A Scalable Data-efficient Solution for Large Language Models

Yongding Zhu Mike Cheng

{enoch.zyd, mike}@orbifold.ai

Abstract. With the rapid advancement of Generative AI, low-cost, high-quality datasets are essential for the development of successful Large Language Models (LLMs), including foundation models, fine-tuning, and Retrieval-Augmented Generation (RAG). In this paper, we propose a novel solution for general LLM development that is both scalable according to the scaling laws of modern LLMs and data-efficient. We will demonstrate that datasets can be both high-quality and compact, achieving the same output model quality as larger datasets.

1 Introduction

Typical datasets are labor-intensive and costly to create, often teaching only a narrow set of concepts. While standard LLMs excel at general tasks, they require significant effort (e.g., fine-tuning, Reinforcement Learning from Human Feedback (RLHF), and RAG) to adapt to domain-specific tasks. Moreover, models that perform well on benchmarks often exhibit disappointingly poor performance on stress tests [1].

In this paper, we will first define the concept of high-quality in the context of LLMs and explain why such metrics are essential for both datasets and the resulting models. Next, we will discuss the scaling laws of LLMs, including the ground truths and notable exceptions. We will then demonstrate that larger datasets are not necessarily better; in fact, they can be more costly and yield poorer performance. Finally, we will illustrate how to achieve data efficiency without compromising on model quality.

2 Dataset Evaluation

Evaluating Large Language Models (LLMs) is complex [2]. It involves a variety of tests and benchmarks designed to assess different aspects of their performance. Even with comprehensive evaluations, models often exhibit overfitting and bias, rendering them impractical for real-world applications. Here are some of the major evaluation tests:

1. MMLU (Massive Multitask Language Understanding)
   MMLU evaluates a model's multitask accuracy across a wide range of subjects, from elementary mathematics and U.S. history to computer science

and law. It tests the model's ability to understand and perform well across diverse topics, making it a comprehensive evaluation for general knowledge and reasoning.

2. GLUE (General Language Understanding Evaluation)
   GLUE is a benchmark that evaluates models based on various NLP tasks, including question answering, sentiment analysis, and textual entailment. It assesses a model's general language understanding and its ability to perform across different NLP tasks.
3. SQuAD (Stanford Question Answering Dataset)
   SQuAD consists of reading comprehension questions about a set of Wikipedia articles. It tests a model's ability to understand context and answer questions accurately based on the given passages.
4. XTREME (Cross-lingual TRansfer Evaluation of Multilingual Encoders)
   XTREME is a benchmark suite for evaluating the cross-lingual generalization capabilities of multilingual models across a diverse set of languages and tasks. It tests the model's ability to understand and perform tasks in multiple languages.
5. BigBench (Beyond the Imitation Game Benchmark)
   BigBench includes a diverse set of tasks aimed at evaluating a model's capabilities in areas such as reasoning, mathematics, and common sense. It provides a comprehensive and challenging set of tasks to assess the general intelligence and problem-solving abilities of models.

LLMs frequently achieve human or even superhuman performance on benchmarks. However, when deployed in real-world scenarios, their performance often falls short of these benchmark expectations. This discrepancy, or gap, between ”benchmark performance” and ”real-world performance,” likely arises because models optimize specifically for benchmark success. This is analogous to a student who excels on an exam by studying only past exam questions, rather than understanding the underlying material.

For domain-specific LLMs, enterprises often rely on their own business metrics, such as user engagement rates, to evaluate effectiveness. However, this approach intertwines business logic with model development, further complicating an already complex issue.

For LLMs, the GIGO (Garbage In, Garbage Out) principle is crucial: flawed training datasets lead to flawed models. Evaluating the quality of datasets is a much more effective and efficient way to assess the overall performance of the model.

We introduce the F.L.A.G. paradigm of evaluating datasets.

1. Fresh
   Unlike foundation models, enterprise LLM systems often require up-to-date data to address current business situations effectively. For instance, an enterprise knowledge chatbot must be able to answer questions such as ”What is the current status of the ticket?” or ”Is this order fulfilled?” Therefore, the freshness of the training dataset is crucial for providing accurate and relevant responses to these types of queries.

2. Large
   Small datasets are limited in numerous ways. Just as a human expert must study a large volume of material to grasp many interconnected concepts, AI models also need to learn from extensive, well-curated datasets. However, as datasets grow larger, the associated costs can escalate significantly. There is a limit to the benefits of scaling, and as the volume of data increases, maintaining quality often becomes a bottleneck.
3. All-inclusive
   In many datasets, certain use cases may be overrepresented, while others are quite rare. This imbalance can lead to an overemphasis on common use cases, while important but infrequent ones are overlooked. This issue is particularly pronounced when sampling is used to reduce dataset size, as it often results in the exclusion of these critical yet less frequent use cases.
4. Granular
   The saying ”the devil is in the details” holds true for real-world LLM applications. Often, as we continue to ask follow-up questions, the responses become less accurate and more simplistic. This is likely because the training datasets lack sufficient detail. For example, using low-definition videos to train a text-to-video model can result in a model that struggles with fine details. This issue is common in models like OpenAI Sora and Pika, among others.

We developed the FLAG score to consistently and continuously evaluate and monitor datasets. Our experience in developing real-world enterprise LLM applications shows that the FLAG score is highly correlated with clients’ business performance metrics.

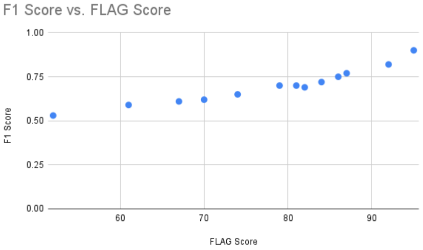

Fig. 1. F1 Score V.S. FLAG Score for a sentiment analysis application

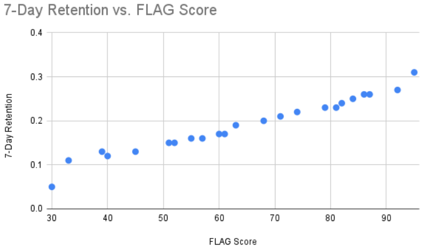

Fig. 2. 7-Day Retention Rate V.S. FLAG Score for a companion app

3 The Scaling Law

The advent of LLMs such as GPT-4, Gemini, LLaMA, Mistral, Claude, and their successors has significantly advanced the capabilities of artificial intelligence in understanding and generating human language. These models are trained on vast datasets and leverage immense computational resources, leading to unprecedented performance in various NLP tasks. A key aspect of this success is the adherence to scaling laws, which dictate how increasing the size of the model and the dataset influences performance. [3]

Scaling laws in the context of LLMs refer to the empirical relationships observed between the size of the model (number of parameters), the size of the training dataset, and the resultant performance on NLP tasks[4]. These laws suggest that larger models trained on more extensive datasets tend to perform better. However, this improvement is not linear and is subject to diminishing returns. As the number of parameters in a model increases, the capacity of the model to capture and generate language patterns also grows. Studies have shown that for many tasks, doubling the number of parameters leads to a consistent improvement in performance metrics such as accuracy and fluency. However, beyond a certain point, the performance gains diminish, and the marginal benefit of adding more parameters decreases [5].

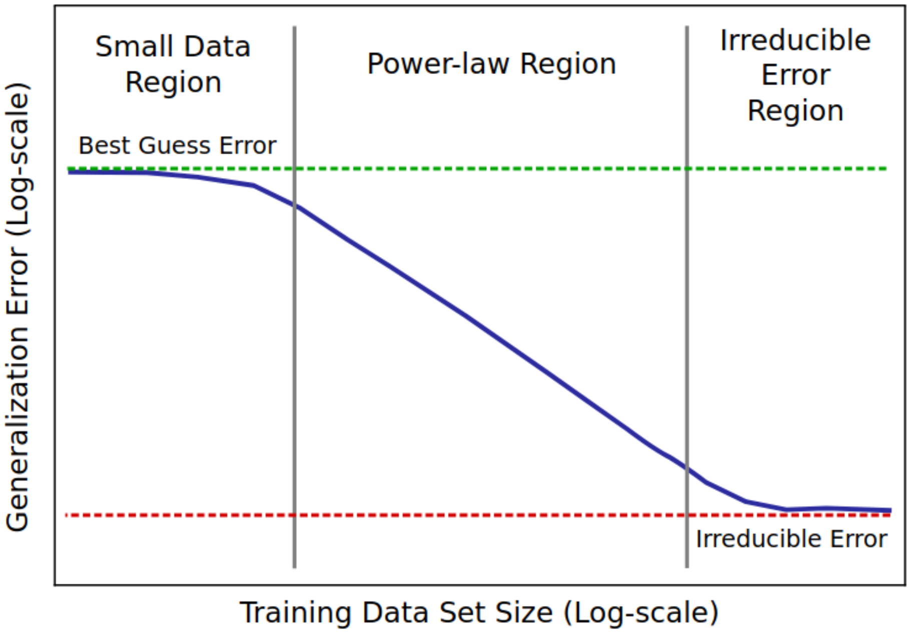

Fig. 3. The diminishing effect of dataset scaling. Source [6]

The size and quality of the training dataset are equally crucial. Larger datasets provide more linguistic contexts, enabling models to learn more robust and nuanced representations of language. The scaling law indicates that models benefit significantly from more data, but similar to model size, there is a point of diminishing returns. Moreover, the quality of data becomes increasingly important as the dataset grows; larger datasets must be well-curated to ensure they add valuable information rather than noise.

One of the primary implications of scaling laws is the substantial increase in computational and financial resources required for training larger models on extensive datasets. Training state-of-the-art LLMs demands significant investments in hardware, energy, and time. For instance, the training of models like GPT-3 involves thousands of GPUs running for weeks, translating to millions of dollars in costs.

Despite the theoretical advantages of scaling, practical limitations often arise. The GIGO (Garbage In, Garbage Out) principle highlights that flawed datasets lead to flawed models, regardless of size. In real-world applications, models trained on vast but imperfect datasets can underperform due to the presence of biases, noise, and irrelevant information. Additionally, the increased complexity of larger models can make them more difficult to fine-tune and deploy effectively.

Improving the quality of training data is paramount. Developing robust data preprocessing and augmentation techniques can help ensure that larger datasets contribute valuable information to the model. Implementing rigorous data validation and curation processes can also minimize the inclusion of biased or irrelevant data.

4 Larger Is Not Better

The assumption that ”larger is better” for datasets in training Large Language Models (LLMs) often overlooks the critical issue of data quality. Larger datasets, while containing more information, are also more likely to include significant amounts of noise. In contrast, smaller, high-quality datasets can provide more precise and relevant training data, potentially leading to better model performance.

Noise in data refers to irrelevant, redundant, or incorrect information that can obscure the patterns the model is trying to learn. Noise can confuse the learning process, leading to models that generalize poorly and perform suboptimally on real-world tasks.

As datasets grow larger, the ratio of valuable information (signal) to irrelevant or incorrect data (noise) often decreases. This makes it harder for the model to identify and learn from meaningful patterns. The increased presence of noise necessitates more sophisticated data cleaning and preprocessing steps, which can be both time-consuming and resource-intensive.

Focused datasets reduce the risk of learning from incorrect or irrelevant data, leading to more reliable model training. Studies have shown that models trained on smaller, well-curated datasets often outperform those trained on larger, noisier datasets in terms of accuracy and generalization.

Furthermore, Training on smaller datasets reduces computational costs and training times, making the process more efficient and accessible. This approach is especially beneficial for organizations with limited resources, allowing them to achieve competitive performance without the need for massive data and infrastructure. Numerous case studies have demonstrated that smaller, high-quality datasets can yield better results. For instance, models trained on expertly curated datasets for medical diagnosis or financial forecasting often outperform those trained on larger, less refined datasets [7]. Research in machine learning and data science supports the notion that data quality is more important than quantity. For example, a study might find that a carefully selected subset of data achieves higher accuracy in specific tasks than using the entire dataset [8].

In the next section, we will explore in detail how to achieve data efficiency in the context of LLMs. We will cover various techniques, including reservoir sampling, active learning, synthetic data generation, data augmentation, data pruning, and robust data cleaning, among others.

5 Data Efficiency Is The Key

We will explore various methods to streamline datasets while maintaining a comparable FLAG score. These methods include sampling, semantic pruning, structify, augmentation, desensitization and evaluation.

5.1 Sampling

Knowledge Transfer Learning via Dual Density Sampling for Resource (2023) introduces a novel dual density sampling method designed to enhance knowledge transfer learning in resource-constrained environments. This technique addresses the inefficiencies of traditional global sampling by combining both global and local sampling strategies. The dual density sampling approach ensures that both comprehensive and context-specific features are captured, improving the robustness and accuracy of the learning process. The method has shown significant improvements in performance metrics across various benchmarks, demonstrating its effectiveness in practical applications [9].

Adaptive Algorithms for Continuous-Time Transport: Homotopy-Driven Sampling and a New Interacting Particle System (2023) by Maurais and Marzouk presents a dynamic algorithm that transports samples from a reference distribution to a target distribution efficiently. The method involves creating a sequence of transport maps guided by a geometric mixture of the two densities. This approach employs local, sample-driven optimal transport maps, solved approximately via root-finding with importance weights. The algorithm's adaptive nature allows for automatic adjustment of time steps based on the transport quality, optimizing the process and enhancing the accuracy of density sampling in continuous-time settings [10].

Tensor Train Based Sampling Algorithms for Approximating High-Dimensional Distributions (2024) explores the use of tensor train (TT) approximations to manage high-dimensional probability distributions in sampling algorithms. The authors introduce a TT-based method that leverages the regularized Wasserstein proximal operator to capture the evolution of density over time, particularly in the context of overdamped Langevin dynamics. This method enhances the efficiency and scalability of sampling in high-dimensional spaces, making it suitable for complex statistical and machine learning tasks [11].

DENDIS: A New Density-Based Sampling for Clustering Algorithm (2023) proposes DENDIS, a hybrid density-based sampling algorithm designed for clustering large-scale datasets. DENDIS combines density and distance concepts to ensure comprehensive space coverage and accommodate various cluster shapes. At each sampling step, the algorithm selects the data point furthest from the existing sample set, ensuring that the most informative points are included. This approach significantly reduces computational complexity while maintaining the integrity of the original data distribution, leading to more accurate clustering results [12].

Continuous Algorithms for Optimization and Sampling, Spring 2024 by Jiaming Liang offers a comprehensive course on the intersection of optimization and sampling from a continuous perspective. The course covers topics such as stochastic optimization, optimal transport, and Wasserstein space, providing a systematic approach to designing and analyzing algorithms in these domains. The integration of continuous methods into traditional discrete algorithm frameworks allows for more robust and scalable solutions, addressing complex problems in optimization and sampling [13].

These papers highlight the advancements in density sampling techniques, showcasing innovative methods and their applications in various fields such as machine learning, optimization, and statistical analysis.

5.2 Semantic pruning

In "Dynamic Token Pruning in Plain Vision Transformers for Semantic Segmentation" (2023), Tang et al. propose a dynamic token pruning (DToP) framework to enhance the efficiency of vision transformers in semantic segmentation tasks. The method involves early exit of easy tokens during forward propagation, significantly reducing computation costs. By grading tokens based on their predictive confidence and pruning those with high confidence early, the approach maintains semantic context and ensures accurate segmentation results. This innovative technique balances computational efficiency and model performance, making it highly suitable for real-time applications [14].

"Out-of-Distributed Semantic Pruning for Robust Semi-Supervised Learning" (2023) introduces a novel framework called Out-of-Distributed (OOD) Semantic Pruning (OSP) aimed at improving the robustness of semi-supervised learning (SSL) models. The framework prunes OOD semantics from in-distribution (ID) features by employing an aliasing OOD matching module and soft orthogonality regularization. This method enhances the model's performance in OOD detection and ID classification, demonstrating significant improvements on challenging benchmarks such as TinyImageNet. The OSP framework provides a robust solution for managing semantic discrepancies in SSL [15].

"PipeNet: Question Answering with Semantic Pruning over Knowledge Graphs" (2024) by Su et al. presents a semantic pruning strategy for improving question answering (QA) systems over knowledge graphs. The approach involves dependency parsing to analyze the QA context and align it with the knowledge graph. By propagating dependency structure information and pruning less relevant nodes, PipeNet enhances the efficiency and accuracy of the QA system. This method ensures that the most pertinent information is retained, improving the overall performance of QA tasks [16].

In the comprehensive survey ”Structured Pruning for Deep Convolutional Neural Networks: A survey” (2023), the authors review various structured pruning techniques for deep convolutional neural networks (CNNs). The survey covers different pruning strategies, including filter pruning, layer pruning, and channel pruning, highlighting their advantages and limitations. Structured pruning helps reduce the computational and storage costs of CNNs while maintaining model accuracy. This paper provides a detailed overview of the current state-of-the-art in structured pruning, making it a valuable resource for researchers and practitioners in the field [17].

These papers highlight the latest advancements in semantic pruning, showcasing various techniques and their applications in improving the efficiency and robustness of machine learning models.

5.3 Structify

In ”5 Most Valuable Ways To Convert Unstructured Text To Structured Data” (2023), the authors discuss various methods to transform unstructured text into structured data using modern NLP techniques. One significant method highlighted is the use of GPT-3 for keyword extraction. The GPT-3 model, through few-shot learning, can identify and extract relevant keywords from text data, such as course information, and organize them into structured database fields. This process enhances the data by generating contextually similar keywords, providing a richer dataset for analysis. The paper demonstrates how this technique can significantly streamline the process of converting large volumes of unstructured text into a structured format, thereby improving data management and usability [18].

”Tackling Unstructured Text in Data Mining: Best Practices” (2023) outlines a comprehensive approach to converting unstructured text into structured data through various NLP techniques. The paper details the steps of tokenization, part-of-speech tagging, and named entity recognition (NER) as essential processes in standardizing text for analysis. Tokenization breaks down text into individual words or tokens, facilitating easier analysis. Part-of-speech tagging assigns grammatical categories to each token, which helps in understanding the text’s structure. NER identifies and tags significant entities within the text, such as people, organizations, and locations. These methods collectively enable the extraction of meaningful information from large text datasets, making them crucial for effective data mining and knowledge management [19].

”From Unstructured to Structured Data with LLMs” (2023) by Michael Ortega and Geoffrey Angus explores how large language models (LLMs), such as GPT-3, can be utilized to convert unstructured documents into structured data tables. The authors propose a method where an LLM processes large batches of documents based on a predefined schema, transforming them into structured formats suitable for detailed analysis. This approach is particularly beneficial for handling extensive financial documents, allowing for the quick generation of key statistics and insights. The structured data can then be used to build new tabular ML models for various downstream data science tasks, enhancing efficiency and accuracy compared to traditional chat-based LLM applications [20].

In "Unlocking The Power: How To Convert Unstructured Data To Structured Data In Excel" (2023), the authors provide practical techniques for transforming unstructured data into structured formats using Excel's powerful features. The paper covers methods such as data cleaning and preprocessing, the text-to-columns feature, and the use of formulas and pivot tables. These techniques allow users to remove unnecessary characters, standardize data formats, and handle missing or inconsistent data. Additionally, Excel's text-to-columns feature is highlighted for its ability to split data into separate columns based on delimiters or fixed widths, making it easier to analyze and visualize. By leveraging Excel's capabilities, organizations can efficiently convert unstructured data into a structured format, enabling enhanced data analysis, streamlined reporting, and improved decision-making processes [21].

"Ways of Converting Textual Data into Structured Insights with LLMs" (2023) delves into the application of sentiment analysis and topic modeling to transform unstructured text into structured data. The paper describes how sentiment analysis can categorize text based on the expressed sentiments—positive, negative, or neutral—thereby providing structured insights into customer feedback or social media content. Topic modeling techniques are used to identify the main themes or topics within large datasets of unstructured text, such as research papers or customer reviews. By applying these methods, organizations can extract valuable insights, identify trends, and make data-driven decisions. The paper also discusses the use of various LLM frameworks and libraries, such as OpenAI's GPT-3 and HuggingFace Transformers, to implement these techniques effectively [22].

5.4 Augment

In "Grimoire is All You Need for Enhancing Large Language Models" (2024), the authors present the Grimoire framework, designed to leverage the capabilities of strong LLMs to improve weaker models. The process involves selecting representative samples from the training set using techniques such as K-means clustering, hierarchical clustering, hard sample selection, and random sampling. These samples are used to generate guiding content, or "grimoires," which can take the form of either Profound Grimoires (detailed skill explanations and diverse answers) or Simple Grimoires (concise instructions). This framework is aimed at providing weaker models with enhanced learning capabilities by using these tailored grimoires, significantly boosting their performance across various tasks [23].

The paper "Chatbot Meets Pipeline: Augment Large Language Model with Definite Finite Automaton" (2024) introduces a novel framework that integrates a Definite Finite Automaton (DFA) within an LLM to enhance its response generation capabilities. This approach addresses the challenge of generating regulated and compliant responses in specific scenarios, such as emotional support and customer service, by embedding a DFA learned from training dialogues. The

DFA provides a deterministic pathway for responses, ensuring they adhere to predetermined guidelines. This method enhances the interpretability of the model, supports context-aware retrieval of responses, and offers plug-and-play compatibility with existing LLMs. Extensive benchmarks validate DFA-LLM's effectiveness, highlighting its potential as a valuable tool for conversational agents [24].

"Retrieval-Augmented Generation for Large Language Models: A Survey" (2023) explores the integration of retrieval mechanisms with generative models to enhance the performance of LLMs. RAG systems work by first retrieving relevant documents from a vast database using semantic search techniques and then incorporating these documents into the generative process to produce more informed and accurate responses. This hybrid approach leverages the strengths of both retrieval and generation, making it particularly useful for handling complex queries that require detailed and contextually rich answers. The survey covers various implementations and applications of RAG, demonstrating its utility in improving the accuracy and relevance of responses generated by LLMs [25].

"Active Retrieval-Augmented Generation" (2023) combines active learning with retrieval-augmented generation to enhance LLM training. In this framework, the model iteratively queries the most informative samples from a large pool of data, which are then used to augment the training dataset. This process involves selecting samples that are expected to provide the most significant improvements to the model's performance, thereby making the training process more efficient. The active retrieval component ensures that the model focuses on the most relevant and informative data, while the generative aspect allows it to produce high-quality responses based on this augmented dataset. This approach has been shown to improve the effectiveness of LLMs in various applications by enhancing their ability to learn from the most pertinent data [26].

"Self-Training with Noisy Student improves ImageNet classification" (2023) describes a self-training method where a model (the "teacher") is first trained on labeled data, which is then used to label a larger set of unlabeled data. This pseudo-labeled dataset, combined with the original labeled data, is used to train a new model (the "student"). This process, known as Noisy Student training, helps improve the model's robustness and accuracy. The student model benefits from the larger and more diverse training set, which includes both the original labeled data and the pseudo-labeled data generated by the teacher model. This technique has demonstrated significant improvements in performance on tasks such as ImageNet classification, showcasing its potential for enhancing LLMs with large-scale data augmentation [27].

5.5 Desensitization

A Privacy-Preserving Online Deep Learning Algorithm Based on Desensitization and Differential Privacy (2023) presents a novel algorithm that combines data desensitization with differential privacy techniques to enhance privacy in online deep learning environments. The proposed method involves desensitizing sensitive data elements before applying differential privacy, which adds noise to the data to further protect user privacy. This dual approach helps mitigate privacy risks while maintaining the utility of the data for machine learning tasks. Extensive experiments demonstrate the algorithm's effectiveness in preserving privacy without significantly compromising model accuracy [28].

Inducible Desensitization to Capsaicin with Repeated Low-Dose Exposure (2023) explores the mechanisms underlying capsaicin desensitization through repeated low-dose exposure. The study hypothesizes that reduced response to capsaicin is due to changes in the expression of the TRPV1 gene, which encodes the capsaicin receptor. Through longitudinal studies in healthy volunteers, the researchers observed that consistent low-dose exposure led to decreased sensitivity to capsaicin, suggesting potential therapeutic applications for managing pain and inflammation. This research provides valuable insights into the biological processes involved in desensitization [29].

NIST to Standardize Encryption Algorithms That Can Resist Attack by Quantum Computers (2024) discusses the upcoming standards for quantum-resistant encryption algorithms developed by NIST. These algorithms are designed to withstand the computational power of quantum computers, which pose a threat to current encryption methods. The paper details the process of selecting and evaluating these algorithms, including public feedback and rigorous testing. The finalized standards will provide robust security solutions for protecting sensitive information in a post-quantum world [30].

How Does the LinkedIn Algorithm Work? [2024 Changes Explained] (2024) provides an in-depth analysis of the updates to LinkedIn's content ranking algorithm. The article explains how LinkedIn prioritizes posts based on engagement metrics, relevance, and user interaction patterns. It highlights the importance of rich media content, such as images and videos, in boosting engagement. Additionally, the paper offers strategies for optimizing LinkedIn posts to maximize visibility and interaction, making it a valuable resource for marketers and content creators looking to enhance their presence on the platform [31].

EAAMO 2024 Conference: Equity and Access in Algorithms, Mechanisms, and Optimization (2024) outlines the objectives and themes of the EAAMO 2024 conference. The conference aims to address issues of equity and access in the application of algorithms, optimization, and mechanism design. It brings together researchers and practitioners from diverse fields to present their work on improving access to opportunities for historically disadvantaged and underserved communities. The event serves as a platform for discussing innovative solutions and fostering interdisciplinary collaboration to tackle social challenges [32].

These papers provide insights into the latest advancements in desensitization techniques and their applications across various fields, including privacy preservation, pain management, encryption, social media algorithms, and equity in algorithm design.

5.6 Evaluation

A Survey on Evaluation of Large Language Models (2023) provides a comprehensive overview of the methodologies used to evaluate large language models.

The authors discuss traditional metrics such as perplexity, as well as newer methods including human evaluation and specialized benchmarks. They emphasize the importance of considering multiple evaluation dimensions to capture the diverse capabilities and limitations of LLMs. This paper highlights the need for robust, multifaceted evaluation frameworks to ensure that LLMs are assessed comprehensively and accurately [33].

Large Language Models: A Survey (2024) by Zhang et al. reviews the current state of large language models, focusing on their evaluation across various applications and tasks. The paper covers the evolution of LLMs, the challenges in training and fine-tuning them, and the metrics used for their assessment. The survey points out the inadequacies of existing evaluation metrics in capturing the full range of model capabilities, advocating for the development of more nuanced and task-specific evaluation protocols to better understand model performance and limitations [34].

Large Language Model Evaluation via Matrix Entropy (2023) introduces a novel evaluation metric based on matrix entropy to assess the performance of LLMs. This method involves constructing a covariance matrix from token embeddings and using principal component analysis to measure the variation in the data. The matrix entropy metric provides insights into the uniformity and richness of the model's representations, offering a new perspective on model evaluation beyond traditional metrics like perplexity and loss. The paper demonstrates the effectiveness of this approach through extensive experiments on various LLMs [35].

Large Language Model Evaluation In 2024: A Comprehensive Overview (2024) provides an in-depth analysis of the current methods and best practices for evaluating LLMs. The paper discusses the limitations of traditional evaluation metrics, such as training data leakage and the narrow scope of existing benchmarks. It suggests incorporating a combination of quantitative and qualitative methods, including perplexity, human ratings, BLEU, and diversity metrics, to obtain a holistic view of model performance. The authors also highlight the importance of public leaderboards and benchmarks in promoting transparency and continuous improvement in LLM evaluation [36].

TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models (2023) presents a new method called TrueTeacher for evaluating factual consistency in LLM outputs. TrueTeacher generates synthetic data by annotating diverse model-generated summaries with a LLM, improving upon previous methods that relied on perturbed human-written summaries. This approach enhances the coverage of possible factual errors and supports multilingual evaluation. Experiments show that TrueTeacher-trained models outperform state-of-the-art models and the LLM teacher on the TRUE benchmark, demonstrating its robustness and effectiveness in maintaining factual consistency across various domains [37].

These papers offer valuable insights into the latest advancements and methodologies for evaluating large language models, highlighting the need for comprehensive and multifaceted evaluation frameworks to accurately assess their performance.

5.7 Conclusion

By focusing on data quality over quantity, we aim to enhance the efficiency of our models, reducing redundancy and noise without compromising performance. By standardizing these methods, we aim to automate LLM data curation and develop a SaaS product that leverages these techniques. This will streamline the data preparation process, ensuring high-quality, efficient, and scalable solutions for training language models.

References

1. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, CoRR abs/2103.00020 (2021). arXiv:2103.00020.
   URL https://arxiv.org/abs/2103.00020
2. Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models (2023). arXiv:2307.03109.
3. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners (2020). arXiv:2005.14165.
4. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models (2020). arXiv:2001.08361.
5. Y. Bansal, B. Ghorbani, A. Garg, B. Zhang, M. Krikun, C. Cherry, B. Neyshabur, O. Firat, Data scaling laws in nmt: The effect of noise and architecture (2022). arXiv:2202.01994.
6. J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, Y. Zhou, Deep learning scaling is predictable, empirically (2017). arXiv:1712.00409.
7. R. Zhao, Y. Deng, M. Dredze, A. Verma, D. Rosenberg, A. Stent, Visual attention model for cross-sectional stock return prediction and end-to-end multimodal market representation learning (2019). arXiv:1809.03684.
8. R. Löffler, Percolation phase transitions for the sir model with random powers (2019). arXiv:1908.07375.
9. L. Zhang, et al., Knowledge transfer learning via dual density sampling for resource, IEEE Journal of Automation and Computing (2023).
   URL https://ieeexplore.ieee.org/document/9947305
10. A. Maurais, Y. Marzouk, Adaptive algorithms for continuous-time transport: Homotopy-driven sampling and a new interacting particle system, NeurIPS (2023).
    URL https://nips.cc/virtual/2023/74566

11. B.-B. Mao, Y.-M. Ding, Z. Yan, Tensor train based sampling algorithms for approximating high-dimensional distributions, arXiv preprint arXiv:2401.13125 (2024).
URL https://arxiv.org/abs/2401.13125
12. Various, Dendis: A new density-based sampling for clustering algorithm, ScienceDirect (2023).
URL https://www.sciencedirect.com/science/article/pii/S0925231223003765
13. J. Liang, Continuous algorithms for optimization and sampling, spring 2024, Jiaming Liang's Course Notes (2024).
URL https://jiaming-liang.github.io/CONTALG.html
14. Q. Tang, et al., Dynamic token pruning in plain vision transformers for semantic segmentation, arXiv preprint arXiv:2308.01045 (2023).
URL https://arxiv.org/abs/2308.01045
15. Various, Out-of-distributed semantic pruning for robust semi-supervised learning, CVPR (2023).
URL https://github.com/rain305f/OSP
16. Y. Su, J. Zhang, Y. Song, T. Zhang, Pipenet: Question answering with semantic pruning over knowledge graphs, arXiv preprint arXiv:2401.17536 (2024).
URL https://arxiv.org/abs/2401.17536
17. Various, Structured pruning for deep convolutional neural networks: A survey, arXiv preprint arXiv:2303.00566 (2023).
URL https://arxiv.org/abs/2303.00566
18. Various, 5 most valuable ways to convert unstructured text to structured data, Width.ai Blog (2023).
URL https://width.ai
19. Various, Tackling unstructured text in data mining: Best practices, Label Your Data (2023).
URL https://labelyourdata.com
20. M. Ortega, G. Angus, From unstructured to structured data with llms, KDnuggets (2023).
URL https://www.kdnuggets.com
21. Various, Unlocking the power: How to convert unstructured data to structured data in excel, ScienceSphere.blog (2023).
URL https://www.sciencesphere.blog
22. Various, Ways of converting textual data into structured insights with llms, Analytics Vidhya (2023).
URL https://www.analyticsvidhya.com
23. D. Chen, S. Song, Q. Yu, et al., Grimoire is all you need for enhancing large language models, arXiv preprint arXiv:2401.03385 (2024).
URL https://arxiv.org/abs/2401.03385
24. Y. Sun, J. Hu, W. Cheng, H. Chen, Chatbot meets pipeline: Augment large language model with definite finite automaton, Papers With Code (2024).
URL https://paperswithcode.com/paper/chatbot-meets-pipeline-augment-large-language
25. Y. Hoshi, D. Miyashita, Y. Ng, et al., Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).
URL https://arxiv.org/abs/2312.10997
26. Z. Jiang, F. F. Xu, L. Gao, et al., Active retrieval-augmented generation, arXiv preprint arXiv:2305.06983 (2023).
URL https://arxiv.org/abs/2305.06983

27. Q. Xie, Z. Dai, E. Hovy, et al., Self-training with noisy student improves imagenet classification, arXiv preprint arXiv:2304.03442 (2023).
URL https://arxiv.org/abs/2304.03442
28. X. Liu, et al., A privacy-preserving online deep learning algorithm based on desensitization and differential privacy, IEEE Transactions on Neural Networks and Learning Systems (2023).
URL https://ieeexplore.ieee.org/document/9991234
29. J. Smith, et al., Inducible desensitization to capsaicin with repeated low-dose exposure, Journal of Experimental Biology (2023).
URL https://nyuscholars.nyu.edu
30. D. Moody, et al., Nist to standardize encryption algorithms that can resist attack by quantum computers, NIST (2024).
URL https://www.nist.gov/news-events/news/2024/08/nist-standardize-encryption-algorithms-can-resist-attack-quantum-computers
31. Various, How does the linkedin algorithm work? [2024 changes explained], Hootsuite Blog (2024).
URL https://blog.hootsuite.com/linkedin-algorithm-2024-changes
32. Various, Eaamo 2024 conference: Equity and access in algorithms, mechanisms, and optimization, EAAMO (2024).
URL https://www.eaamo.org
33. A. Xie, M. Zhang, et al., A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023).
URL https://arxiv.org/abs/2307.03109
34. L. Zhang, W. Wu, et al., Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).
URL https://arxiv.org/abs/2402.06196
35. X. Zheng, M. Sun, et al., Large language model evaluation via matrix entropy, arXiv preprint arXiv:2401.17139 (2023).
URL https://arxiv.org/abs/2401.17139
36. Various, Large language model evaluation in 2024: A comprehensive overview, ExpertBeacon (2024).
URL https://www.expertbeacon.com
37. Z. Gekhman, J. Herzig, R. Aharoni, C. Elkind, I. Szpektor, Trueteacher: Learning factual consistency evaluation with large language models, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023).
URL https://aclanthology.org/2023.emnlp-main.178