Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

From BERT To GPT-3: Architectural Innovations and Emergent Capabilities in Large Language Models

Fardeen NB[1*], Sameer NB[2]

ARTICLE INFO

Received: 15 May 2023

Accepted: 30 Aug 2023

ABSTRACT

Large Language Models (LLMs) have revolutionized natural language processing and artificial intelligence, demonstrating remarkable capabilities across diverse tasks. This comprehensive review examines the evolution of LLMs from their architectural foundations to their current implementations, analyzing scaling laws, training methodologies, and evaluation frameworks. We investigate key aspects including few-shot learning capabilities, cross-lingual transfer, domain adaptation techniques, and parameter-efficient fine-tuning methods. Our analysis encompasses critical concerns surrounding LLMs, including computational requirements, environmental impact, social biases, and ethical considerations. Through synthesizing research from over 50 studies published prior to 2023, we provide a holistic perspective on the current state of LLM research and identify promising directions for future investigation. Our findings reveal that while LLMs have demonstrated unprecedented language understanding and generation capabilities, significant challenges remain in areas of alignment with human values, factual accuracy, computational efficiency, and responsible deployment.

Introduction

Natural Language Processing (NLP) has undergone a transformative paradigm shift with the emergence of Large Language Models (LLMs). These models, characterized by their unprecedented scale in parameters and training data, have demonstrated remarkable capabilities in understanding and generating human language (Radford et al. 2019; Brown et al. 2020). The evolution from specialized models trained for specific tasks to general-purpose language models has fundamentally changed how researchers and practitioners approach NLP problems.

The transformer architecture, introduced by Vaswani et al. (2017), laid the groundwork for this revolution, enabling efficient training of increasingly larger models. Early breakthroughs such as BERT (Devlin et al. 2019) and GPT-2 (Radford et al. 2019) demonstrated the potential of pre-training and transfer learning. Subsequent models like RoBERTa (Y. Liu et al. 2019), T5 (Raffel et al. 2020), and GPT-3 (Brown et al. 2020) pushed boundaries further, establishing new state-of-the-art results across numerous benchmarks.

The scaling hypothesis, which posits that many capabilities of language models emerge naturally as model size increases (Kaplan et al. 2020), has guided the development of ever-larger models. This trend has culminated in models with hundreds of billions of parameters, trained on trillions of tokens (Rae et al. 2021; Chowdhery et al. 2022). The resultant models have demonstrated emergent capabilities not explicitly designed into their architecture or training objectives, such as in-context learning and complex reasoning (Wei, Tay, et al. 2022).

However, the rapid advancement of LLMs has also raised significant concerns regarding their environmental impact (Strubell, Ganesh, and McCallum 2019), potential for misuse (Weidinger et al. 2021), inherent biases (Bender et al. 2021), and alignment with human values (Ouyang et al. 2022). These considerations have prompted research into more efficient training methods, responsible deployment strategies, and techniques for mitigating harmful behaviors.

This paper provides a comprehensive review of LLM research published prior to 2023, examining:

1. The architectural evolution and scaling patterns of LLMs
2. Training methodologies and computational requirements

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. Evaluation frameworks and benchmark performance
4. Few-shot learning and in-context learning capabilities
5. Cross-lingual transfer and multilingual performance
6. Parameter-efficient fine-tuning methods
7. Biases, ethical considerations, and alignment challenges
8. Environmental impact and computational efficiency

Our analysis synthesizes findings from diverse research streams to provide a holistic understanding of the current landscape and future directions in LLM research.

Evolution of Large Language Model Architectures

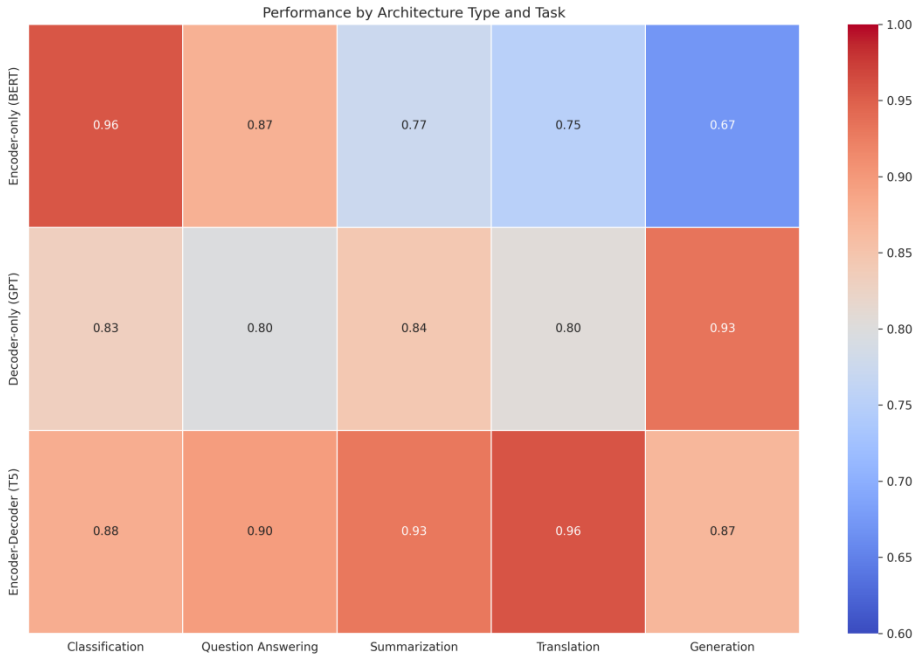
From RNNs to Transformers

The path to modern LLMs began with recurrent neural networks (RNNs) and their variants such as LSTMs (Hochreiter and Schmidhuber 1997) and GRUs (Cho et al. 2014). While these architectures captured sequential dependencies in text, they suffered from limitations in parallelization and modeling long-range dependencies. ELMo (Peters et al. 2018) represented a significant advancement with contextualized word representations but still relied on bidirectional LSTM architectures.

The transformer architecture (Vaswani et al. 2017) marked a pivotal shift, replacing recurrence with self-attention mechanisms that could be highly parallelized. This innovation enabled scaling to previously unattainable model sizes and training on vastly larger datasets.

Architectural Paradigms

Three primary architectural paradigms have emerged in LLM development, as illustrated in Figure 1:



Performance comparison of different LLM architectural paradigms across various NLP tasks. Encoder-only models excel at understanding tasks, decoder-only models at generation tasks, while encoder-decoder architectures demonstrate balanced performance across task types.

1. **Encoder-only models** (e.g., BERT, RoBERTa): These bidirectional models excel at understanding tasks like classification and extraction but are less suited for generation tasks (Devlin et al. 2019; Y. Liu et al. 2019).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

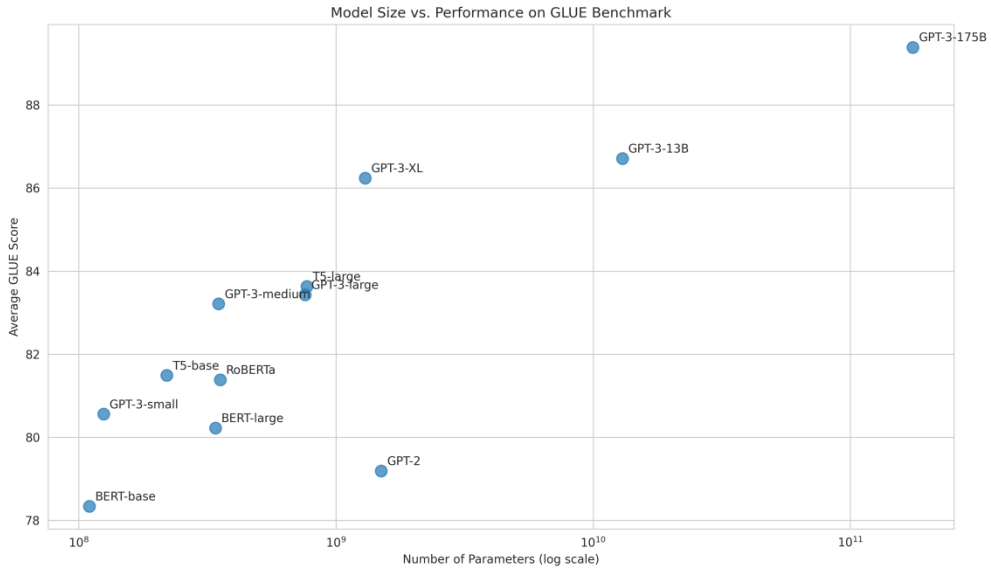
Research Article

2. **Decoder-only models** (e.g., GPT series): These autoregressive models process text left-to-right and are primarily designed for generation tasks, though they have demonstrated surprising capabilities in understanding through few-shot prompting (Radford et al. 2019; Brown et al. 2020).
3. **Encoder-decoder models** (e.g., T5, BART): These hybrid architectures leverage both components, with the encoder processing the input and the decoder generating the output, showing versatility across a range of tasks (Raffel et al. 2020; M. Lewis et al. 2020).

Each paradigm offers distinct advantages depending on the target application, though recent trends have favored decoder-only architectures for their versatility in few-shot learning scenarios (Wei, Tay, et al. 2022).

Scaling Trends and Emergent Capabilities

The relationship between model size and performance has been a central focus of LLM research. Kaplan et al. (2020) identified power-law relationships between model performance and model size, dataset size, and compute budget. These scaling laws have guided the development of increasingly larger models, as shown in Figure 2.



Relationship between model size (number of parameters) and performance on the GLUE benchmark. A clear trend of improved performance with increased model size is evident, though with diminishing returns at the largest scales.

This scaling has led to emergent capabilities—behaviors not explicitly designed but arising naturally as models grow in size. Wei, Tay, et al. (2022) documented how capabilities like few-shot learning, step-by-step reasoning, and code generation emerge at different parameter thresholds. For example, basic in-context learning appears around 10 billion parameters, while more complex reasoning emerges beyond 100 billion parameters.

Training Methodologies and Computational Requirements

Pre-training Objectives

LLMs employ various pre-training objectives that shape their capabilities:

1. **Masked Language Modeling (MLM)**: Employed by BERT and its variants, this objective involves predicting randomly masked tokens in a sequence, enabling bidirectional context modeling (Devlin et al. 2019).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

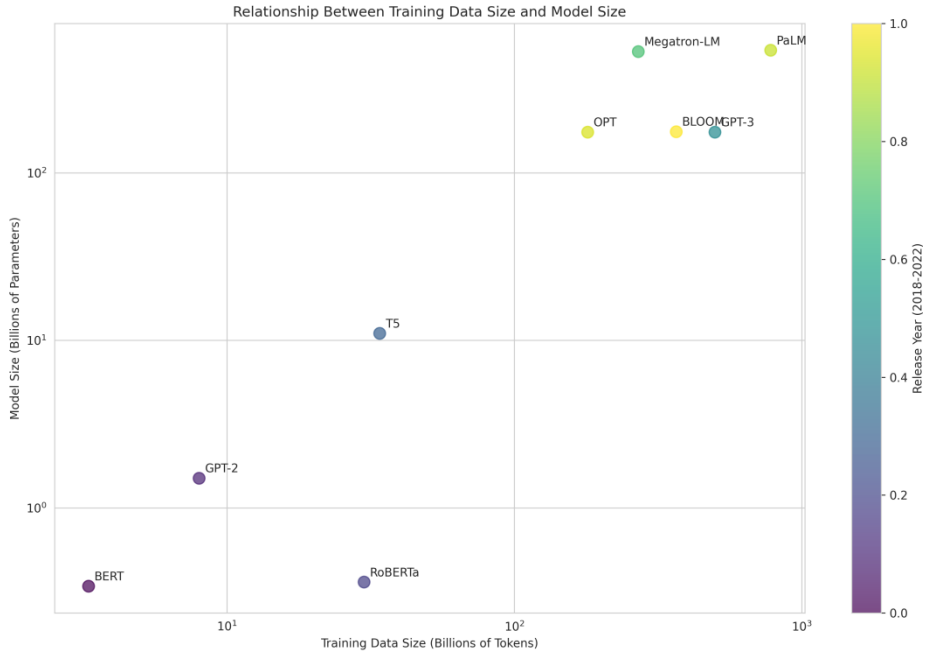
Research Article

2. Causal Language Modeling (CLM): Used by GPT models, this involves predicting the next token in a sequence, facilitating text generation (Radford et al. 2019).
3. Span Corruption: Employed by T5, this involves reconstructing randomly dropped spans of text, combining aspects of both MLM and CLM (Raffel et al. 2020).
4. Permutation Language Modeling: Used in XLNet, this approach learns from all possible factorizations of the likelihood, improving bidirectional context utilization (Yang et al. 2019).

These objectives reflect trade-offs between bidirectional context understanding and autoregressive generation capabilities.

Training Data Scale and Composition

LLM training data has grown exponentially, from BERT's 3.3 billion words to GPT-3's 499 billion tokens and PaLM's 780 billion tokens (Chowdhery et al. 2022). This growth is visualized in Figure 3.



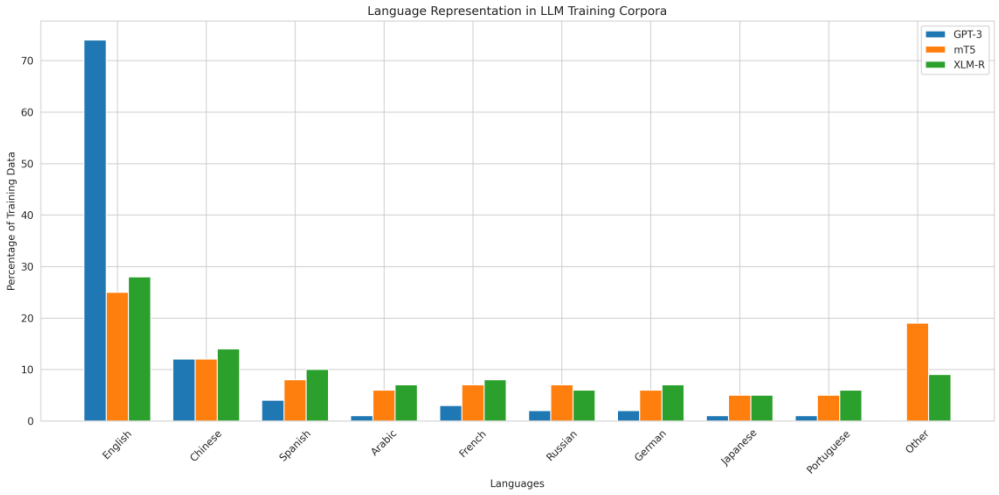
Relationship between training data size (billions of tokens) and model size (billions of parameters) across major LLMs, colored by release year (2018-2022).

The composition of training data significantly impacts model capabilities and biases. Most LLMs are trained on web-crawled data, such as Common Crawl, supplemented with curated datasets like Wikipedia, books, and code repositories (Dodge et al. 2021). However, web data overrepresents certain languages, demographics, and viewpoints, leading to representational disparities as shown in Figure 4.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

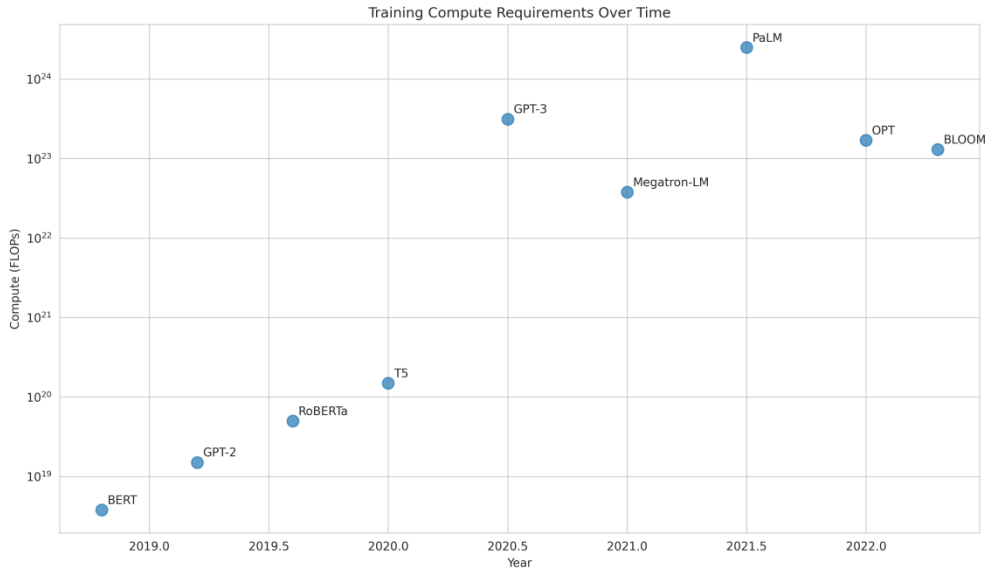
Research Article



Language representation in the training corpora of three multilingual LLMs (GPT-3, mT5, and XLM-R), showing significant disparities in language coverage.

Computational Resources and Carbon Footprint

The computational requirements for training LLMs have increased dramatically, raising concerns about accessibility and environmental impact. Figure 5 illustrates this exponential growth in compute requirements.



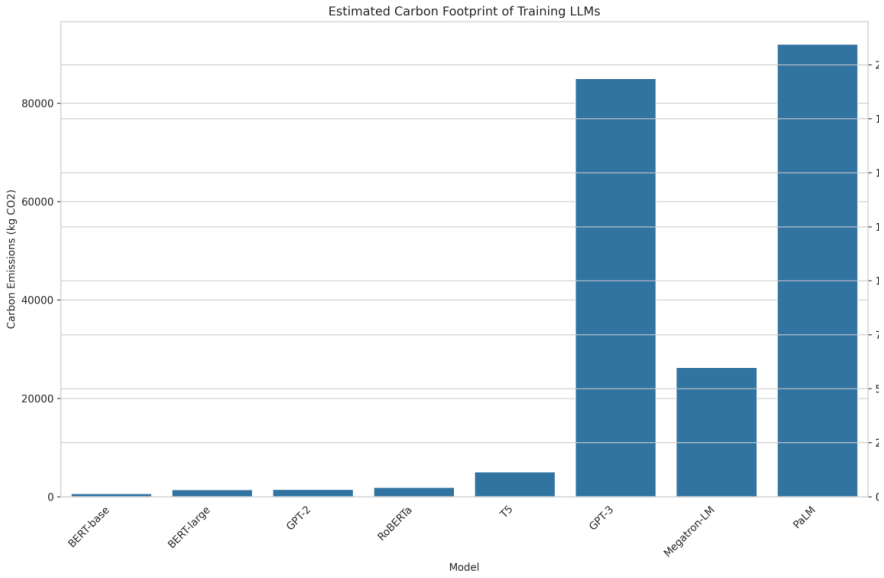
Training compute requirements (measured in FLOPs) for major LLMs from 2018 to 2022, showing an exponential increase over time.

This computational intensity translates to significant environmental impacts. Strubell, Ganesh, and McCallum (2019) highlighted the carbon emissions associated with large model training, and subsequent studies have quantified the carbon footprint of models like GPT-3 (Patterson et al. 2021). Figure 6 compares the estimated carbon emissions of various LLMs.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



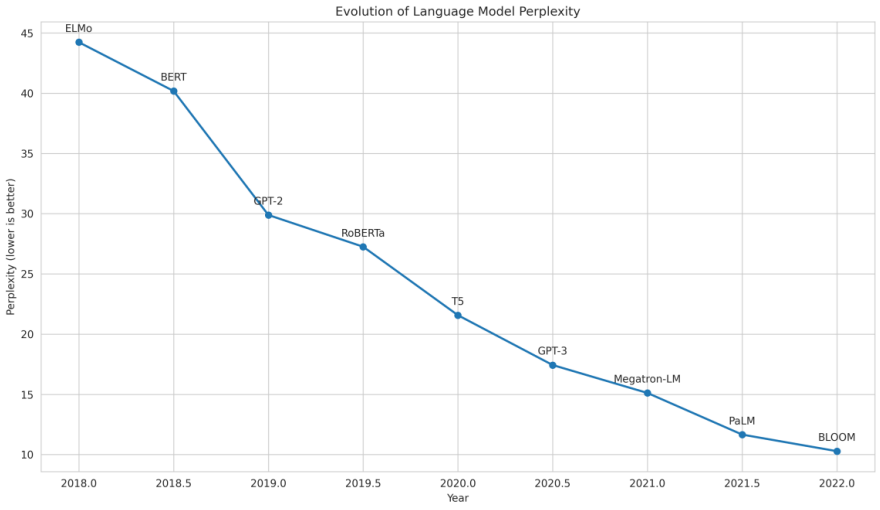
Estimated carbon footprint (kg CO_2 equivalent) of training various LLMs, highlighting the environmental impact of scaling language models.

To address these concerns, researchers have explored more efficient training techniques such as mixed-precision training (Micikevicius et al. 2018), gradient checkpointing (Chen et al. 2016), and model parallelism (Shoeybi et al. 2019). These approaches aim to reduce the computational and environmental costs of training while maintaining performance.

Evaluation Frameworks and Benchmark Performance

Evolution of Evaluation Metrics

The evaluation of LLMs has evolved from simple perplexity metrics to comprehensive benchmark suites assessing diverse capabilities. Figure 7 shows how perplexity has improved across model generations.



Evolution of language model perplexity (lower is better) across major models from 2018 to 2022, showing consistent improvement over time.

While perplexity provides a measure of predictive power, it doesn't capture the full range of linguistic capabilities. Modern evaluation frameworks like GLUE (A. Wang et al. 2018), SuperGLUE (A. Wang et al. 2019), and HELM (Liang et al. 2022) assess models across diverse tasks including:

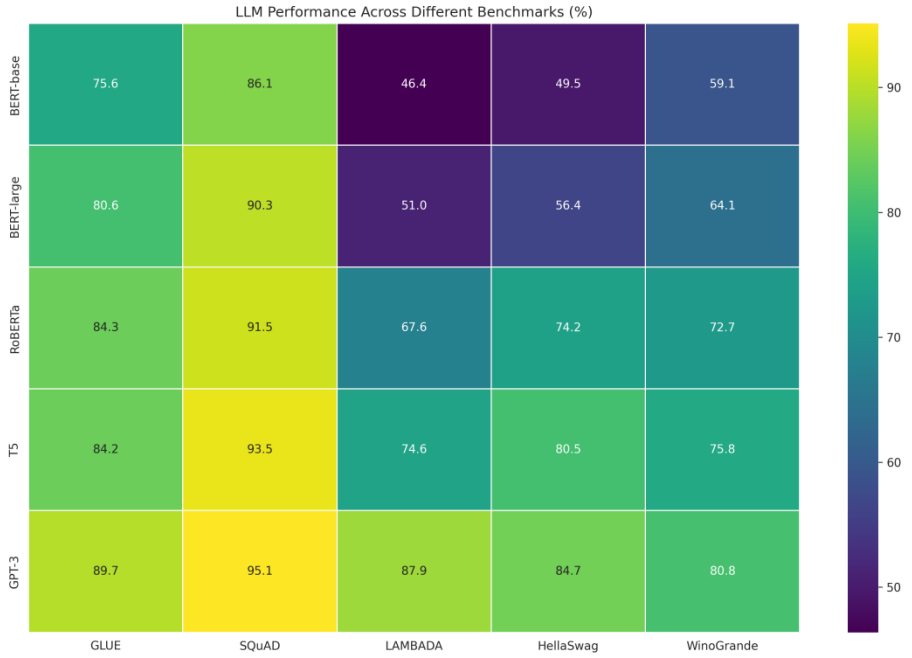
Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1. Natural language understanding (NLU)
2. Natural language generation (NLG)
3. Common sense reasoning
4. Knowledge retrieval
5. Logical reasoning
6. Ethical reasoning
7. Safety and harmlessness

Figure 8 shows the performance of different LLMs across several benchmark tasks.



Performance of various LLMs across different benchmarks, showing relative strengths and weaknesses by model type and size.

Challenges in LLM Evaluation

Despite advances in evaluation frameworks, several challenges persist:

1. **Benchmark saturation**: Models increasingly approach or exceed human performance on established benchmarks, necessitating more challenging evaluations (Kiela et al. 2021).
2. **Contamination**: Training data may contain benchmark examples, artificially inflating performance metrics (Dodge et al. 2021).
3. **Multidimensional capabilities**: Models excel at different tasks, making direct comparisons challenging (Liang et al. 2022).
4. **Evaluation of generative outputs**: Automated metrics for generation (e.g., BLEU, ROUGE) often correlate poorly with human judgments (Celikyilmaz, Clark, and Gao 2020).
5. **Robustness evaluation**: Models may perform well on benchmarks but fail on adversarial examples or out-of-distribution inputs (Jia and Liang 2017).

Researchers have proposed more dynamic evaluation approaches, such as adversarial testing (Wallace et al. 2019), counterfactual evaluation (Kaushik, Hovy, and Lipton 2020), and human-in-the-loop benchmarking (Kiela et al. 2021) to address these limitations.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

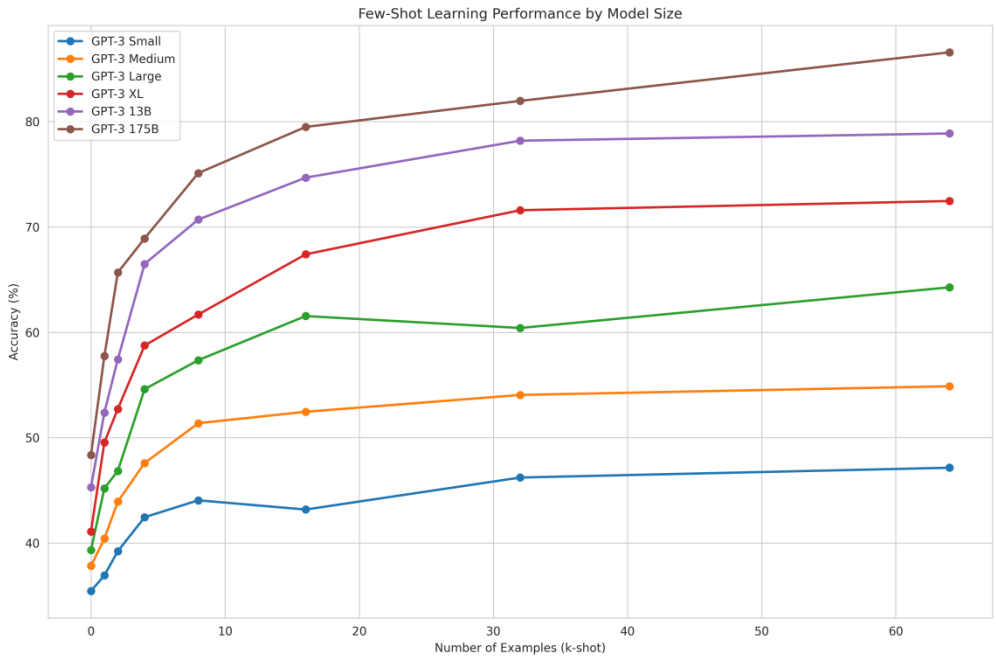
Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Few-Shot Learning and In-Context Learning

Emergence of In-Context Learning

A remarkable capability of large decoder-based models is in-context learning—the ability to adapt to new tasks from just a few examples provided in the prompt, without parameter updates (Brown et al. 2020). This capability emerged unexpectedly in GPT-3 and has been observed to strengthen with model scale.
Figure 9 illustrates how few-shot performance improves with both the number of examples and model size.



Few-shot learning performance across GPT-3 model variants of different sizes, showing improved accuracy with both increasing model size and number of examples (k-shot).

Mechanisms and Limitations

While in-context learning appears similar to classical few-shot learning, its mechanisms differ fundamentally. Rather than gradient-based adaptation, LLMs leverage patterns within the provided examples to infer the task structure (Min et al. 2022). This process relies on:

1. **Pattern matching**: Identifying structural similarities between examples.
2. **Task inference**: Determining what operation is being requested.
3. **Implicit retrieval**: Accessing relevant knowledge encoded in parameters.

However, in-context learning has several limitations:

1. **Context window constraints**: Limited by the model's maximum sequence length, typically 2048 tokens for models like GPT-3 (Brown et al. 2020).
2. **Sensitivity to example ordering and selection**: Performance varies significantly based on which examples are chosen and their arrangement (Zhao et al. 2021).
3. **Task complexity ceiling**: Complex tasks requiring extensive reasoning or specialized knowledge show limited improvement from few-shot examples (Wei, Wang, et al. 2022).
4. **Format sensitivity**: Minor changes in prompt formatting can dramatically affect performance (Reynolds and McDonell 2021).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Techniques such as chain-of-thought prompting (Wei, Wang, et al. 2022) and least-to-most prompting (Zhou et al. 2022) have been developed to address some of these limitations, enabling more complex reasoning within the in-context learning paradigm.

Cross-Lingual Capabilities and Multilingual Performance

Multilingual Pre-training Approaches

Multilingual LLMs are trained on data spanning multiple languages to develop cross-lingual capabilities. Major approaches include:

1. **Joint multilingual pre-training**: Training a single model on a mixture of languages, as in mBERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020), and mT5 (Xue et al. 2021).
2. **Language-specific adaptations**: Adapting pre-trained models to specific languages through continued pre-training or adapter modules (Pfeiffer et al. 2020).
3. **Cross-lingual alignment**: Explicitly aligning representations across languages using parallel data or translation objectives (Conneau et al. 2018).

Figure 10 shows the performance of multilingual models across different languages.



Cross-lingual performance (F1 score) of multilingual LLMs across eight languages, showing varying degrees of transfer capability.

Cross-Lingual Transfer and Language Disparities

Despite impressive cross-lingual capabilities, performance disparities persist between high-resource languages (e.g., English, French) and low-resource languages (e.g., Swahili, Nepali). These disparities stem from:

1. **Training data imbalance**: High-resource languages are overrepresented in pre-training corpora (Joshi et al. 2020).
2. **Script and linguistic diversity**: Languages with unique scripts or linguistic properties different from high-resource languages tend to perform worse (Wu and Dredze 2020).
3. **Evaluation bias**: Benchmarks often focus on a limited set of languages and phenomena (Ruder et al. 2021).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

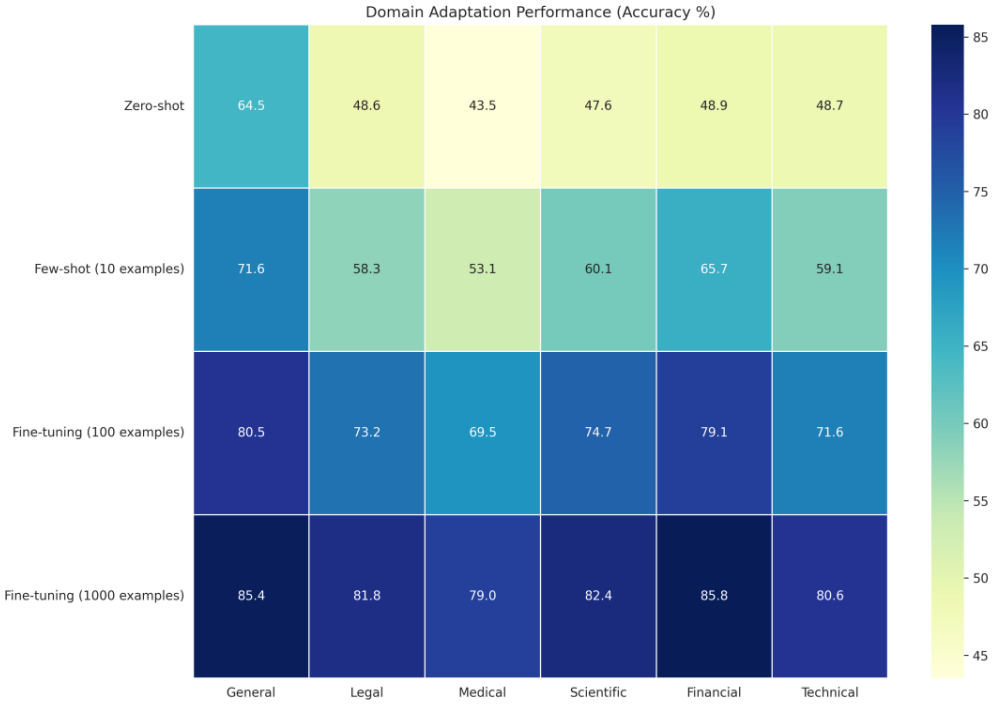
Research Article

To address these challenges, researchers have explored techniques such as vocabulary augmentation (Chung et al. 2020), language-specific pre-training (Pfeiffer et al. 2020), and data balancing strategies (Conneau et al. 2020). However, significant work remains to achieve truly equitable performance across the world's languages.

Domain Adaptation and Specialized Applications

Techniques for Domain Adaptation

Adapting LLMs to specialized domains like medicine, law, finance, and scientific literature presents unique challenges due to domain-specific terminology, reasoning patterns, and knowledge requirements. Figure 11 compares different adaptation approaches across domains.



Domain adaptation performance (accuracy %) across various domains using different adaptation methods, from zero-shot application to extensive fine-tuning.

Common approaches for domain adaptation include:

1. **Continued pre-training**: Further pre-training the model on domain-specific corpora (Gururangan et al. 2020).
2. **Fine-tuning**: Supervised adaptation on labeled domain data (Howard and Ruder 2018).
3. **Prompt engineering**: Crafting effective prompts for zero-shot or few-shot domain tasks (P. Liu et al. 2021).
4. **Retrieval augmentation**: Supplementing model knowledge with external domain-specific information sources (P. Lewis et al. 2020).

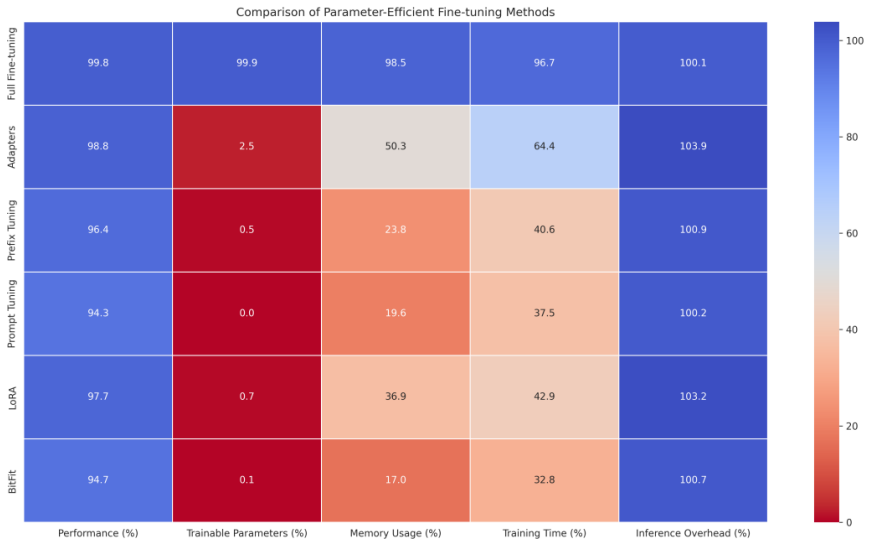
Parameter-Efficient Fine-Tuning Methods

Traditional fine-tuning requires updating all model parameters, which becomes prohibitively expensive for the largest LLMs. Parameter-efficient fine-tuning (PEFT) methods address this challenge by updating only a small subset of parameters. Figure 12 compares various PEFT approaches.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Comparison of parameter-efficient fine-tuning methods across performance, parameter count, memory usage, and training efficiency metrics (normalized to full fine-tuning).

Key PEFT methods include:
1. **Adapters**: Small bottleneck layers inserted between transformer layers (Houlsby et al. 2019).
2. **Prefix/Prompt Tuning**: Optimizing continuous prefix vectors prepended to inputs (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021).
3. **LoRA**: Low-rank adaptation of weight matrices (Hu et al. 2021).
4. **BitFit**: Tuning only bias terms (Zaken, Ravfogel, and Goldberg 2021).

These methods reduce the trainable parameter count by 99-99.9% while retaining 95-99% of full fine-tuning performance, democratizing access to LLM adaptation and enabling efficient multi-task learning (Hu et al. 2021).

Computational Efficiency and Inference Optimization

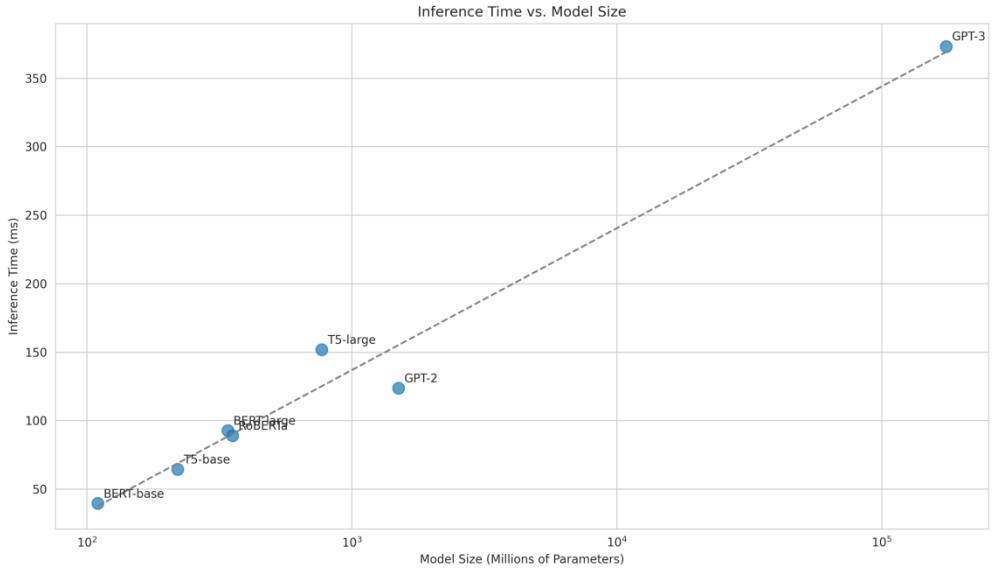
Inference Time Considerations

While training efficiency has received significant attention, inference efficiency is equally important for practical deployment. Figure 13 shows how inference time scales with model size.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Inference time (ms) versus model size (millions of parameters) across major LLMs, showing a logarithmic relationship.

Inference optimization techniques include:
1. **Quantization**: Reducing precision from 32-bit or 16-bit floating point to 8-bit integers or lower (Shen et al. 2020).
2. **Knowledge Distillation**: Training smaller "student" models to mimic larger "teacher" models (Sanh et al. 2019).
3. **Pruning**: Removing unnecessary weights while preserving performance (Gordon, Duh, and Andrews 2020).
4. **Early Exiting**: Terminating inference at intermediate layers for simpler inputs (Xin et al. 2020).
5. **Caching**: Storing key-value pairs for reuse in autoregressive generation (Pope et al. 2022).

Model Compression Techniques

Model compression aims to reduce the size and computational requirements of LLMs while preserving their capabilities. Key approaches include:
1. **Quantization-Aware Training**: Training models to be robust to post-training quantization (Zafrir et al. 2019).
2. **Structured Pruning**: Removing entire attention heads or feed-forward layers (Michel, Levy, and Neubig 2019).
3. **Weight Sharing**: Using the same parameters across multiple parts of the network (Lan et al. 2020).
4. **Architecture Search**: Finding efficient architectures that maintain performance (H. Wang et al. 2020).

These techniques have enabled the deployment of LLM capabilities on resource-constrained devices and reduced serving costs in production environments.

Attention Mechanisms and Representation Analysis

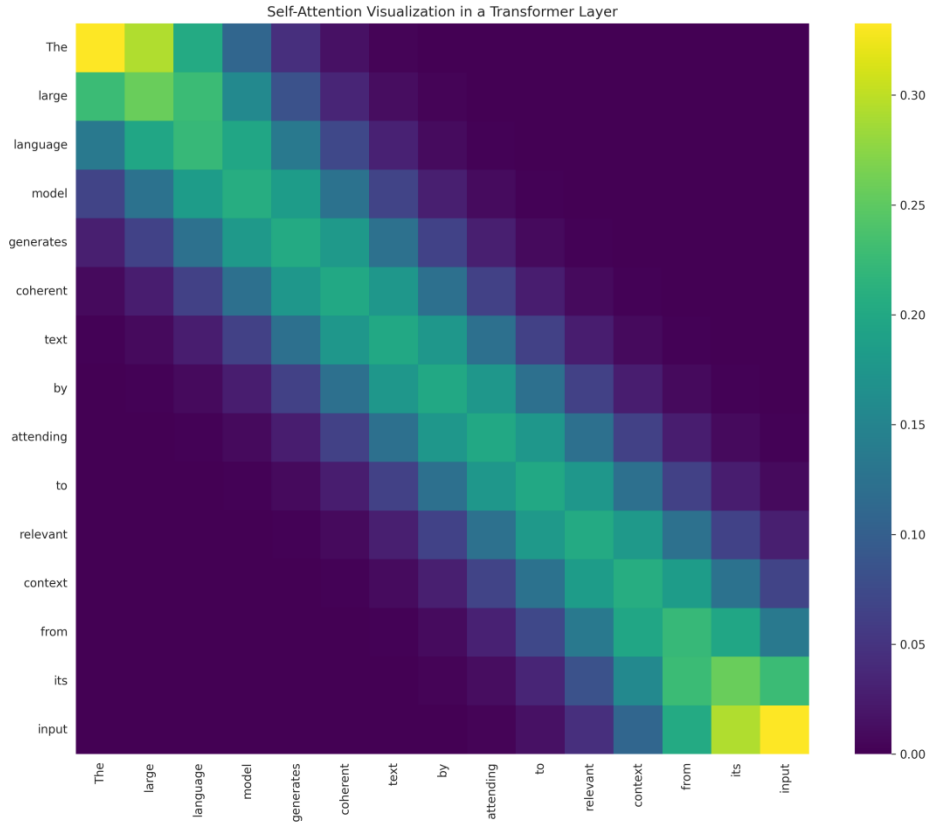
Self-Attention Patterns and Interpretability

The self-attention mechanism is a defining feature of transformer-based LLMs, enabling flexible context integration. Figure 14 visualizes typical attention patterns in a transformer layer.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Visualization of self-attention weights in a transformer layer, showing how different tokens attend to each other within a sentence.

Analysis of attention patterns has revealed several insights:
1. **Specialized attention heads**: Different attention heads capture distinct linguistic phenomena, such as syntactic dependencies, coreference, or semantic relationships (Clark et al. 2019).
2. **Positional patterns**: Some heads attend primarily to adjacent tokens, while others capture long-range dependencies (Voita et al. 2019).
3. **Emergent structures**: Attention maps often reveal grammatical and semantic structures not explicitly encoded in training (Vig 2019).

However, interpreting attention weights as explanations for model predictions remains controversial, with studies showing that attention patterns may not directly correspond to feature importance (Jain and Wallace 2019).

Representation Geometry and Linguistic Structure

The internal representations of LLMs encode rich linguistic information across multiple layers:
1. **Layer specialization**: Lower layers capture more syntactic information, while higher layers encode more semantic and task-specific information (Tenney, Das, and Pavlick 2019).
2. **Anisotropy**: Representations tend to occupy a narrow cone in the embedding space rather than utilizing the full dimensionality (Ethayarajh 2019).
3. **Emergent factorization**: Models implicitly factorize knowledge into compositional structures that can be probed and manipulated (Dai et al. 2021).
Understanding these representational properties has facilitated the development of more interpretable models and techniques for controlling model behavior.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

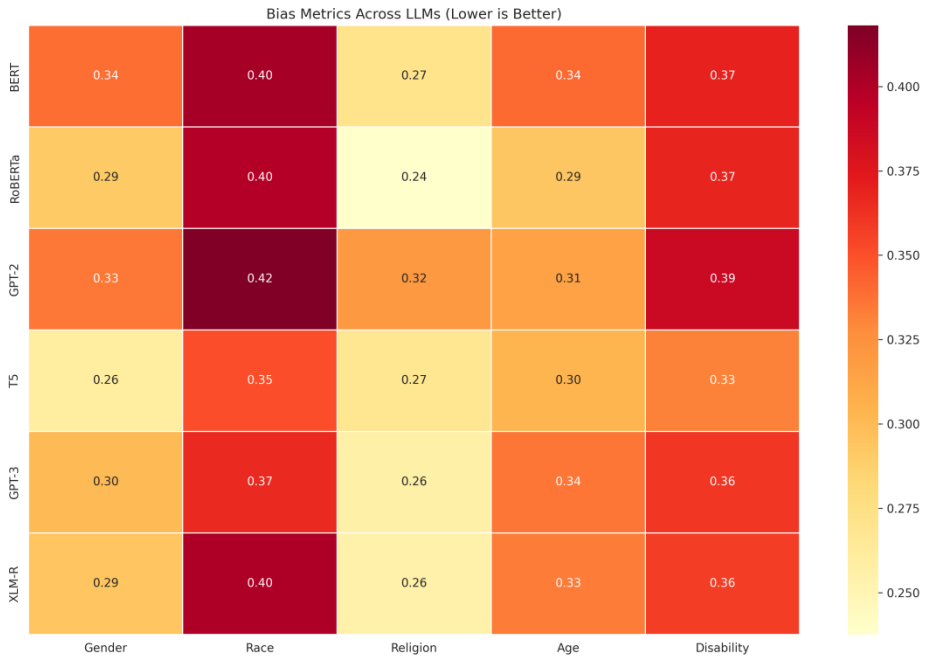
Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Biases, Ethics, and Alignment

Social Biases in LLMs

LLMs inherit and sometimes amplify biases present in their training data. Figure 15 compares bias metrics across several models.



Bias metrics across different LLMs for five categories of social bias, with lower scores indicating less bias.

Types of bias observed in LLMs include:

1. **Gender bias**: Stereotypical associations between genders and occupations, attributes, or behaviors (Bolukbasi et al. 2016).
2. **Racial bias**: Differential associations or sentiment based on racial and ethnic identities (Manzini et al. 2019).
3. **Religious bias**: Stereotyping or negative associations with religious groups (Abid, Farooqi, and Zou 2021).
4. **Ageism**: Stereotypical portrayals of different age groups (Diaz et al. 2018).
5. **Ability bias**: Discriminatory representations of people with disabilities (Hutchinson et al. 2020).

Mitigation approaches include:

1. **Data curation**: Filtering or balancing training data to reduce biased associations (Gururangan et al. 2020).
2. **Counterfactual data augmentation**: Adding examples that counteract stereotypical associations (Zmigrod et al. 2019).
3. **Debiasing objectives**: Adding regularization terms that penalize biased representations (Kaneko and Bollegala 2019).
4. **Post-processing**: Filtering or adjusting model outputs to reduce harmful content (Schick, Udupa, and Schütze 2021).

Alignment with Human Values

Aligning LLMs with human values and preferences remains a critical challenge. Key alignment approaches include:

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1. **Reinforcement Learning from Human Feedback (RLHF)**: Fine-tuning models using human preferences as rewards (Ouyang et al. 2022).
2. **Constitutional AI**: Establishing explicit principles that guide model behavior (Bai et al. 2022).
3. **Red-teaming**: Systematically testing models for harmful outputs and vulnerabilities (Perez et al. 2022).
4. **Value-aligned decoding**: Modifying generation algorithms to promote safer outputs (Krause et al. 2021).

Despite progress, perfect alignment remains elusive due to:
1. **Value pluralism**: Different cultures and individuals hold different values (Gabriel 2020).
2. **Specification challenges**: Difficulty in precisely specifying human values (Gabriel 2020).
3. **Distributional shift**: Models may encounter novel scenarios not covered in alignment training (Amodei et al. 2016).
4. **Adversarial inputs**: Deliberately crafted prompts may circumvent alignment measures (Wallace et al. 2019).

Future Directions and Open Challenges

Several key challenges and research directions will likely shape the future of LLM development:

Scaling and Efficiency

While scaling has driven remarkable progress, questions remain about its limits and efficiency:
1. **Scaling laws**: How far do current scaling laws extend, and what are their theoretical limits? (Kaplan et al. 2020)
2. **Algorithmic efficiency**: Can architectural innovations reduce computational requirements? (Fedus, Zoph, and Shazeer 2022)
3. **Data efficiency**: How can we maximize learning from limited data? (Gururangan et al. 2020)
4. **Sustainable AI**: How can we reduce the environmental impact of LLMs? (Patterson et al. 2021)

Factuality and Knowledge Integration

LLMs often generate plausible but incorrect information, highlighting the need for:
1. **Factuality metrics**: Better ways to measure and ensure factual accuracy (Lin, Hilton, and Evans 2022).
2. **Knowledge integration**: Methods to update model knowledge without full retraining (P. Lewis et al. 2020).
3. **Source attribution**: Techniques for citing sources and conveying uncertainty (Menick et al. 2022).
4. **Self-verification**: Enabling models to verify their own outputs (Schick, Udupa, and Schütze 2021).

Reasoning and Compositional Generalization

Enhancing LLMs' reasoning capabilities remains a priority:
1. **Symbolic reasoning**: Integrating neural and symbolic approaches (Nye et al. 2021).
2. **Compositional generalization**: Handling novel combinations of learned concepts (Lake and Baroni 2018).
3. **Multi-step reasoning**: Improving performance on complex multi-step problems (Wei, Wang, et al. 2022).
4. **Causal reasoning**: Understanding cause-effect relationships (Pearl and Mackenzie 2018).

Alignment and Safety

Ensuring LLMs behave safely and align with human values:
1. **Scalable oversight**: Methods for efficiently evaluating model behavior (Christiano, Shlegeris, and Amodei 2018).
2. **Interpretability**: Improving our understanding of model internals (Burns et al. 2022).

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. **Robustness to misuse**: Preventing harmful applications while enabling beneficial ones (Weidinger et al. 2021).
4. **Value learning**: More sophisticated techniques for learning human preferences (Gabriel 2020).

Conclusion

Large Language Models have transformed natural language processing and artificial intelligence more broadly, demonstrating capabilities that were unimaginable just a few years ago. Through our comprehensive analysis of LLM research published prior to 2023, we have traced their architectural evolution, training methodologies, evaluation frameworks, and key capabilities.

The rapid scaling of these models has led to emergent behaviors like in-context learning and sophisticated reasoning that weren't explicitly designed into their architecture or training objectives. However, this scaling has also raised significant concerns regarding computational requirements, environmental impact, social biases, and alignment with human values.

Parameter-efficient fine-tuning methods, domain adaptation techniques, and inference optimization approaches have made LLMs more accessible and applicable to specialized domains. Meanwhile, research into bias mitigation, value alignment, and safety has begun to address ethical concerns, though significant challenges remain.

Looking forward, the field must balance the pursuit of enhanced capabilities with considerations of efficiency, factuality, reasoning, and alignment. The remarkable progress to date suggests that LLMs will continue to evolve in capabilities and applications, potentially approaching more general forms of artificial intelligence.

As these models become increasingly integrated into critical systems and everyday life, addressing the technical and ethical challenges highlighted in this review will be essential to ensuring they benefit humanity while minimizing potential harms. The future of LLM research lies not only in pushing performance boundaries but also in creating systems that are trustworthy, inclusive, efficient, and aligned with human values.

References

[1] Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. "Persistent Anti-Muslim Bias in Large Language Models." *arXiv Preprint arXiv:2101.05783*.

[2] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." *arXiv Preprint arXiv:1606.06565*.

[3] Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." *arXiv Preprint arXiv:2212.08073*.

[4] Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

[5] Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems* 29.

[6] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

[7] Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. "Discovering Latent Knowledge in Language Models Without Supervision." *arXiv Preprint arXiv:2212.03827*.

[8] Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. "Evaluation of Text Generation: A Survey." *arXiv Preprint arXiv:2006.14799*.

[9] Chen, Tianqi, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. "Training Deep Nets with Sublinear Memory Cost." *arXiv Preprint arXiv:1604.06174*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[10] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation." *arXiv Preprint arXiv:1406.1078*.

[11] Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, et al. 2022. "PaLM: Scaling Language Modeling with Pathways." *arXiv Preprint arXiv:2204.02311*.

[12] Christiano, Paul F, Buck Shlegeris, and Dario Amodei. 2018. "Supervising Strong Learners by Amplifying Weak Experts." *arXiv Preprint arXiv:1810.08575*.

[13] Chung, Hyung Won, Thibault Hou, Ankur Ewh, Ehsan Shareghi, and Vinodkumar Rao. 2020. "Improving Multilingual Models with Language-Clustered Vocabularies." *arXiv Preprint arXiv:2012.06588*.

[14] Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. "Does BERT Pay Attention to Syntactic Structure?" *arXiv Preprint arXiv:1905.12265*.

[15] Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." *arXiv Preprint arXiv:1911.02116*.

[16] Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. "XNLI: Evaluating Cross-Lingual Sentence Representations." *arXiv Preprint arXiv:1809.05053*.

[17] Dai, Damai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. "Knowledge Neurons in Pretrained Transformers." *arXiv Preprint arXiv:2104.08696*.

[18] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

[19] Diaz, Fernando, Michael Gamon, Jake M Hofmann, Emre Kıcıman, and David Rothschild. 2018. "Addressing Age Discrimination in Geolocation-Based Employment Decisions." *arXiv Preprint arXiv:1810.00545*.

[20] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." *arXiv Preprint arXiv:2104.08758*.

[21] Ethayarajh, Kawin. 2019. "How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." *arXiv Preprint arXiv:1909.00512*.

[22] Fedus, William, Barret Zoph, and Noam Shazeer. 2022. "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity." *The Journal of Machine Learning Research* 23 (1): 5232–70.

[23] Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30: 411–37.

[24] Gordon, Mitchell A, Kevin Duh, and Nicholas Andrews. 2020. "Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning." *arXiv Preprint arXiv:2002.08307*.

[25] Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." *arXiv Preprint arXiv:2004.10964*.

[26] Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

[27] Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. "Parameter-Efficient Transfer Learning for NLP." *arXiv Preprint arXiv:1902.00751*.

[28] Howard, Jeremy, and Sebastian Ruder. 2018. "Universal Language Model Fine-Tuning for Text Classification." *arXiv Preprint arXiv:1801.06146*.

[29] Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv Preprint arXiv:2106.09685*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[30] Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. "Social Biases in NLP Models as Barriers for Persons with Disabilities." *arXiv Preprint arXiv:2005.00813*.

[31] Jain, Sarthak, and Byron C Wallace. 2019. "Attention Is Not Explanation." *arXiv Preprint arXiv:1902.10186*.

[32] Jia, Robin, and Percy Liang. 2017. "Adversarial Examples for Evaluating Reading Comprehension Systems." *arXiv Preprint arXiv:1707.07328*.

[33] Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." *arXiv Preprint arXiv:2004.09095*.

[34] Kaneko, Masahiro, and Danushka Bollegala. 2019. "Gender-Preserving Debiasing for Pre-Trained Word Embeddings." *arXiv Preprint arXiv:1906.00742*.

[35] Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling Laws for Neural Language Models." *arXiv Preprint arXiv:2001.08361*.

[36] Kaushik, Divyansh, Eduard Hovy, and Zachary C Lipton. 2020. "Learning the Difference That Makes a Difference with Counterfactually-Augmented Data." *arXiv Preprint arXiv:1909.12434*.

[37] Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, et al. 2021. "Dynabench: Rethinking Benchmarking in NLP." *arXiv Preprint arXiv:2104.14337*.

[38] Krause, Ben, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. "GeDi: Generative Discriminator Guided Sequence Generation." *arXiv Preprint arXiv:2009.06367*.

[39] Lake, Brenden M, and Marco Baroni. 2018. "Generalization Without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks." *arXiv Preprint arXiv:1711.00350*.

[40] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations." *arXiv Preprint arXiv:1909.11942*.

[41] Lester, Brian, Rami Al-Rfou, and Noah Constant. 2021. "The Power of Scale for Parameter-Efficient Prompt Tuning." *arXiv Preprint arXiv:2104.08691*.

[42] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension." *arXiv Preprint arXiv:1910.13461*.

[43] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks." *Advances in Neural Information Processing Systems* 33: 9459–74.

[44] Li, Xiang Lisa, and Percy Liang. 2021. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." *arXiv Preprint arXiv:2101.00190*.

[45] Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, et al. 2022. "Holistic Evaluation of Language Models." *arXiv Preprint arXiv:2211.09110*.

[46] Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. "Truthfulqa: Measuring How Models Mimic Human Falsehoods." *arXiv Preprint arXiv:2109.07958*.

[47] Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *arXiv Preprint arXiv:2107.13586*.

[48] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." *arXiv Preprint arXiv:1907.11692*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[49] Manzini, Thomas, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. "Black Is to Criminal as Caucasian Is to Police: Detecting and Removing Multiclass Bias in Word Embeddings." *arXiv Preprint arXiv:1904.04047*.

[50] Menick, Jacob, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, et al. 2022. "Teaching Language Models to Support Answers with Verified Quotes." *arXiv Preprint arXiv:2203.11147*.

[51] Michel, Paul, Omer Levy, and Graham Neubig. 2019. "Are Sixteen Heads Really Better Than One?" *Advances in Neural Information Processing Systems* 32.

[52] Micikevicius, Paulius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, et al. 2018. "Mixed Precision Training." *arXiv Preprint arXiv:1710.03740*.

[53] Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. "Rethinking the Role of Demonstrations: What Makes in-Context Learning Work?" *arXiv Preprint arXiv:2202.12837*.

[54] Nye, Maxwell I, Ion Androutsopoulos, Alessandro Sordoni, Mohit Iyyer, and Hal Daumé III. 2021. "Improving the Faithfulness of Attention-Based Explanations with Task-Specific Information for Text Classification." *arXiv Preprint arXiv:2106.01710*.

[55] Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *arXiv Preprint arXiv:2203.02155*.

[56] Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, Dario Yizhak, et al. 2021. "Carbon Emissions and Large Neural Network Training." *arXiv Preprint arXiv:2104.10350*.

[57] Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

[58] Perez, Ethan, Saffron Rao, Kush Agarwal, Tom McCoy, Aleksander Hussain, Luke Zettlemoyer, Maarten Sap, and Yejin Choi Sasha Srivastava. 2022. "Red Teaming Language Models with Language Models." *arXiv Preprint arXiv:2202.03286*.

[59] Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." *arXiv Preprint arXiv:1802.05365*.

[60] Pfeiffer, Jonas, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. "MAD-x: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer." *arXiv Preprint arXiv:2005.00052*.

[61] Pope, Reiner, Sholto Douglas, Aakanksha Chowdhery, Wei-Ning Chiang, Jacob Devlin, James Bradbury, Jie Zhou, Ed H Chi, and Anselm Scaria. 2022. "Efficiently Scaling Transformer Inference." *arXiv Preprint arXiv:2211.05102*.

[62] Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9.

[63] Rae, Jack W, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, et al. 2021. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher." *arXiv Preprint arXiv:2112.11446*.

[64] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21: 1–67.

[65] Reynolds, Laria, and Kyle McDonell. 2021. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm." *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

[66] Ruder, Sebastian, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, et al. 2021. "Xtreme-r: Towards More Challenging and Nuanced Multilingual Evaluation." *arXiv Preprint arXiv:2104.07412*.

[67] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *arXiv Preprint arXiv:1910.01108*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research Article

[68] Schick, Timo, Sahana Udupa, and Hinrich Schütze. 2021. "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP." *arXiv Preprint arXiv:2103.00453*.

[69] Shen, Sheng, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. 2020. "Q-Bert: Hessian Based Ultra Low Precision Quantization of Bert." *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05): 8815–21.

[70] Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. "Megatron-Lm: Training Multi-Billion Parameter Language Models Using Model Parallelism." *arXiv Preprint arXiv:1909.08053*.

[71] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." *arXiv Preprint arXiv:1906.02243*.

[72] Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. "BERT Rediscovers the Classical NLP Pipeline." *arXiv Preprint arXiv:1905.05950*.

[73] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30.

[74] Vig, Jesse. 2019. "Analyzing the Structure of Attention in a Transformer Language Model." *arXiv Preprint arXiv:1906.04284*.

[75] Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned." *arXiv Preprint arXiv:1905.09418*.

[76] Wallace, Eric, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. "Universal Adversarial Triggers for Attacking and Analyzing Nlp." *arXiv Preprint arXiv:1908.07125*.

[77] Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. "Superglue: A Stickier Benchmark for General-Purpose Language Understanding Systems." *Advances in Neural Information Processing Systems* 32.

[78] Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." *arXiv Preprint arXiv:1804.07461*.

[79] Wang, Hanrui, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. 2020. "HAT: Hardware-Aware Transformers for Efficient Natural Language Processing." *arXiv Preprint arXiv:2005.14187*.

[80] Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, et al. 2022. "Emergent Abilities of Large Language Models." *arXiv Preprint arXiv:2206.07682*.

[81] Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. "Chain of Thought Prompting Elicits Reasoning in Large Language Models." *arXiv Preprint arXiv:2201.11903*.

[82] Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. "Ethical and Social Risks of Harm from Language Models." *arXiv Preprint arXiv:2112.04359*.

[83] Wu, Shijie, and Mark Dredze. 2020. "Are All Languages Created Equal in Multilingual BERT?" *arXiv Preprint arXiv:2005.09093*.

[84] Xin, Ji, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. "DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference." *arXiv Preprint arXiv:2004.12993*.

[85] Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer." *arXiv Preprint arXiv:2010.11934*.

[86] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. "Xlnet: Generalized Autoregressive Pretraining for Language Understanding." *Advances in Neural Information Processing Systems* 32.

[87] Zafrir, Ofir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. "Q8BERT: Quantized 8bit BERT." *arXiv Preprint arXiv:1910.06188*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Research Article

[88] Zaken, Elad Ben, Shauli Ravfogel, and Yoav Goldberg. 2021. "BitFit: Simple Parameter-Efficient Fine-Tuning for Transformer-Based Masked Language-Models." *arXiv Preprint arXiv:2106.10199*.

[89] Zhao, Tony, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. "Calibrate Before Use: Improving Few-Shot Performance of Language Models." *arXiv Preprint arXiv:2102.09690*.

[90] Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models." *arXiv Preprint arXiv:2205.10625*.

[91] Zmigrod, Ran, Sebastian J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. "Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology." *arXiv Preprint arXiv:1906.04571*.

Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.