# Overcoming the challenges of Large Language Models: Introducing a novel proposition for Synthetic Data Validation

Urvashi Bhargava
*Computer Science and Engineering Department*
*PES University*
Bengaluru, India
pes2202100886@pesu.pes.edu

Y Teresha
*Computer Science and Engineering Department*
*PES University*
Bengaluru, India
pes2202101007@pesu.pes.edu

Nishank Koul
*Computer Science and Engineering Department*
*PES University*
Bengaluru, India
pes2202101224@pesu.pes.edu

Chandrashekhar Pomu Chavan
*Computer Science and Engineering Department*
*PES University)*
Bengaluru, India
cpchavan@pes.edu

***Abstract*—The market debut of ChatGPT gave rise to the development and deployment of various other Large Language Models (LLMs) that achieve state-of-the-art performance across various tasks. The growing popularity of these models has captivated some to attempt to construct or enhance their own LLM. We must be aware of the significant problems that already exist and that we might face along the way. This paper aims to identify and investigate the main challenges in this field, provide existing solutions, and propose novel approaches to mitigate them. A unique Truth-Table proposition for validating synthetic data is presented examining two models, along with a bidirectional knowledge graph-based solution for curing the reverse curse problem, data generation strategies, domain adaptation methods, and the use of a custom dataset to address model hallucinations. The methodology and findings of this study provide valuable insights for users, researchers, and industry experts who are interested in LLMs. It serves as a reference for future research on current models, refining models or developing domain-specific ones.**

***Keywords—Large Language Models, Challenges, Strategies, Synthetic Data Validation, Reversal Curse, Model Hallucinations, Knowledge Graph, Data Scarcity.***

## I. INTRODUCTION

LLMs are a fragment of deep learning that comprehends, predicts, and generates new human-like text using vast datasets and natural language understanding techniques. The primary goal of LLMs is to evaluate and provide contextually precise content. Foundational models are pre-trained on massive datasets with billions or trillions of parameters. Further, these pre-trained models are finetuned to perform specific tasks. They can generate language that is both cohesive and contextually suitable. Integrating LLMs into company planning improves client involvement, fosters more efficient communication, and improves overall corporate efficiency.

## II. LARGE LANGUAGE MODELS

### *A. Advancements in LLMs*

Many significant attributes have contributed to recent advancements in computational infrastructure, training approaches, and model designs resulting in the remarkable growth in the field of LLMs, which include:

Model size scaling: Over time, the volume of data in the market has grown substantially. For instance, older versions like the GPT-2 have hundreds of millions of parameters whereas newer models like the GPT-4 or GEMINI have almost a billion or trillion parameters. It has been proved that the larger the model, the greater its performance [10].

Multimodal Capabilities: This new evolution broadens models' scope beyond typical text-based interpretation and development. It enables the model to process data in the form of images, videos, and audio. This is obvious in models such as Contrastive Language-Image Pretraining (CLIP) and DALL-E.

Fine-Tuning Techniques: The pre-trained models provide a firm basis but fine-tuning the model to adapt to tasks raises the overall performance of LLMs significantly. This is known as transfer learning, and it enables a model to apply what it has learned to a new task.

### *B. Next-Generation Models*

The newest advancements in AI (Artificial Intelligence) innovation include SORA by OpenAI, Gemini by Google, and Microsoft Copilot by Microsoft:

Sora is a text-to-video model [11] that creates videos up to 1 minute with various aspect ratios. It uses latent diffusion models for high-resolution video generation, cascade evolution for long videos, recaptioning techniques of DALLE 3 for improved language understanding, diffusion transformer for flexibility and

scalability in data, and native video transformer for training videos on native aspect ratios.

Gemini is a multimodal model that seamlessly understands and processes text, code, audio, image, and video. It includes many versions for performing different tasks like Ultra for complex operations, Pro for scalability, and Nano for on-device efficiency. It has dedicated features to filter out harmful content and demonstrates state-of-the-art performance across benchmarks.

Microsoft Copilot smoothly integrates LLMs, Microsoft graphs, and other applications. LLMs interpret user prompts and the Microsoft graph allows safe access to secure organizational data, regulating real-time assistance across Microsoft applications.

## III. Major Challenges in Large Language Models

### A. Scarcity of high-quality data

The lack of high-quality data is a significant obstacle during the pre-training of the model. Without enough data, the model finds it difficult to identify reliable patterns and accurate relationships, resulting in mediocre performance. The added data complexity in multilingual and multimodal LLMs amplifies this issue. The lack of extensive datasets in specific languages makes it more difficult for multilingual LLMs to grasp the underlying concepts. This is especially because English dominates as the primary language of the internet. Similarly, multimodal LLMs find it challenging to efficiently align several data modalities (text, image, video, and audio) when data is scarce, which makes learning representations extremely challenging.

### B. Synthetic Data Validation

The lack of validation strategies for synthetic data generation, particularly in scenarios where data is scarce causes significant obstacles during the foundational phases of building a model. This functionality is often absent in the integration of synthetic data into model training pipelines. The lack of dependable validation procedures creates consequential issues in assuring the strength of models during their development phases. In section 4, we present our proposal to tackle this problem.

### C. Limited Domain Specific Understanding

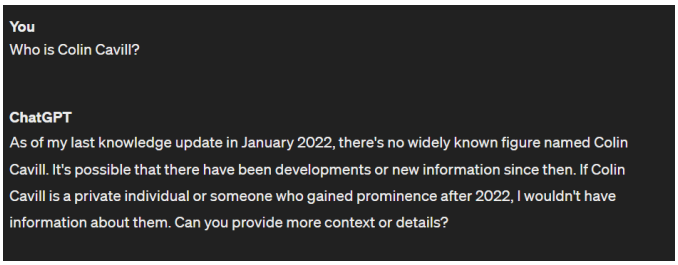
LLMs have a remarkable performance in numerous tasks, yet they lack a comprehensive understanding of specific concepts in underrepresented topics, which is due to the vast knowledge base that currently exists. Domain-specific LLMs grasp the unique phrases within their domain better than generalist LLMs. For example, the general term" resolution" could be misinterpreted as a solution or a formal statement, whereas a software development focused LLM would understand it as the smallest display unit of image or video detail, enhancing accuracy. The limited availability of domain-specific datasets for pre-training or fine-tuning LLMs also poses a great data-related challenge. The reliability of LLMs domains such as biomedical [1], legal [4], or finance [2], requires high accuracy of question-answering. The newly pre-trained or finetuned domain-specific models like BloombergGPT, StarCoder, and BioBERT perform better on their domain-specific benchmarks [1, 5, 3].

### D. Model Hallucinations

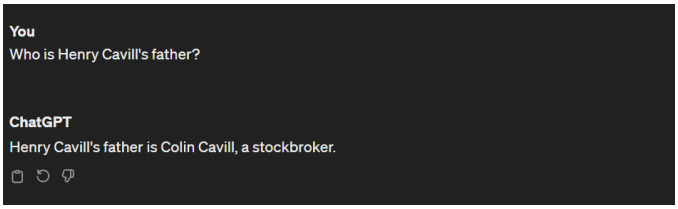
When a model is hallucinating, it produces outputs that include fabricated information and errors [8] due to the absence of access to instantaneous data. This problem becomes evident when the model does not have the required domain specific knowledge or is presented with prompts that are not included in its training data. As a result, the model produces text that seems reasonable but is incorrect. To effectively address model hallucinations, it is necessary to understand the prompts that make the model hallucinate.

### E. Reversal Curse

Reversal Curse is observed in LLMs like GPT or Llama. It tells us that the model cannot naturally create statements of the structure (attributes belong to object) if it is trained on sentences like (object has attributes). It explains a shortcoming in the model's logical deduction power that a phrase structure (A has B) cannot be automatically applied to its opposite case (B belongs to A) until and unless the model is directly trained with this exposure. Tests [9] conducted on various LLMs have provided evidence of the Reversal Curse.




(a)




(b)

Fig. 1. GPT 3.5 accurately recognizes Colin Cavill as Henry Cavill's father in scenario (b), yet fails to make the same identification in scenario (a).

## IV. Strategies and Solutions

### A. Data Generation and Acquisition

Social media sites can be used to generate datasets. Reddit offers an extensive variety of user-created information on a wide range of subjects and languages, making it an excellent source for gathering abundant data. To gather information obtained via Reddit's API, we can use PRAW (Python Reddit API Wrapper), a Python library specifically created to simplify access to Reddit's API while following its rules and regulations. By leveraging PRAW, researchers may efficiently gather data from Reddit, which can then be utilized to train and improve LLMs. Twitter can also be used as a data source.

Hugging Face is an extremely popular platform that offers open-source datasets and models. Users can use the platform for hosting, training, and deploying models. The primary domain of interest is natural language processing. A user-friendly library called Transformers allows the usage of powerful models. It offers online courses, workshops, and provides extensive documentation which is accessible to everyone.

Gretel.ai is a Software platform that helps in generating synthetic datasets that closely resemble real data. Users can use this platform to build their domain specific datasets without affecting the real data. Moreover, developers can estimate the accuracy of the synthetic data they produce. Its ability to generate synthetic data improves with the generalization of the foundational training data.

### *B. Bi-directional Knowledge-Graphs for Curing Reversal Curse*

To address the Reversal Curse challenge observed in LLMs, we present a proposition that allows the incorporation of a bidirectional knowledge graph that would seamlessly include both the original and the reversed association of entities during the training phase. This graph includes nodes (entities) and edges (relationships between nodes) related to each other bidirectionally. This amplifies the learning of implicated relationships which enables the models to learn contextually rich representations mitigating the reversal curse challenge. This model allows bidirectional exposure and generalization capabilities of data during the training phase which helps in fostering improved model performance and developing a more robust understanding of relationships between the entities.

| Child | Parent | can_reverse |
|---|---|---|
| Sasha Calle | Samira Calle | false |
| Kate Mara | Timothy Christopher Mara | true |
| Josh Hartnett | Daniel Hartnett | false |
| Monica Barbaro | Heidi Wagner | false |

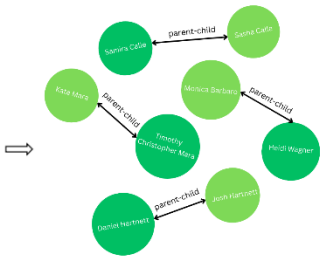



Fig. 2. A visual representation of reversal curse pairs being connected.

### *C. Domain Adaptation*

Pre-training Models can gain domain-specific knowledge through pre-training, which involves understanding a diverse range of concepts and specialized activities within that sector. It aids in generating user-friendly and precise solutions that are suitable and relevant to that circumstance. The improved precision of the model's generating capabilities strengthens its understanding of the specific subject area. To achieve domain-specificity, we can utilize solely domain specific data instead of general data during the initial training of the model [5]. Another approach is to combine both types of data in a self-supervised learning set of tasks [3]. Alternatively, we can apply a" cold start" to a pre-trained generic LLM, which then becomes proficient in tasks through transfer learning [1].

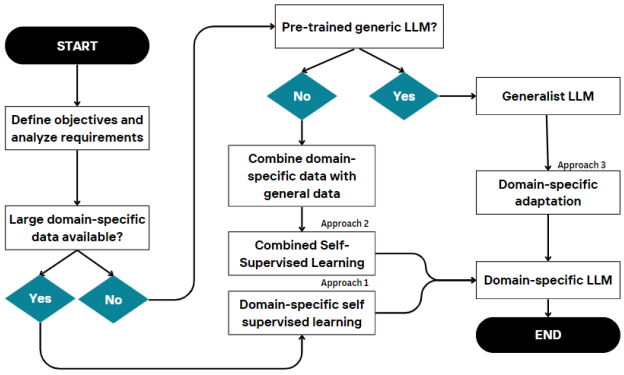



Fig. 3. Charting the Course: Pre-Training Approaches for Domain Adaptation

Fine-tuning With embeddings, models can grasp the semantic similarities between languages, even without encountering any explicit data in those languages. By fine-tuning existing LLMs, the process of creating language models that adapt to a specific domain can be accelerated significantly. The process involves utilizing general data obtained from pre-training, and then adjusting the model's parameters in response to specific input related to the domain. Less labelled data and less training time can be effective for fine-tuning. Due to its low cost, this methodology has become the most widely used method for constructing domain specific LLMs. Nevertheless, fine-tuning can result in significant performance implications, particularly in specialized domains.

Reading Comprehension Approach Recent research proved that conversion of raw corpora into a reading comprehension targeting a specific domain is superior to fine-tuning and pre-training models on domain-specific raw corpora [7]. It is proposed that, if raw domain-specific data is converted to a reading comprehension while identifying the title, domain keywords, natural language inference explaining the entailment reason, cause & effect, and semantic similarity, the inferences drawn show much better prompting abilities compared to fine-tuning. This was inspired by how humans learn through reading comprehension which involves reading followed by practice to increase memory power and enhances one's capacity to respond to queries based on newly acquired information.

### *D. Model Hallucinations*

Hallucination-Resilient Training This is carried out by training the model on an already existing dataset of hallucinogenic prompts available on the internet. By doing this we address the challenge of hallucination by training the model to respond with an acknowledgment of ambiguity such as 'I am not certain about the response this' or 'I do not know the answer to your question' thereby significantly reducing the generation of inaccurate information and improving the reliability of the model.

Fine-tuning on domain-specific data Acquired domain-specific datasets with accurate information are necessary for this approach. By using transfer learning approaches adjust the parameters of the model to capture the distinctions of the selected domain and fine-tune the pre-trained LLM on domain-specific data. By doing this, we notice an improvement in

contextual understanding and a decrease in hallucinations, particularly when producing responses inside the fine-tuned domain. We can compare the performance of the baseline and the fine-tuned model by using metrics like accuracy, precision, recall, and F1 score to see if the model has overcome the challenge of hallucinations.

## V. Truth-Table Proposition for Synthetic Data Validation

### A. Leveraging agreement between Language Models

Proposition 1. Given the substantial concordance observed in similarity metrics, including histogram analysis, BLEU and ROUGE scores, between Language Models LM1 and LM2, based on the synthetically generated datasets D1 and D2 respectively, it is postulated that the accuracy of the synthetic data generated by both models can be validated by checking for similarity between the two. This assertion draws parallels with the phenomenon of cumulative validation inherent in human reasoning, thereby suggesting a positive correlation between agreement in similarity metrics SIM (D1, D2). It follows the truth table of logical AND, implying that if both models indicate truth (T), then the accuracy of the synthetic data is likely to be validated (T) considering the ground truth.

### B. Experimental Design

The datasets used in this study were chosen with great care to cover a wide range of fields. To create synthetic mathematical data, we used the GSM8K dataset from Kaggle. We customized the dataset to meet our needs by extracting the numerical solution from the output, which made calculations easier. This dataset is renowned for its extensive compilation of math questions and corresponding answers from primary education. We meticulously chose it to incorporate mathematical topics that are pertinent to our research. We employed sophisticated methods such as regular expression parsing to extract precise numerical values from the dataset. Furthermore, we employed the b-mc2/SQL create-context dataset from Hugging Face to generate synthetic SQL query data. This collection comprises structured data on database schemas and their corresponding natural language queries, which aids in the generation of synthetic SQL queries. We carefully chose and analyzed the datasets to improve their appropriateness for generating synthetic data. By ensuring that the synthetic data created in our study is both robust and relevant, we can ensure its high quality and usefulness.

### C. Implementation

This section details the methodology employed to evaluate the agreement between language models (LLMs) on synthetic data accuracy. We focus on two distinct domains in this experiment: SQL query data and mathematical data.

SQL Query Synthetic Data: We leveraged the T5 text-to-code pre-trained model (t5-small) from the Hugging Face Transformers library. This model is specifically designed to translate natural language descriptions into various programming languages, including SQL. This model was fine-tuned on the b-mc2/SQL create-context dataset to learn schema-query relationships. Language Models: For comparison purposes, another pre-trained text-to-SQL model, PipSQL (PipableAI/pip-sql-1.3b), was employed. PipSQL is specifically designed for converting natural language descriptions into SQL queries. Evaluation Process: A python script utilizing the fine-tune T5 model takes an input prompt with two sections:

1) Database Schema Definition, which defines the database tables, columns, and data types like 'CREATE TABLE' statements.
2) Natural Language Question, which describes the desired information to be retrieved from the database using a natural language question.

By combining these elements, the script allows the T5 model to comprehend the structure of the data and the user's information need, generating a corresponding accurate SQL query.

Mathematical Synthetic Data: We utilized a pre-defined Lang Chain prompt template to generate synthetic responses to mathematical questions. Lang Chain is a framework for prompting LLMs in a structured manner. This approach ensures consistency and reduces variability in the generated responses. Language Models: Two LLMs were employed for synthetic data generation: "gpt-3.5-turbo-instruct" and "davinci-002". Based on numerical performance, models were chosen. Evaluation Process: Since the answer's column might include texts along with the numerical answer (e.g., "The answer is 42"), a regular expression was applied to isolate the final number. Regular Expression used for output extraction:

$$\text{pattern} == r' [-+]? \backslash d*\backslash.\backslash d+|\backslash d+' \quad (1)$$

Following this extraction, the numerical values obtained from the LLM responses were compared to assess their agreement on the answer.

### D. Performance Metrics

We evaluated the performance of the models using the metrics defined below:

Histogram analysis for similarity requires a comparison of feature distributions within datasets.

ROUGE (Recall Oriented Understudy for Gisting Evaluation) is a tool for checking how similar computerized text is to human-written summaries. Each variant of ROUGE provides us with the recall, precision, and F1 scores.

The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of computerized text. It provides a quantitative measure of the generated text's similarity to human written references.

Mean Absolute Error and Mean Squared Error are used to quantify the difference between expected values and factual values, with lower values indicating better performance. methodology employed to evaluate the agreement between expected values and factual values, with lower values indicating better performance.

### E. Results

The results in mathematical domain underscore the critical importance of model selection for numerical tasks, affirming our proposition that an agreement of (F, F) leads to (F).

Conversely, in the SQL Query domain, both models excelled in terms of BLEU and ROUGE scores. The histograms of both models closely resembled each other upon visual analysis, validating our proposition that an agreement of (T, T) leads to (T).

TABLE I. PERFORMANCE METRICS

| Domain | Model | Metrics |
|---|---|---|
| Math | gpt-3.5-turbo-instruct | Mean Absolute Error: 2534.89<br>Mean Squared Error: 280,373,691.10 |
| Math | davinci-002 | Mean Absolute Error: 2490.37<br>Mean Squared Error: 262,743,986.78 |
| SQL Query | t5-small | BLEU Score: 0.7455<br>ROUGE-1: $r_1$=0.9353, $p_1$=0.9463, $f1_1$=0.9387<br>ROUGE-2: $r_2$=0.8440, $p_2$=0.8523, $f1_2$=0.8455<br>ROUGE-L: $r_3$=0.9249, $p_3$=0.9351, $f1_3$=0.9280 |
| SQL Query | pipsql | BLEU Score: 0.4595<br>ROUGE-1: $r_1$=0.7316, $p_1$=0.8162, $f1_1$=0.7579<br>ROUGE-2: $r_2$=0.5321, $p_2$=0.5780, $f1_2$=0.5412<br>ROUGE-L: $r_3$=0.7272, $p_3$=0.8104, $f1_3$=0.7529 |

These findings emphasize the significance of model choice and performance variability across different domains, highlighting the complexity of synthetic data generation and model evaluation.

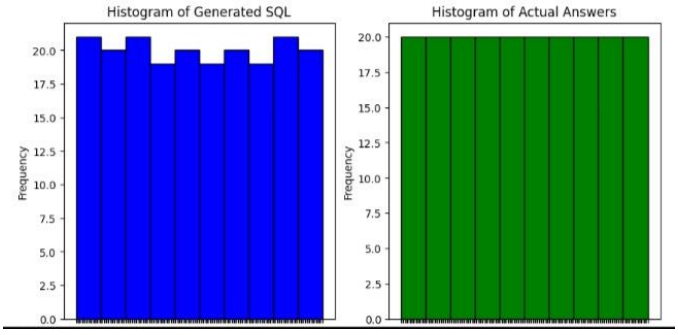



Fig. 4. SQL queries synthetically generated by PIP SQL and T5-Small.

This subsection explores the rationale behind our proposition: agreement between language models (LLMs) on synthetic data characteristics leads to increased confidence in its accuracy. We base this on two key points:

Analogy to Human Consensus Humans rely on consensus to establish truth. If multiple independent sources agree on a statement, the likelihood of it being accurate increases. Similarly, when two LLMs, functioning as independent sources, show agreement on the similarity of synthetic data to real-world data distributions through various metrics, our confidence in the data's quality strengthens.

Justification through AND Truth Table Consider a simplified truth table representing the relationship between model agreement (LM1 agrees and LM2 agrees) and synthetic data accuracy: Table 2 highlights that high agreement (both models agree - T, T) corresponds to a true (T) scenario for synthetic data accuracy. Conversely, disagreement (any combination with F) leads to uncertainty about the data's quality. This aligns with the concept of an AND operation in logic, where both conditions (model agreement) need to be true for the outcome (high confidence in accuracy) to be true.

TABLE II. TRUTH TABLE FOR INTER MODEL AGREEMENT

| LM1 agrees | LM2 agrees | Synthetic Data Accuracy |
|---|---|---|
| True (T) | True (T) | Likely True (T) |
| True (T) | False (F) | Uncertain (F) |
| False (F) | True (T) | Uncertain (F) |
| False (F) | False (F) | Uncertain (F) |

This framework provides a theoretical foundation for our proposition. We suggest that there is a high probability of the synthetic data being accurate by analyzing similar responses of the LLMs.

### *F. Applications*

The utilization of synthetic data production has a wide range of applications in many businesses, facilitating innovation and optimization in numerous disciplines. Here are four examples demonstrating its versatility:

Improving Educational Platforms: Customizing mathematical problems to suit the specific requirements of each student using synthetic data creation has the potential to transform online education by promoting tailored learning experiences and enhancing student understanding.

Database Query Optimization: The creation of artificial data enables the evaluation and improvement of the speed and effectiveness of SQL queries, giving database managers the ability to boost the efficiency of operations and the overall performance of the system.

Personalized Medical Diagnosis: The creation of artificial medical records tailored to individual patient profiles helps healthcare practitioners enhance diagnostic algorithms and treatment strategies, resulting in better patient outcomes through customized treatments.

Financial Risk Assessment: Synthetic data simulation allows financial firms to evaluate and reduce risk by studying various market situations, improving decision-making processes, and ensuring financial stability.

### *G. Limitations*

We understand and accept certain limitations in our proposal. It can be challenging to select the most suitable model for a specific topic. Fine-tuning models for each unique task is not always feasible. Most importantly, the lack of ground truth in many scenarios makes it difficult to identify situations where both the models exhibit similar but low agreement (F, F), leading to uncertainty about the data's quality. Additionally, in cases where one model shows high agreement (T) while the other exhibits low agreement (F), the overall confidence level remains unclear, requiring further investigation.

## VI. LIMITATIONS

Recognizing the limits that come with every research project is crucial. Although our proposed methods display

potential in tackling the difficulties of big language models, it is important to be mindful of some limitations and factors to consider. The efficacy of synthetic data validation can vary based on the intricacy and variety of the target domain. Moreover, the dependence on pre-existing datasets for training and evaluating models may create biases and limits that impact the capacity to apply our findings to a wider context. By openly recognizing these constraints, our objective is to offer an impartial and practical evaluation of our research contributions.

## VII. Conclusion

Our detailed analysis of the challenges and the proposed propositions and solutions related to LLMs such as bidirectional knowledge graphs and truth table propositions for synthetic data validation tells us about the effectiveness of the model and our process of evaluating data precision. We have identified a considerable number of challenges, including the limited availability of data, validation of synthetic data, comprehension of domain-specific knowledge, instances of model hallucinations, and the occurrence of the reversal curse phenomena. Although our suggested propositions and solutions show promise in addressing these challenges, we acknowledge the complexities and limitations that are inherent with the development and effective implementation of LLMs. Our paper seeks to be a valuable resource for researchers and industry experts adding on to the improvement in reliability and effectiveness of LLMs.

## References


[1] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240 (2020)

[2] Li, Y., Wang, S., Ding, H., & Chen, H.: Large language models in finance: A survey. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 374-382) (2023, November)

[3] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., & Mann, G.: Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023)

[4] Zheng, L., Guha, N., Anderson, B. R., Henderson, P., & Ho, D. E.: When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In Proceedings of the eighteenth international conference on artificial intelligence and law (pp. 159-168) (2021, June)

[5] Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., & de Vries, H.: StarCoder: may the source be with you!. arXiv preprint arXiv:2305.06161 (2023)

[6] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., & Natarajan, V.: Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138 (2022)

[7] Cheng, D., Huang, S., & Wei, F.: Adapting large language models via reading comprehension. arXiv preprint arXiv:2309.09530 (2023)

[8] Azamfirei, R., Kudchadkar, S. R., & Fackler, J.: Large language models and the perils of their hallucinations. Critical Care, 27(1), 1-2 (2023)

[9] Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O.: The Reversal Curse: LLMs trained on" A is B" fail to learn" B is A". arXiv preprint arXiv:2309.12288 (2023)

[10] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A.: A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435 (2023)

[11] Karaarslan, E., & Aydin, O. Generate Impressive Videos with Text Instructions: A Review of OpenAI Sora, Runway, Midjourney, and Comparable Models (2024)