Reusing, Recycling and Reducing Large Models for Developing Green and Responsible Language Technology

Ainhoa Vivel-Couso

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain

Abstract

Natural Language (NL) is the most common and efficient tool for humans to transmit information. Natural Language Processing (NLP), which includes NL Understanding (NLU) and NL Generation (NLG), is one of the main challenges in Artificial Intelligence (AI) and has a growing economic impact on the current digital transformation. Despite their impressive capabilities, large pre-trained language models present serious drawbacks from a research, environmental, and ethical advancement perspective. The primary research objective of this doctoral thesis is to advance the state of the art in NL technology by (i) developing efficient methods to extend existing models to new domains, genres, and languages for the official languages of Spain (Spanish, Catalan, Basque, and Galician) and English; (ii) exploring new ways to pre-train and fine-tune language models efficiently in terms of parameters, thereby reducing the carbon footprint required to train such models; (iii) addressing the explainability of large pre-trained language models for NLG tasks; (iv) developing a series of advanced domain-based content applications across multiple languages, sectors and domains (e.g., Meteorology and Health) with an emphasis on explainability and evaluation tasks; and (v) defining (and overseeing compliance with) guidelines and requirements for the development of Responsible NLP with an Ethical, Legal, Social, Economical, and Cultural (ELSEC) perspective.

Keywords

natural language generation, multilingualism, artificial intelligence, energy efficiency

1. Justification of the Proposed Research

The NLP community has contributed to the emergence of new disruptive techniques and tools that are revolutionizing research on AI. Thus, the NLP community is currently undergoing a paradigm shift with the production and exploitation of large, pre-trained language models based on transformers [1, 2]. Despite their impressive capabilities, large pre-trained language models present drawbacks from an environmental and ethical perspective. For example, computing large pre-trained models from scratch is highly demanding and has a significant carbon footprint [3]. Additionally, these models are black boxes, meaning we do not have a clear understanding of how they function, when they fail, what emergent properties they might present, or new ways to efficiently exploit these models. Fortunately, research on Explainable AI can help better understand and utilize these models [4, 5, 6].

Some authors refer to these models as foundation models to highlight their central yet incomplete nature [7]. Furthermore, these models are costly to train and develop, both financially—due to the cost of hardware, electricity, and cloud computing time—and environmentally, due to the carbon footprint required to power modern servers with multi-GPU hardware. This also means that only a limited number of organizations with ample resources in terms of funding, computing capabilities, NLP experts, and corpora can currently afford to develop and implement such models. A growing concern is that due to unequal access to computing power, only certain companies and elite research groups can afford modern AI research [8].

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

ainhoa.vivel.couso@usc.es (A. Vivel-Couso)
citius.gal/es/team/ainhoa-vivel-couso/ (A. Vivel-Couso)
0000-0002-5860-4849 (A. Vivel-Couso)

2. Origin and Related Work

For an AI to be considered reliable, there are at least four aspects that can be considered: fairness, robustness, explainability and traceability [9]. The concept of Explainable AI (eXplainable AI, XAI) refers to the ease with which a human being can understand the decisions made by the AI and the actions underlying them [10, 11].

In the case of data-driven machine learning methods, depending on the level of transparency of the AI method considered, we speak of white-box, gray-box or black-box methods. An example of a white-box, or fully transparent, method is the so-called decision tree [12], since the tree generated from data makes explicit in a readable form (equivalent to rules) the knowledge at stake (variables, values) when making a prediction. For their part, methods based on fuzzy rule systems are an example of a gray box because they are interpretable to a certain degree [13]. Finally, neural networks are an example of an opaque or black-box method, since it is not possible, especially in complex models such as Deep Learning (DL) architectures, to interpret in a readable or understandable way the relevant elements that the model considers when making a prediction. There are different ways of explaining this type of methods, both in terms of their inner workings and the justification of their output (post-hoc methods) [14, 5, 15].

In the field of application of this dissertation (i.e., Reusing, Recycling and Reducing Pre-trained Models for Developing and Evaluating Green Data-to-text Systems) there are a few publications in the literature where the environmental impact of large languages models (LLMs) is analyzed [16, 17, 18]. However, our proposal in this dissertation provides a new approach to the problem, since it addresses the comparison of knowledge transfer methods to reduce environmental impact. To this end, we will study whether fine-tuning can be replaced by another type of knowledge transfer method with a lower environmental impact by being able to generate narratives that are equally understandable, natural and fluid. We devote the remaining of this section to present the definitions and basic concepts of the methodologies we will use in our research: NLG, transfer learning and automatic evaluation of narratives.

2.1. Natural Language Generation

Natural language generation (NLG) is a process that consists of the automatic construction of narratives. This process is traditionally [19] divided into a series of stages:

1. **Text planning**. The first stage consists of the strategic generation of narratives. Specifically, it studies what to say in writing.
2. **Sentence planning**. The second stage determines how to organize the information to be conveyed.
3. **Linguistic realization**. The last stage is based on the application of syntactic and morphological rules to generate a correct text.

Numerous technologies exist for NLG [20]. We will focus on the use of pre-trained models [21] which have been trained in large datasets for specific tasks. They learn general patterns and features from extensive data, often using unsupervised ML techniques. After pre-training, language models can be further adapted or fine-tuned [22] for specific tasks, offering a head start in performance for tasks like language understanding, image recognition, or other applications in AI. Pre-trained models have become popular in various domains due to their ability to capture and transfer knowledge from diverse datasets, improving efficiency and performance for specific applications [23]. We will use them as a basis because a lot of resources have been used during their training. By reusing everything they have learned, we reduce the environmental impact of training a new model from scratch.

As a preliminary analysis[1], we studied the use of LLMs embedded in the following conversational assistants:

[1] I finished my master's studies in February 2024 with the defense of my Master's Thesis. I am now starting my PhD thesis on this topic.

- **Bing AI**[2]. Tests performed on 13/11/2023 at 11:30. Bing AI uses Copilot with GPT-4.0.
- **ChatGPT**[3]. Tests performed on 13/12/2023 at 11:40. The free version of ChatGPT uses GPT-3.5.
- **Google BARD**[4]. Tests performed on 13/12/2023 at 11:45. The specific language model that Google BARD uses is not publicly available, but it is known to be based on the Transformer architecture developed by Google AI. Currently, BARD has been completely replaced by Gemini[5].

The interaction with these systems is based on prompting. Doing several tests on them, it has been empirically appreciated that ChatGPT is the one that generates better narratives. It fits the data, follows the instructions properly and does not tend to hallucinate. While ChatGPT and similar language models have proven to be valuable and versatile, there are some challenges and limitations associated with their usage.

Conversational assistants can generate incorrect or unreliable information due to their reliance on learned data patterns, making independent verification crucial. They are sensitive to input phrasing, leading to different responses with slight changes in wording, and can reflect biases present in their training data. There is a risk of hallucination, where the model generates credible-sounding but inaccurate content. In addition, handling complex or technical queries can be challenging for these models, as they are designed for general use. Prompt engineering, or the specific framing of queries, significantly impacts the model's responses. Ambiguity in queries may lead the model to make incorrect assumptions rather than seeking clarification. Additionally, updates or changes to the model can alter its response patterns, and accessing the most advanced models often requires subscription plans.

When using this type of conversational assistants, it is crucial to be aware of these limitations. For these reasons, we will look for freely worldwide available models which may produce more controlled narratives. Finally, it is worth noting that we will pay attention only to Text-to-Text (T2T) systems because they are the most appropriate for the use case with meteorological data under consideration. To be more specific, we will use Sequence-to-Sequence (seq2seq) language models [24], a type of neural network architecture designed for tasks that involve mapping input sequences to output sequences. These models are particularly popular in NLP tasks such as machine translation, text summarization, and chatbot development. We will reuse pre-trained T2T systems to generate weather descriptions. These descriptions will be based on the meteorological characteristics of each geographical area throughout the year.

2.2. Transfer Learning

Transfer learning (TL) [25] is a ML technique where a model trained on one task is adapted for a second related task. Instead of training a model from scratch, transfer learning leverages the knowledge gained from solving a different but related task. This approach is particularly useful when we have a limited amount of labeled data for the target task and it is especially important to reduce the energy cost and carbon footprint. Training DL models from scratch can be computationally intensive and resource-consuming. Leveraging pre-trained models and fine-tuning them for specific tasks can be more efficient in terms of computational resources and time.

The process of TL typically involves two steps:

1. **Pre-training**. Train a model on a large dataset and a related task. This model is often referred to as the *pre-trained model* or *base model*. For example, a model might be pre-trained on a massive dataset for text generation.
2. **Knowledge transfer**. Use the pre-trained model as a starting point and use some technique to adjust the model so that the information learned can be reused.

[2] https://www.microsoft.com/en-us/bing

[3] https://chat.openai.com

[4] https://bard.google.com/

[5] https://gemini.google.com/

TL is especially valuable in DL, where models have many layers and parameters. Popular pre-trained models, such as those based on convolutional neural networks (CNNs) for image-related tasks or models like BERT for NLP, have shown great success in TL scenarios [26].

TL speeds up training and allow you to capture domain-specific features. However, what interests us most for this Thesis is to study the computational cost, specifically the energy cost, of the different ways of transferring knowledge. This section will discuss the different alternatives existing in the current state of the art for knowledge transfer.

Fine-tuning (FT) is the process of making small adjustments to achieve the desired output or performance [22]. It is well known and widely used in ML, especially in DL. In the context of DL, it involves the use of weights of a trained neural network to program another DL algorithm from the same domain. Thus, FT consists in taking a pre-trained model and training at least one internal model parameter (i.e., weights). On the other hand, in the context of LLMs, what FT typically transforms is a general-purpose base model (e.g., GPT-3) into a specialized model for a particular use case (e.g., summarization).

While FT is a common and effective approach for adapting pre-trained language models to specific downstream tasks, it requires retraining the entire model, which is usually computationally expensive. Therefore, there are several alternative methods and strategies that we can consider, depending on the use case and data availability. The choice of method depends also on factors such as the complexity of the task, the amount of task-specific data available, and the computational resources at our disposal. Accordingly, we have to try empirically different methods in the search for the most suitable one for the particular NLP task under study.

There are many well-known methodologies for TL such as Prompt Engineering [27], Prompt-free Methods [28], Meta-Learning [29], and Reinforcement Learning from Human Feedback [30]. Below, we go deeper with those methods which are the most pertinent for the scope of this Thesis:

- **Zero-shot Learning** (ZSL) is a problem setup in DL where, at test time, a learner observes samples from classes which were not observed during training [31, 32]. Zero-shot methods generally work by associating observed and non-observed classes through some form of auxiliary information, which encodes observable distinguishing properties of objects. If we want to perform tasks that the model has never seen during pre-training, we can explore ZSL techniques. These methods aim to make predictions without task-specific FT. ZSL relies on models' ability to generalize to new tasks by understanding textual descriptions or examples of those tasks. ZSL is a valuable technique that showcases the generalization and adaptability of pre-trained language models to a wide array of NLP tasks. Nevertheless, even if ZSL offers significant advantages, it may have limitations in cases where the task descriptions are ambiguous or the model's pre-trained knowledge does not align well with the target tasks. In such cases, few-shot learning or FT on a small amount of task-specific data may be necessary to achieve optimal performance.
- **Few-shot Learning** (FSL), also referred to as low-shot learning (LSL) by some researchers, is a ML method ready to exploit a training dataset which contains limited information [33, 34]. Thanks to FSL, we can train models to perform tasks with only a few examples. This approach leverages the ability of models to generalize from limited examples, and it is particularly useful when, due to the lack of data, FT on a large dataset is not feasible. FSL is an alternative approach to FT language models that aims to perform tasks with very limited labeled data, often as few as one or a few examples per class or category. Instead of extensively FT a model, FSL focuses on enabling models to generalize effectively from a small amount of task-specific data.
- **Adapters** (ADT) add new modules between layers of a pre-trained model [35, 36]. This means that parameters are copied over from pre-training (meaning they remain fixed) and only a few additional task-specific parameters are added for each new task, all without affecting previous ones. In standard FT, the new top-layer and the original weights are co-trained. In contrast, in adapter-tuning, the parameters of the original network are frozen and therefore may be shared by many tasks. ADT modules have two main features: a small number of parameters, and a near-identity initialization. This TL technique provides a flexible and efficient way to extend pre-trained language models for a variety of NLP tasks.

2.3. Metrics for Text Evaluation

Text evaluation metrics are used to assess the quality of generated text. There are numerous techniques for evaluating language models [37, 38]. The choice of metrics depends on the specific task or goal, since different metrics capture various aspects of text quality. Text evaluation metrics can be broadly categorized into human evaluation and automatic evaluation metrics:

- **Human evaluation** entails the application of human judgment to assess the quality of generated text. Human evaluators contribute subjective ratings or judgments on various aspects of text quality (e.g., fluency, coherence, relevance, or overall quality). While human evaluation offers rich and subjective insights, it is characterized by being time-consuming, expensive, and susceptible to individual evaluator bias [39].
- **Automatic evaluation** employs computational methods to assess the quality of generated text, offering valuable insights into the strengths and weaknesses of the model's outputs. This analytical approach provides detailed quantitative feedback, facilitating a comprehensive understanding of the context in which the generated text exists. A key distinction from human metrics lies in the fact that automatic metrics can be computed by a machine, streamlining the evaluation process.

Given the large number of model checkpoints slated for evaluation, conducting human assessments was discarded for this Thesis. Consequently, the assessment of texts generated by diverse models employing various TL techniques will be conducted through automated evaluation only.

Although a multitude of automatic evaluation metrics exists, it is important to note that, given the nature of our problem as seq2seq, not all metrics are applicable. For example, Perplexity [40] stands out as one of the most prevalent metrics for appraising language models. Nevertheless, it is pertinent to acknowledge that this metric is tailored specifically to autoregressive language models, often referred to as causal language models. However, perplexity is not commonly used to evaluate seq2seq models, especially those like MT5 (Multilingual Translation Transformer) [41], where the task involves transforming an input sequence into an output sequence.

For evaluating seq2seq models like MT5, metrics such as BLEU (Bilingual Evaluation Understudy) [42] or ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [43] are the most commonly used. These metrics compare the generated sequence against a reference or target sequence and are better suited to capture the overall quality and fluency of the generated text.

3. Description of the Proposed Research

The main objective of this doctoral thesis is to develop and validate the necessary tools to pave the way for Responsible NLP technology. To achieve this primary objective, the following specific objectives will be addressed: (I) develop efficient methods for extending existing models to new domains, genres, and languages for the official languages of Spain and English; (II) explore new ways to pre-train and fine-tune language models efficiently; (III) address the explainability of large pre-trained language models; (IV) develop advanced applications in multiple sectors and domains (e.g., Meteorology and Health) with an emphasis on explainability and evaluation tasks; and (V) define and oversee the compliance with guidelines and requirements for the development of Responsible NLP Technology from an ELSEC perspective.

4. Methodology and Proposed Experiments

To address all the scientific and technical challenges in this doctoral thesis, we will follow the principles of agile software development. This involves engaging users in the design process, seeking continuous improvement, encouraging a quick and flexible response to changes, and supporting the frequent delivery of functional software along with related technical and user manuals. Of course, a thorough review of the state of the art will be conducted at the beginning of each sprint. Thus, requirements analysis tasks are not intended to be completed before subsequent tasks.

The entire process is a dynamic cycle around increasingly complex prototypes, covering the following three steps:

1. **Research and Design Period**. Define the research objective to be achieved in this iteration and investigate related work. The objective will be adjusted according to potential improvements in state-of-the-art methods or to address gaps in the research line. A simple proposal will be written to introduce previous studies and highlight the novelty in this thesis.
2. **Development and Validation Period**. Develop new algorithms and interfaces customized for the target users.
3. **Optimization and Data Collection Period**. Test the new developments with experimental data and optimize them by adjusting the hyperparameters.

The experimental process comprises the following stages:

- **Standby Power Measurement** involves quantifying the energy consumption of the GPUs when no processes are running. This measurement is expressed in watts-hour.
- **Data Preprocessing** is done to generate the datasets used for defining the Baseline and training the models.
- **Baseline Definition**. The training involves establishing a reference model to serve as a baseline. This model will undergo traditional and resource-intensive training.
- **Knowledge Transfer**. The generation process involves creating alternative language models through less resource-intensive training, employing diverse knowledge transfer techniques.
- **Automatic Text Generation**. Automatically generating narratives from test data for each trained model, including the Baseline.
- **Evaluation**. Evaluating texts generated by all models, including the Baseline, using automatic metrics.

In our doctoral thesis, both public and private data (confidential and/or personal) will be handled in use cases exclusively for research purposes. All data will be appropriately managed and anonymized before analysis, and the data protection officer of USC will be contacted in case of any conflict. Additionally, authorization from the USC Research Ethics Committee will be sought before conducting any experiments that may involve ethical issues.

Furthermore, data collection and usage will comply with the General Data Protection Regulation (GDPR) and the new European AI regulation (AI Act). Finally, we will apply the Assessment List for Trustworthy AI, developed by the High-Level Expert Group on AI established by the European Commission (EU HLEG), to evaluate the compliance of the new AI systems developed with the requirements (Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity, Non-discrimination, and Fairness; Societal and Environmental Well-being; Accountability) in their Ethical Guidelines for Trustworthy AI. Additionally, the doctoral thesis supervisor leads the Trustworthy AI laboratory at CiTIUS, which is affiliated with the Z-inspection® initiative. This is a bottom-up holistic inspection process for ethical AI that can be applied to a variety of domains such as business, healthcare, the public sector, and many others. Z-inspection uses the EU HLEG guidelines for Trustworthy AI and is listed in the OECD AI tools and metrics catalog.

5. Specific Research Elements Proposed for Discussion

Our doctoral thesis is focused on advancing NLP technology, particularly in the development of responsible NLP, efficient model training, and application across various domains. The key areas are:

1. **Technological Innovation in NLP**: (I) review the methodologies for extending language models to new domains, genres, and languages; (II) evaluate the efficiency of new pre-training and fine-tuning techniques; and (III) assess the advances in explainability of large pre-trained models.

2. **Responsible AI and Ethical Considerations**: (I) discuss the compliance with GDPR, AI Act, and the principles of FAIR; (II) examine the use of the Assessment List for Trustworthy AI and the Z-inspection® initiative; and (III) consider the environmental impact and the measures taken to reduce the carbon footprint.
3. **Application and Impact**: (I) review the practical applications in sectors like meteorology and healthcare; (II) analyze the socio-economic impact of recycling pre-trained models for various domains; and (III) discuss the open-source distribution of software and models and their potential for broader industry adoption.
4. **Data Handling and Ethics**: (I) review the protocols for handling public and private data, including anonymization and ethical approvals; and (II) examine the strategies for ensuring data privacy and security in compliance with legal standards.

With this, we want to ensure that all critical aspects of our thesis are thoroughly evaluated. The elements proposed for discussion in this Symposium are:

1. What methods for extending language models do you recommend?
2. How to measure the efficiency in pre-training and fine-tuning? Rigth now, I'm using CodeCarbon to take measurements.
3. How can we guarantee the compliance with legal standards (GDPR, AI Act, etc.)? And with ethical guidelines (FAIR, Z-inspection)?
4. Any suggestion for domain-specific applications? I've been working with meteorological data mainly.

Acknowledgments

This research work is supported under Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", but also under Grants PID2020-112623GBI00 and PID2021-123152OB-C21 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future". We also acknowledge the support of the Galician Ministry of Culture, Education, Professional Training and University (Grants ED481A-2024-059, Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and ED431C2022/19 co-funded by the European Regional Development Fund, ERDF/FEDER program).

References

[1] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, J. Zhu, Pre-trained models: Past, present and future, AI Open 2 (2021) 225–250. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000231. doi:https://doi.org/10.1016/j.aiopen.2021.08.002.

[2] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3605943. doi:10.1145/3605943.

[3] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://aclanthology.org/P19-1355. doi:10.18653/v1/P19-1355.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[5] S. Ali et al., Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. Https://doi.org/10.1016/j.inffus.2023.101805.

[6] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. URL: https://www.sciencedirect.com/science/article/pii/S1566253523001148. doi:https://doi.org/10.1016/j.inffus.2023.101805.

[7] R. B. et al., On the opportunities and risks of foundation models, ArXiv preprint (2021). URL: https://crfm.stanford.edu/assets/report.pdf.

[8] N. Ahmed, M. Wahed, The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research, 2020. arXiv:2010.15581.

[9] S. Barro, A. Bugarín, J. Alonso, La confianza en las máquinas inteligentes, Thomson Reuters Aranzadi, 2020.

[10] A. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. Https://doi.org/10.1016/j.inffus.2019.12.012.

[11] D. Gunning et al., DARPA's explainable AI (XAI) program: A retrospective, Applied AI Letters 2 (2021) e61. Https://doi.org/10.1002/ail2.61.

[12] J. Quinlan, Induction of decision trees, Machine learning 1 (1986) 81–106.

[13] J. Alonso, C. Castiello, L. Magdalena, C. Mencar, Explainable Fuzzy Systems - Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, volume 970, Springer International Publishing, 2021. Https://doi.org/10.1007/978-3-030-71098-9.

[14] R. Guidotti et al., A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51 (2019) 1–42. URL: https://dl.acm.org/doi/10.1145/3236009. doi:10.1145/3236009, https://doi.org/10.1145/3236009.

[15] G. Ras, N. Xie, M. V. Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, Journal of Artificial Intelligence Research 73 (2022) 329–396.

[16] L. Weidinger et al., Taxonomy of risks posed by language models, in: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 214–229.

[17] A. Luccioni, S. Viguier, A.-L. Ligozat., Estimating the carbon footprint of bloom, a 176b parameter language model, Journal of Machine Learning Research 24 (2023) 1–15.

[18] M. Rillig et al., Risks and benefits of large language models for the environment, Environmental Science & Technology 57 (2023) 3464–3466. Https://doi.org/10.1021/acs.est.3c01106.

[19] E. Reiter, R. Dale, Building Natural Language Generation Systems, Studies in Natural Language Processing, Cambridge University Press, 2000. doi:10.1017/CBO9780511519857, https://doi.org/10.1017/CBO9780511519857.

[20] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170.

[21] H. Wang et al., Pre-trained language models and their applications, Engineering (2022). Https://doi.org/10.1016/j.eng.2022.04.024.

[22] J. Dodge et al., Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, arXiv:2002.06305 (2020). Https://doi.org/10.48550/arXiv.2002.06305.

[23] R. Tinn et al., Fine-tuning large neural language models for biomedical natural language processing, Patterns 4 (2023). Https://doi.org/10.1016/j.patter.2023.100729.

[24] G. Neubig, Neural machine translation and sequence-to-sequence models: A tutorial, arXiv:1703.01619 (2017). Https://doi.org/10.48550/arXiv.1703.01619.

[25] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI Global, 2010, pp. 242–264.

[26] M. Mozafari, R. Farahbakhsh, N. Crespi, A BERT-based transfer learning approach for hate speech detection in online social media, in: Proceedings of the Eighth International Conference on Complex Networks and Their Applications, Springer, 2020, pp. 928–940.

[27] H. Strobelt et al., Interactive and visual prompt engineering for ad-hoc task adaptation with large language models, IEEE Transactions on Visualization and Computer Graphics 29 (2022) 1146–1156. Https://doi.org/10.1109/TVCG.2022.3209479.

[28] R. Mahabadi et al., Prompt-free and efficient few-shot learning with language models, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, 2022, pp. 3638–3652. URL: https://aclanthology.org/2022.acl-long.254. doi:10.18653/v1/2022.acl-long.254, https://doi.org/10.18653/v1/2022.acl-long.254.

[29] X. Wu, L. Varshney, A meta-learning perspective on transformers for causal language modeling, arXiv:2310.05884 (2023).

[30] Y. Bai et al., Training a helpful and harmless assistant with reinforcement learning from human feedback, arXiv:2204.05862 (2022). Https://doi.org/10.48550/arXiv.2204.05862.

[31] J. Lu et al., What makes pre-trained language models better zero-shot learners?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL, Toronto, Canada, 2023, pp. 2288–2303. URL: https://aclanthology.org/2023.acl-long.128. doi:10.18653/v1/2023.acl-long.128, https://doi.org/10.18653/v1/2023.acl-long.128.

[32] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, Advances in Neural Information Processing Systems 35 (2022) 462–477.

[33] X. Lin et al., Few-shot learning with multilingual generative language models, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, ACL, Abu Dhabi, United Arab Emirates, 2022, pp. 9019–9052. URL: https://aclanthology.org/2022.emnlp-main.616. doi:10.18653/v1/2022.emnlp-main.616, https://doi.org/10.18653/v1/2022.emnlp-main.616.

[34] T. Brown et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[35] N. Houlsby et al., Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.

[36] R. He et al., On the effectiveness of adapter-based tuning for pretrained language model adaptation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), ACL, Online, 2021, pp. 2208–2222. URL: https://aclanthology.org/2021.acl-long.172. doi:10.18653/v1/2021.acl-long.172, https://doi.org/10.18653/v1/2021.acl-long.172.

[37] G. Melis, C. Dyer, P. Blunsom, On the state of the art of evaluation in neural language models, arXiv:1707.05589 (2017).

[38] Y. Chang et al., A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology (2023). Https://doi.org/10.1145/3641289.

[39] E. Reiter, R. Robertson, L. Osman, Lessons from a failure: Generating tailored smoking cessation letters, Artificial Intelligence 144 (2003) 41–58. URL: https://www.sciencedirect.com/science/article/pii/S0004370202003703. doi:https://doi.org/10.1016/S0004-3702(02)00370-3, https://doi.org/10.1016/S0004-3702(02)00370-3.

[40] Y. Bengio, R. Ducharme, P. Vincent, A neural probabilistic language model, Advances in neural information processing systems 13 (2000).

[41] L. Xue et al., mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Online, 2021, pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[42] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, ACL, 2002, pp. 311–318. Https://doi.org/10.3115/1073083.1073135.

[43] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81. Https://aclanthology.org/W04-1013.