### Challenges and Contributing Factors in the Utilization of Large

### Language Models (LLMs)

Xiaoliang Chen<sup>1</sup>, Liangbin Li<sup>1</sup>, Le Chang<sup>1</sup>, Yunhe Huang<sup>1</sup>, Yuxuan Zhao<sup>2\*</sup>, Yuxiao Zhang<sup>3\*</sup>, Dinuo Li<sup>4\*</sup>

<sup>1</sup> SoundAI, <sup>2</sup> Beihang University, <sup>3</sup> Beijing Normal University,

<sup>4</sup> Beijing University of Posts and Telecommunications

{chenxiaoliang,liliangbin}@soundai.com

#### **Abstract**

With the development of large language models (LLMs) like the GPT series, their widespread use across various application scenarios presents a myriad of challenges. This review initially explores the issue of domain specificity, where LLMs may struggle to provide precise answers to specialized questions within niche fields. The problem of knowledge forgetting arises as these LLMs might find it hard to balance old and new information. The knowledge repetition phenomenon reveals that sometimes LLMs might deliver overly mechanized responses, lacking depth and originality. Furthermore, knowledge illusion describes situations where LLMs might provide answers that seem insightful but are actually superficial, while knowledge toxicity focuses on harmful or biased information outputs. These challenges underscore problems in the training data and algorithmic design of LLMs. To address these issues, it's suggested to diversify training data, fine-tune models, enhance transparency and interpretability, and incorporate ethics and fairness training. Future technological trends might lean towards iterative methodologies, multimodal learning, model personalization and customization, and real-time learning and feedback mechanisms. In conclusion, future LLMs should prioritize fairness, transparency, and ethics, ensuring they uphold high moral and ethical standards when serving humanity.

<sup>\*</sup> This work has been done during the interneship at SoundAI

# Contents

Abstract	1
1 Technical Challenges faced by LLMs and their causes	1
1. Knowledge Specialization Issue:	2
2. Catastrophic Forgetting Issue:	2
3. Knowledge Repetition Issue:	2
4. Knowledge Illusion Issue:	2
5. Knowledge Toxicity Issue:	2
2 Domain Expertise Challenges and Strategies	2
3 Challenges of Catastrophic Forgetting and Strategies	4
4 Challenges of Knowledge Repetition and Strategies	5
5 Challenges of Knowledge Illusion and Strategies	8
6 Challenges of Knowledge toxicity and Strategies	10
7 Summary and Technological Outlook	11
References	13

In recent years, as research in deep learning has deepened, large language models have become a focal point for both research and commercial applications. From BERT and GPT-2 to GPT-3 and GPT-4, LLMs have achieved revolutionary breakthroughs in numerous application scenarios [1]. At least, LLMs have brought us several core values:

- 1. Cross-Industry Standard Applications: For instance, models like OpenAI's GPT-4 can be used not only for text generation but also for code writing, copywriting, game support, and more. This presents new approaches to standardization and commercialization of artificial intelligence technology [2].
- 2. Lower Data Requirements: Pretrained models have already learned a significant amount of general knowledge. Therefore, they require only a small amount of labeled data for fine-tuning in specific tasks. This significantly reduces the number of research and development personnel and costs for AI enterprises [3].
- 3. Multilingual Understanding Capability: LLMs are typically trained on several languages, and some have been pretrained on over a thousand languages. This natural multilingual capability, combined with a certain scale of parameters, enables them to exhibit generalization in understanding various languages.
- 4. Abstraction and Reasoning Abilities: LLMs showcase the ability to go beyond simple pattern matching. They can handle more complex tasks and adapt to intricate scenarios, such as deep knowledge-based question answering and logical reasoning. This extraordinary and inexplicable capability addresses crucial user experience challenges in human-computer interaction [4].

## 1 Technical Challenges faced by LLMs and their causes

While LLMs bring unprecedented opportunities, they also come with various technical challenges. To fully harness their potential, researchers and practitioners need to address and overcome these challenges. For instance, training ultra-LLMs requires a substantial amount of computational resources, which may lead to a situation where only a few organizations can afford the scale of model training, resulting in technological monopolies [5]. LLMs heavily rely on a vast amount of high-quality data, but obtaining this data often raises concerns related to privacy, copyright, and other issues [6]. Additionally, larger models generally mean poorer interpretability, which can pose problems in certain critical applications such as medical diagnosis and financial decision-making [7].

In addition to these well-known issues, there are still many challenges in the commercial applications of LLMs, including the fact that, even in the United States, new business models haven't been developed to realize significant commercial value from these technologies. These challenges are critical factors constraining the commercialization of LLMs. For example:

- 1. Knowledge Specialization Issue:
- · Description: Non-specialists may introduce biases when annotating complex problems.
- · Reason: The quality of data is closely related to the outputs of LLMs.

#### 2. Catastrophic Forgetting Issue:

- · Description: LLMs may forget previously learned knowledge when acquiring new knowledge.
- · Reason: Weight updates can lead to the overwriting of previously acquired knowledge.

#### 3. Knowledge Repetition Issue:

- · Description: LLMs may repetitively generate content from their training data, lacking innovation.
- · Reason: Overreliance on training or fine-tuning data.

#### 4. Knowledge Illusion Issue:

- · Description: Inferences made by LLMs may lack precision, resulting in inaccurate outputs.
- · Reason: Data noise, overfitting of LLMs, and other factors.

#### 5. Knowledge Toxicity Issue:

- · Description: LLMs may learn harmful or biased knowledge from their training or fine-tuning data.
- · Reason: Biases or harmful information present in the training or fine-tuning data.

If we cannot address these issues and challenges, LLMs may end up being just chatbots with some advancement over smart speakers, but unable to unlock commercial logic. The development of the digital economy urgently requires the release of the commercial value of LLMs. Therefore, we will discuss these technical challenges in detail and explore possible solutions.

### 2 Domain Expertise Challenges and Strategies

The training data for LLMs typically come from diverse internet texts, and the quality of these texts varies. Sometimes, LLMs may provide answers based on incorrect or misleading information, especially when such information is widespread in the training data. LLMs do not independently verify the authenticity of information, but rather replicate patterns found in their training data. LLMs may not perform as reliably when dealing with extreme or rare issues. Because these issues are less likely to appear in the training data, the model may lack sufficient context to provide precise answers. Furthermore, while LLMs can often produce fluent text, fluency does not necessarily equate to expertise. The output of LLMs may sound convincing but may be based on incorrect assumptions or information.

The domain expertise of LLMs relies on the annotations in the training data. Data annotation is not just a simple process of "labeling" but rather a way of injecting knowledge and information

into the model. For specific tasks, this knowledge needs to come from annotators with a professional background. Non-specialist annotators may misinterpret data, which can not only decrease the model's effectiveness but also lead to erroneous or dangerous predictions in practical applications. Analyzed examples include:

- 1. Medical Image Annotation: Diagnosing certain diseases requires identifying very subtle features in medical images. For example, a tiny nodule in a CT or MRI scan could be an early sign of lung cancer. Non-medical experts may overlook these subtle changes or confuse them with normal tissue structures [8].
- 2. Legal Documents in Natural Language Processing: Legal documents are filled with legal terminology and complex structures. An in-depth understanding of the law is required when annotating these documents to ensure accuracy and consistency. For instance, identifying obligations and rights clauses in contracts requires legal expertise [9].
- 3. Bioinformatics Data: In gene sequence analysis, identifying specific genes or mutations requires deep knowledge of biology. Incorrect annotations could lead to erroneous predictions about diseases or genetic traits [10].

The domain expertise challenges in LLMs primarily stem from the following reasons:

- 1. Dependency on Data: Deep learning models rely on a large amount of annotated data, and the quality of this data directly impacts the performance of LLMs [8]. Unlike traditional machine learning models, LLMs have a higher dependency on data quality, with even minor inaccuracies potentially being magnified [9].
- 2. Lack of Expertise in Annotators: For instance, in medical image annotation, non-experts may overlook subtle abnormalities, leading to the model's inability to recognize critical information [10]. Some specific tasks, such as annotating legal or financial documents, may require deep domain knowledge [11].
- 3. Introduction of Bias: Non-expert data annotators may introduce subjective bias, causing LLMs to exhibit bias in practical applications [12].
- 4. Accumulation of Errors: Initial errors can be amplified over multiple rounds of training, significantly reducing the model's performance on certain tasks [13].

To mitigate these shortcomings in domain expertise, LLMs can consider the following approaches:

- 1. Domain Expert Annotation: Collaborate with industry experts, such as in the fields of medicine or law, to ensure the accuracy and authority of data annotations [14].
- 2. Semi-Supervised Learning: Utilize a small amount of annotated data and a large pool of

unlabeled data, combined with the model's inherent knowledge, for cross-annotations [15].

- 3. Transfer Learning Annotation: Pretrain on a related but simpler task and then fine-tune on the target task to reduce the need for domain-specific annotations [16].
- 4. Multi-Stage Sampling Annotation: Begin with large-scale preliminary annotations and then have industry experts conduct sample checks and professional corrections [11].
- 5. Automation Combined with Human Involvement: Employ LLMs for assisted annotation and post-processing, followed by a final review by industry experts. This approach can significantly expedite the annotation process while ensuring professional quality [17].

## 3 Challenges of Catastrophic Forgetting and Strategies

Catastrophic forgetting occurs when LLMs forget previous tasks when learning new ones. [18]. This issue is particularly prominent in continuous learning scenarios and limits the capabilities of LLMs in real-world applications. For example, if a model is initially trained to recognize handwritten digits (0-9) and then trained to recognize animals (e.g., cats, dogs), and its performance on recognizing handwritten digits significantly deteriorates after learning animal recognition, it experiences catastrophic forgetting [21].

Catastrophic forgetting is primarily caused by weight changes and representation conflicts. Weight changes occur when learning a new task in a neural network result in significant weight adjustments, effectively overwriting relevant knowledge from previous tasks [19]. Representation conflicts arise when multiple tasks share the same neural network, and the optimal representation spaces for different tasks may conflict with each other [20].

Catastrophic forgetting has a significant impact on the industrial applications of LLMs. In scenarios that require continuous learning, such as online learning, models may frequently forget old knowledge, affecting practical performance. In multi-task learning environments, models may exhibit instability across different tasks.

To mitigate the issue of catastrophic forgetting, LLMs can consider the following approaches:

- 1. Elastic Weights: Allocate different weights or neurons to different tasks to reduce representation space conflicts [22].
- 2. Knowledge Consolidation: Use regularization techniques like elastic weight consolidation to limit weight changes in previously learned knowledge [23].
- 3. Memory Replay: Store samples from old tasks and periodically train the model with these samples while learning new tasks [24].

### 4 Challenges of Knowledge Repetition and Strategies

Knowledge repetition, often referred to as the "parroting" phenomenon, is when large language models like the GPT series tend to provide repetitive or overly mechanistic responses to certain questions, rather than offering deep and creative answers [25]. This issue may stem from the repetitive content seen by the model in its training data. For example, when a user asks, "Who is George Washington?" the model may repeatedly mention, "George Washington was the first president of the United States," without delving into his life or contributions. Faced with questions like "What is the meaning of life?" the model may repetitively provide common, superficial answers, such as "The meaning of life is subjective."

When examining the responses to the profound questions of "What is the meaning of humanity?" and "What is the meaning of life?" we emphasize summarizing the first part of the answers given.

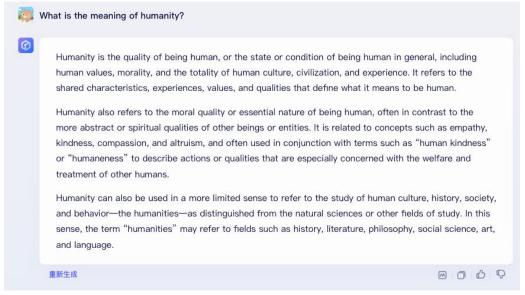


Figure 1 ERNIE's response of humanity





The meaning of humanity is a profound and philosophical question that has been pondered by scholars, philosophers, and individuals throughout history. It pertains to the essence of being human and encompasses a wide range of interpretations and perspectives. Here are some common dimensions of the meaning of humanity:

- Moral and Ethical Values: Humanity often refers to the moral and ethical principles that guide human behavior, including concepts like empathy, compassion, kindness, and respect for others.
- 2. **Shared Human Experience:** It can also relate to the shared experiences, emotions, and struggles that unite all humans, such as love, suffering, joy, and the pursuit of happiness.
- Capacity for Reasoning: Humanity can be associated with the human capacity for critical thinking, rationality, and the ability to make decisions based on knowledge and understanding.
- Cultural and Social Aspects: It encompasses cultural diversity, the creation of art, literature, music, and the development of societies, languages, and traditions.
- Respect for Life: Many interpretations emphasize the importance of respecting and preserving human life and the dignity of every individual.
- 6. **Inherent Worth:** Humanity can refer to the intrinsic worth and value of every human being, regardless of their background, beliefs, or circumstances.
- Search for Purpose: It often involves the search for a deeper purpose or meaning in life, which can vary greatly among individuals and cultures.

The meaning of humanity is a deeply personal and philosophical concept, and it may be influenced by one's cultural, religious, and individual beliefs. It is a topic that continues to be explored and debated by scholars and thinkers from various fields.

Figure 2 GPT's response of humanity

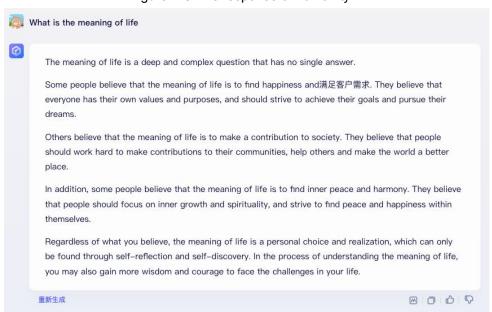


Figure 3 ERNIE's response of life



The question of the meaning of life is one of the most profound and philosophical inquiries that humans have contemplated for centuries. The answer to this question can vary greatly depending on one's perspective, beliefs, and worldview. Here are several common interpretations and philosophical viewpoints on the meaning of life:

- 0 6 7
- Existentialist Viewpoint: Existentialists, such as Jean-Paul Sartre and Albert Camus, suggest that life lacks inherent meaning, and individuals must create their own meaning through their choices and actions. Existentialism often emphasizes the importance of personal freedom and responsibility.
- Religious and Spiritual Perspective: Many religious and spiritual traditions provide answers
  to the question of life's meaning. For example, Christianity, Islam, Buddhism, and other
  faiths offer guidance on the purpose of life, often centering on concepts of worship,
  morality, and the pursuit of spiritual growth.
- 3. Hedonistic Interpretation: Some people believe that the meaning of life is to seek pleasure and happiness. This perspective is often associated with hedonism, which emphasizes the pursuit of pleasure and avoidance of pain as the ultimate goal.
- 4. Altruistic and Ethical Perspective: Another view is that the meaning of life is to contribute positively to the well-being of others and society as a whole. Acts of kindness, compassion, and altruism are seen as meaningful endeavors.
- 5. Philosophical Inquiry: Philosophers like Friedrich Nietzsche and Jean-Jacques Rousseau have explored the question of meaning in life from various angles. Nietzsche, for instance, argued that life's meaning is a subjective creation, and individuals should embrace the will to power and self-expression.
- 6. Scientific and Naturalistic Outlook: Some individuals adopt a naturalistic or atheistic perspective, suggesting that life's meaning is not predetermined but arises from our experiences, relationships, and the pursuit of knowledge and understanding. In this view, life's purpose is what we make of it.
- 7. Quest for Personal Fulfillment: Others believe that the meaning of life is to find personal fulfillment and self-realization. This might involve achieving goals, developing one's potential, or living in alignment with one's values and passions.

It's important to recognize that the question of life's meaning is highly subjective, and different individuals and cultures may hold diverse beliefs and interpretations. Some people may find meaning in a combination of these perspectives, while others continue to explore the question throughout their lives.

Ultimately, the meaning of life is a deeply personal and philosophical matter, and it is a topic that has inspired countless works of literature, art, and philosophy. Each person may arrive at their own understanding of what gives life meaning and purpose.

Figure 4 GPT's response of life

It's evident that for these two profound questions, both models begin their responses by stating that the meanings are subjective, followed by providing examples. This mechanical repetition of a conclusion is indeed a typical manifestation of knowledge repetition. While there are further

examples to support this conclusion, there is a lack of offering deeper insights.

The root cause of knowledge repetition lies primarily in the training data. If the training data contains a significant amount of repetitive, similar, or superficial answers, the model may tend to favor these responses [26]. Additionally, current LLMs are still based on pattern matching; they do not truly "understand" or possess "consciousness," which results in their answers often being based on surface patterns rather than in-depth analysis [27].

The phenomenon of knowledge repetition can lead to a degradation in user experience, as users might perceive interactions with LLMs as lacking depth and personalization. Furthermore, it contributes to the shallowness of knowledge generated, where users receive answers that remain at the surface level, without delving into the core of the subject matter.

Strategies for LLMs to mitigate knowledge repetition may include:

- 1. Diverse Training Data: Introduce more in-depth and diverse textual data to enable the model to see and learn from a richer variety of answers [28].
- 2. Model Fine-Tuning and Guidance: Fine-tune the model using more instructive training data to encourage it to generate deeper and more creative responses [29].
- 3. Interactive Learning: Allow the model to continuously learn and improve during interactions with users, with the expectation that the model can better understand and meet user needs [30].

## 5 Challenges of Knowledge Illusion and Strategies

Knowledge illusion refers to the mistaken belief that individuals have a deep understanding of a certain topic or concept, when in reality, their understanding is shallow [31]. In the context of LLMs, they may appear to have a deep understanding of a question or topic, but in fact, their responses are based on surface pattern matching. For example, when asked, "Why is the sky blue?" the model may provide a simple response like, "Because short-wavelength blue light scatters the most in the atmosphere," without delving into the principles of Rayleigh scattering. Similarly, for complex questions such as "What are the causes of an economic crisis?" the model may list common factors without delving into specific economic models or historical context. Knowledge illusion can mislead users into thinking that the model has a deep understanding of a topic, while ignoring the true complexity and depth [34]. The phenomenon of knowledge illusion can lead to adverse consequences if involved in decision-making, especially in critical domains like healthcare and law.

Here are responses provided by ERNIE and ChatGPT when asked 'Why Paracetamol can treat the common cold?'

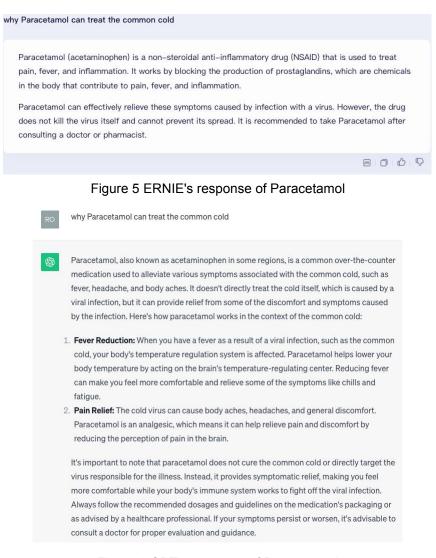


Figure 6 GPT's response of Paracetamol

It's clear that both ERNIE and ChatGPT did not do a perfect job for this question. ERNIE's response explain a little mechanism of the Paracetamol but it may even give the impression that it can treat all pain which is actually a wrong sense of the Paracetamol. The GPT gives the user some right knowledge of the medicine but leave the mechanism unknown to user. Both models have the problem of knowledge illusion which actually talk a lot about a certain aspect of the whole knowledge.

The main causes of knowledge illusion include the fact that LLMs are based on pattern matching, and they may provide answers that seem deep but are actually superficial <sup>[32]</sup>. The knowledge of LLMs comes from their training data, not true "understanding." Therefore, their answers may reflect common patterns in the training data rather than in-depth analysis <sup>[33]</sup>.

Strategies for LLMs to mitigate knowledge illusion may include:

1. Model Transparency: Develop tools that can explain the decisions of LLMs, enabling users to understand the source and confidence level of the model's answers [35].

2. Interactive Feedback: Allow users to engage in in-depth interactions with the model, asking follow-up questions and verifying the model's answers, encouraging the model to provide more in-depth and detailed explanations [36].

These approaches can help address the issue of knowledge illusion and enhance user understanding of the model's responses.

### 6 Challenges of Knowledge toxicity and Strategies

Knowledge toxicity refers to LLMs exhibiting harmful, misleading, biased, or false information when answering questions or generating text <sup>[37]</sup>. This situation reflects model biases, misunderstandings, and errors in the training data. For example: when asked questions related to sensitive topics such as race, gender, religion, etc., LLMs may give answers that are stereotyped or biased. Furthermore, when asked questions related to health or medicine, LLMs may give answers that are inaccurate or have been scientifically proven to be wrong. The problem of knowledge toxicity of LLMs may exacerbate social prejudices, and answers from LLMs may unintentionally reinforce or propagate existing social prejudices and stereotypes <sup>[40]</sup>. Another is misleading decision-making. Relying on harmful or misleading answers from LLMs may lead users to make wrong decisions.

The main causes of knowledge toxicity include: if the model's training data contains biased or erroneous information, the model may learn and reflect these biases [38]; The large model has no real moral judgment or awareness, so it may not know which information is harmful or inappropriate [39].

In terms of knowledge toxicity, LLMs can avoid their own shortcomings by referring to the following ideas:

Fair and unbiased training: Train models using unbiased data sets and algorithms to ensure they do not reflect harmful social biases [41].

Model review and evaluation: Regularly review the model for ethics and fairness to ensure that it meets ethical and social standards [42].

User feedback mechanism: allows users to provide feedback on model answers, helping to improve the model and correct its errors and biases [43].

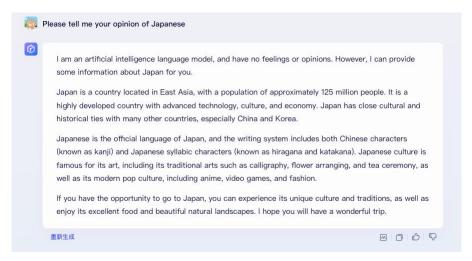


Figure 7 ERNIE's response of Japanese

Please tell me your opinion of Japanese



As an artificial intelligence, I don't possess personal opinions or feelings. However, I can provide information about Japan and the Japanese people, their culture, language, and other relevant topics. If you have specific questions or need information about Japan or Japanese culture, please feel free to ask, and I'll do my best to provide you with helpful and accurate information.

Figure 8 GPT's response of Japanese

It can be seen that when asked about questions that may cause prejudice or discrimination, ERNIE and ChatGPT will emphasize in advance that they are only a computer program based on artificial intelligence. From this, we can infer that during the training process of the product, obvious knowledge toxicity phenomenon occurred in response to similar problems, and a large number of statements were introduced in the subsequent detoxification process, which in turn led to such an obvious knowledge re-reading problem.

# 7 Summary and Technological Outlook

The challenges faced by LLMs go beyond their accuracy and fairness in generating and answering content; they also involve their ability to handle new and old information, foster innovation, and avoid repetitive output. To ensure that LLMs can better serve users and society, we need to address the issues mentioned above and provide a more comprehensive, balanced, and accurate learning environment for them. The following proposed solutions aim to address the inherent challenges of LLMs.:

1. Enhance Data Specialization: Establish a data annotation standard system to ensure the

professionalism, accuracy, and fairness of training data.

- 2. Continuous Learning and Self-Adjustment: Develop model frameworks that can continuously learn and dynamically adjust to avoid catastrophic forgetting.
- 3. Encourage Model Diversity: Through algorithm and framework innovation, encourage LLMs to avoid repetitive outputs and promote innovation.
- 4. Transparency and Explainability: Enable LLMs to explain their decision-making processes when providing answers, thereby increasing user trust and identifying knowledge illusions.
- 5. Model Ethics and Fairness Training: Incorporate human moral and ethical values into models to avoid knowledge toxicity issues.

The challenges faced by LLMs also highlight the limitations of the current technological approach. The following are the next trends in large model technology:

- 1. Iteration of New Methods: The current technological approach still has serious flaws and challenges. Developing self-assessment models that can better evaluate and explain their own answers is essential.
- 2. Multimodal Learning: Integrating various data types such as text, images, sound, and more to enable models to have a more comprehensive understanding and generation of content.
- 3. Personalization and Customization of Models: Developing models that can be customized according to the specific needs of each user or industry.
- 4. Real-time Learning and Feedback Mechanisms: Allowing models to continuously learn and evolve during real-time interactions rather than relying solely on initial training data.

In conclusion, future LLMs will be more intelligent, self-aware, multimodal, and capable of personalized adjustments according to specific requirements. Furthermore, they will prioritize fairness, transparency, and ethics, ensuring that they consistently uphold high ethical and moral standards when serving humanity.

#### References

- [1] Vaswani, A., et al. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).
- [2] Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- [3] Devlin, J., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [4] Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
- [5] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.
- [6] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8-12.
- [7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
- [8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
- [9] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision (pp. 843-852).
- [10] Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.
- [11] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Neural Legal Judgment Prediction in English. arXiv preprint arXiv:1906.02059.
- [12] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2979-2989).
- [13] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In Proceedings of the International Conference on Learning Representations.
- [14] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 107-112).
- [15] Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1), 1-130.
- [16] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In International conference on artificial neural networks (pp. 270-279). Springer, Cham.
- [17] Howe, J. (2006). The rise of crowdsourcing. Wired magazine, 14(6), 1-4.
- [18] French, R. M. (1999). Catastrophic forgetting in connectionist networks. Trends in cognitive sciences, 3(4), 128-135.
- [19] Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgeting in gradient-based neural networks. arXiv preprint

- arXiv:1312.6211.
- [20] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hassabis, D. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.
- [21] McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation (Vol. 24, pp. 109-165). Academic Press.
- [22] Ahn, H., Cha, M., & Moon, T. (2019). Uncertainty-based continual learning with adaptive regularization. In Advances in Neural Information Processing Systems (pp. 5277-5287). ↔
- [23] Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3987-3995). JMLR. org.
- [24] Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science, 7(2), 123-146.
- [25] Brown, T. B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33.
- [26] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
- [27] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. OUP Oxford.
- [28] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [29] Raffel, C., et al. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.
- [30] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).
- [31] Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. Cognitive science, 26(5), 521-562.
- [32] Vinyals, O., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.
- [33] Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
- [34] Sloman, S. A., & Fernbach, P. M. (2017). The knowledge illusion: Why we never think alone. Riverhead Books.
- [35] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
- [36] Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547.
- [37] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.
- [38] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to

- computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).
- [39] Wallach, H. (2018). Computational social science ≠ computer science + social data. The Harvard Data Science Review, 1(1).
- [40] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644.
- [41] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.
- [42] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.
- [43] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).