecting the varying levels of complexity and clarity in the data. For example, the "Economics" category achieved a high accuracy of 96.49%, indicating that the model can confidently classify questions in this domain. This is likely due to the consistent language patterns and well-defined boundaries of economic topics. In contrast, the "Environment" category, despite having a relatively strong overall performance, required the most training. This is because environmental questions often exhibit subtle nuances and overlap with other categories, making it easier to detect faulty or misclassified questions during evaluation. On the other hand, "Technology" and "Miscellaneous" categories struggled with accuracies of 56.14% and 59.52%, respectively. These low accuracies suggest that these categories are more ambiguous or contain questions that are less distinct in their linguistic or contextual features. Technology questions often involve evolving terminology and cross-domain references, while the "Miscellaneous" category likely suffers from a lack of focus and heterogeneity. Addressing these shortcomings would require better feature engineering, improved training data curation, and possibly fine-tuning of the model architecture to better detect and differentiate subtle patterns in these categories.

Figure 8 presents the confusion matrix and the qq plot. The confusion matrix reveals fascinating insights into how paradoxical scientific questions are misclassified across disciplines, highlighting the inherent complexity of categorizing problematic scientific inquiries. The significant cross-discipline confusion, particularly between Physics-Technology (8 instances) and Medicine-Biology (13 instances), likely reflects questions that deliberately blur disciplinary boundaries or contain logical inconsistencies. For example, the confusion between Medicine and Biology might represent questions that incorrectly conflate cellular mechanisms with clinical outcomes, while the Physics-Technology overlap could indicate questions that misapply physical principles to technological applications. The relatively high accuracy in Economics (96.5%) suggests that paradoxical economic questions may have more distinctive characteristics that make them easier to identify, perhaps due to their often quantifiable nature. The Q-Q plot's deviations from normality take on new meaning in this context. The non-normal distribution at the extremes, particularly visible above the 1.5 theoretical quantile, suggests that some faulty scientific questions are either obviously flawed (leading to very high classification accuracy) or so subtly paradoxical that they consistently fool the classification system. The generally linear trend in the middle range (-1.0 to 1.0) indicates that most paradoxical questions follow predictable patterns of ambiguity,

but the outliers represent particularly challenging cases where the questions may span multiple disciplines in logically inconsistent ways. This distribution pattern could be valuable for identifying different categories of scientific paradoxes and their relative difficulty of classification.

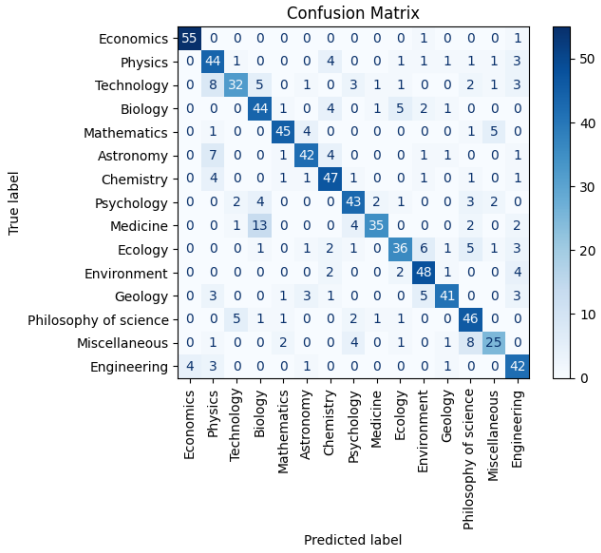

Figure 7: The confusion matrix of the model

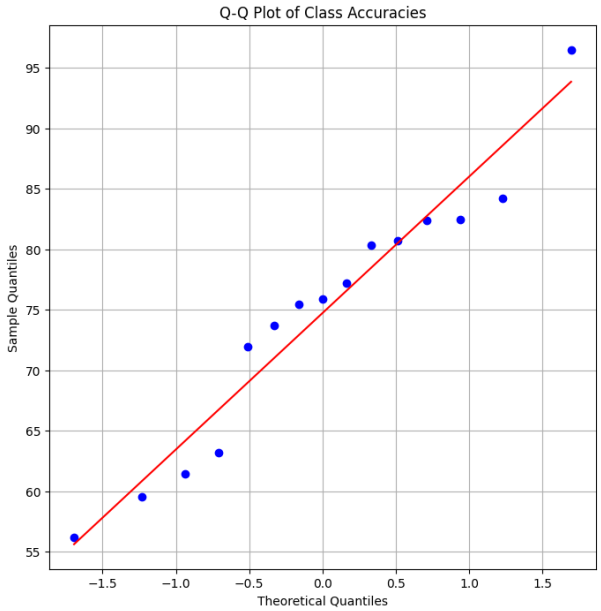

Figure 8: the q-q plot over the categoris

8 Conclusion

In conclusion, the study highlights the effectiveness of employing the BERT-base model combined with a cyclic learning rate scheduler to evaluate text generation models and classify complex scientific questions. The integration of cyclic learning enabled dynamic adaptation of the learning rate, enhancing model optimization and generalization, and contributing to robust performance metrics, particularly at epoch 8 with a validation accuracy of 83.33% and a test accuracy of 74.94%. The analysis of classification performance across categories revealed that domains like "Economics" achieved high accuracy due to well-defined patterns, while "Technology" and "Miscellaneous" struggled due to ambiguity and heterogeneity. The confusion matrix and Q-Q plot provided deeper insights into cross-domain misclassifications, emphasizing the challenges posed by questions that blur disciplinary boundaries or exhibit subtle paradoxes. These findings underscore the importance of tailored feature engineering, refined data curation, and enhanced model architectures to address category-specific challenges and improve detection accuracy. Overall, the study offers valuable methodological insights and practical implications for improving the evaluation and classification of complex and ambiguous textual data.

References

Sara Abdali, Jia He, CJ Barberan, and Richard Anarfi. 2024. Can llms be fooled? investigating vulnerabilities in llms. *arXiv preprint arXiv:2407.20529*.

Sungmin Aum and Seon Choe. 2021. srbert: automatic article classification model for systematic review using bert. *Systematic reviews*, 10:1–8.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, et al. 2024. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976*.

Ernest Davis. 2016. How to write science questions that are easy for people and hard for computers. *AI magazine*, 37(1):13–22.

Christian Grévisse, Maria Angeliki S Pavlou, and Jochen G Schneider. 2024. Docimological quality analysis of llm-generated multiple choice questions in computer science and medicine. *SN Computer Science*, 5(5):636.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Md Saiful Islam and Long Zhang. 2024. A review on bert: Language understanding for different types of nlp task.

Mayara Khadhraoui, Hatem Bellaaj, Mehdi Ben Ammar, Habib Hamam, and Mohamed Jmaiel. 2022. Survey of bert-base models for scientific text classification: Covid-19 case study. *Applied Sciences*, 12(6):2891.

Leslie N Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE.