Business & Information Systems Engineering

Foundation models

--Manuscript Draft--

Manuscript Number:	BUIS-D-23-00126
Full Title:	Foundation models
Article Type:	Catchword
Abstract:	N/A (according to common catchword articles)

Foundation models

Abstract N/A

Keywords **-** *Foundation Models, Artificial Intelligence, Emergent Behavior, Homogenization, Transformers*

Introduction

Foundation models (Bommasani et al., 2021) are large-scale artificial intelligence (AI) models, trained on a large amount of broad, typically unlabeled data. These models extend the features of deep learning models in considerable ways, and they can be seen as the next step in the evolution of AI. Large language models, a key type of foundational models, are trained on vast amounts of data to understand and generate text. Engineers have attributed them properties such as conciousness[1]. Models like ChatGPT, GPT-4 OpenAI (2023) and PaLM 2Google (2023) indicate what foundation models are capable of through their generative power, which moves AI beyond addressing specific discriminative tasks such as recognizing pre-defined set of classes: Foundation models can perform tasks that they were not explicitly trained for Brown et al. (2020). This significantly increases the potential application of AI based systems in a variety of contexts Dwivedi et al. (2023).

There has been a significant surge in knowledge about the conceptual foundations, implementation, training, and application of foundation models over the last months. What these, mostly technical, perspectives neglect, however, is the socio-technical nature of systems involving foundational models. The point is that foundation models in terms of their origin (grounded in large amounts of data), use, and impact cannot be seperated from the context involving human actors, tasks, and social structures such as roles, identities, and regulatory frameworks. First, foundation models change work practices as human actors delegate tasks that traditionally fell into the domain of human action –such as tasks requiring visual or textual creativityBasalla, Schneider, and vom Brocke (2022); Chen, Sun, and Han (2023); Skjuve (n.d.). For example, even layman can design furniture based on constraints expressed in text within secondsRamesh et al. (2021). Second, organizations must evaluate foundational models from perspectives

[1] https://www.nytimes.com/2022/07/23/technology/google-engineer-artificial-intelligence.html

such as automating tasks that were typically reserved for humans, developing new value propositions, and rethinking organizational strategy. For example, (adjusted) foundation models can serve as an interface for healthcare providers, e.g., to retrieve relevant cases, suggest diagnosis and treatments (Rasmy, Xiang, Xie, Tao, & Zhi, 2021). In software engineering foundation models can be used for code generation from textual description or finding mistakes in code, i.e., for debuggingSobania, Briesch, Hanna, and Petke (2023). Third, foundation models are likely to change the market power and the broader economical and political landscape associated with AI Agrawal, Gans, and Goldfarb (2018) because both the set-up and the ongoing operation of foundation models is extremely costly. This can potentially limit their development and control to a few powerful organizations Zuboff (2019). Fourth, foundation models are a type of emergent technology, and regulation will focus on how to control the unintended consequences originating in foundational models but, at the same time, try to not inhibit innovation. These developments warrant a distincit information systems perspective on foundation models e.g.Dwivedi et al. (2023) that considers their socio-technical nature Sarker, Chatterjee, Xiao, and Elbanna (2019). Such a perspective can help analyze the multifarious relationships that we can expect to exist between their technical implementation and a variety of tasks, human skills and interests, and social structures Bostrom and Heinen (1977). Such analysis must not blackbox the systems' features and capabilities, but must consider how they are implicated in changing organizational practice–and how they are changed by that practice. With this catchword article, we address the following questions:

1. What are foundation models and what are their key defining features and capabilities?

2. What are challenges and future research opportunities?

To answer these questions, we draw from a recent report by Stanford University's Center for Research on Foundation Models (Bommasani et al., 2021). This extensive report compiled interdisciplinary contributions of more than hundred researchers from ten departments. We combine insights from this report with perspectives from the information systems literature as well as more recent literature not included in the report. We also point to future research directions. To this end, we focus attention on the socio-technical implications of foundation models. Furthermore, we portray foundation models as a historical evolution of machine learning and articulate how they are different in a compact and concise way.

Foundation Models Foundations

A *foundation model* is "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks"(Bommasani et al., 2021). Foundation models can serve as the base for many downstream applications. They can be adjusted (or fine-tuned) with limited labeled training data to a specific task or support multiple tasks at once(Reed et al., 2022). Foundation models constitute a paradigm shift in AI development and use. They extend and go beyond the capabilities of earlier deep learning models, which is the common reference point for modern, powerful AI applications. Before we explain the technical features of foundation models, we describe their development in light of their broader historical context.

History: From expert systems to foundation models

Foundation models can be seen as a "logical" step in the development of machine learning towards self-supervised, larger deep learning models as shown in Figure 1.

Model Size (#parameters)	None (Hand-designed rules)	Very few	Few	Large	Very large
Features	Hand-designed	Hand-designed	Learnt (Simple features)	Learnt (Hierarchy of features)	Learnt (Hierarchy of features)
Learning	None	Supervised	Supervised	Supervised	Self-supervised + in-context learning
Data	None	Very few labeled-data	Few labeled-data	Large labeled-data	Very large unlabeled data (+ small labeled data)
Adaptability	None	None	Little	Medium	Large

Figure 1: History of machine learning

Before learning from data, decision-making by machines was done by expert systems(Jackson, 1986) starting with the introduction of computers in the 1940ies. They were popularized in the 1960ies through systems such as Dendral (Feigenbaum, 1977). These early AI systems operated on the basis of a set of detailed rules to specify how an input is transformed into an output; the development of these rules, in turn, made them costly and brittle. The next important step was the invention of machine learning in the 1960ies. These systems can operate "without explicitly being programmed", as famously coined by Samuel Jackson in 1959. Early machine learning learned decision rules based on a set of criteria, i.e., features, defined by experts. That is, the definition of these features was subject to human designers.

It required domain and technical expertise (e.g., see (Bay, Tuytelaars, & Gool, 2006; Lowe, 1999) for examples in computer vision). In the 2000s, representation learning became a dominant approach. The main goal of representation learning was to automate the identification of features. These attempts were originally limited to learn relatively simple features.

Deep learning achieved a breakthrough in representation learning. It reduced the need for feature engineering and allowed learning a hierarchy of features (Goodfellow, Bengio, & Courville, 2016, p. 1). Learning such hierarchies also demanded much larger datasets and models, measured by the number of estimated parameters. For example, the well known Imagenet-1k dataset(Deng et al., 2009) from 2009 has about 1 Mio. labeled samples, while prior image datasets such as MNIST from 1994 were more than an order of magnitude smaller. Deep learning also allowed building models in a modular, flexible way by stacking various layers on top of each other. This approach makes it easy to enlarge models or combine models for data of different modalities such as text and images. Deep learning can be seen as an evolution of early neural networks developed in the 1940ies(McCulloch & Pitts, 1943; Schmidhuber, 2015; Wang & Raj, 2017). Its success can be attributed primarily to computation and data processing capabilities, along with technical innovations. Deep learning transcended diverse areas of AI, including computer vision, speech, and natural language processing. Specific deep learning models were developed through different compositions of basic elements for a wide variety of tasks. The most prevalent paradigm for training models was supervised learning, i.e., training on a labeled dataset: Each input is associated with a label commonly provided by a human. Although models were developed based on specific datasets, many learned features stored within a model could be reused across similar datasets through transfer learning (also called "fine-tuning")(Thrun, 1998).

Foundation models emerged around 2017. They are very large deep learning models relying on self-supervised learning. That is, they learn from large amounts of unlabelled data and show the capability of in-context learning(Brown et al., 2020), i.e., trained models can solve tasks without explicitly being trained on them, i.e., without model parameter updates. This universality of models is a key and unexpected characteristic of foundation models. A main success story of these models has been text generation based on user inputs. For example, short texts generated from GPT-3 might be hard to distinguish from those of humans(Clark et al., 2021). Before the area of foundation models, generating general-purpose language was seen as very difficult and only approachable through other linguistic subtasks (Paris, Swartout, & Mann, 2013). A few years later, after the success in natural language processing

(NLP), transformer variants(Dosovitskiy et al., 2020) also improved state-of-the-art results in computer vision. However, the gains were less profound. Transformers have also entered other areas. Speech recognition systems are increasingly built based on a large dataset of audio alone and then adapted with audio and associated transcriptions (Baevski, Zhou, Mohamed, & Auli, 2020).

Key Features of Foundation Models

Foundation models leverage a major breakthrough in natural language processing (NLP): transformers(Vaswani et al., 2017). Technically, transformers constitute a novel deep learning architecture built using existing design elements, i.e., layers. They are flexible in the sense that they make few assumptions about the data characteristics and therefore can learn many functional dependencies.

Foundation models mostly differ from prior deep learning models based on their expressivity (ability to capture diverse information) and scalability (ability to leverage large amounts of data both in terms of computational efficiency as well as being able to keep learning with growing amounts and diversity of data). Other relevant aspects that exhibit fewer differences to prior deep learning models include multi-modality (the processing of multimodal data such as images and text in one model), memory (representing and retrieving knowledge, possible from a model external source), and compositionality (modularity of the model)[2].

Foundation models commonly achieve state-of-the-art results on specific tasks outperforming tailored models. Furthermore, even if adjusted to a task, they tend to require less task-specific (labeled) data to perform well, which makes them attractive since labeled data (i.e., the process of labelling data) can be costly. Their success is due to the sheer size of models and the diversity and amount of training data. They learn from large, unlabeled data in a self-supervised manner. In supervised learning a data item consists of an input and a desired output, i.e., label. For example, for image recognition an image of a cat might have the label cat. *Self-supervised learning* learns from task-agnostic data with self-generated labels (rather than from human annotators). Labels originate from artificially generated tasks. For example, for language models (such as GPT-1 to ChatGPT), the task is to predict the next word in a document. Even in case foundation models can solve a task without further adjustments, additional adaptation often improves performance. For example, raw language models might provide incorrect, toxic, or low quality outputs from a user perspective. That is, they might lack user alignmentD. Ouyang et al. (2020). Ad-

[2]This is a deviation from (Bommasani et al., 2021) that regards compositionality as "ability to generalize through its representation and model modularity". We view compositionality more narrowly as "model modularity".

justing the (pre-trained) model using labeled data from human feedback can improve user alignment and help to better respond to textual instructions. Such data can be collected through demonstrations, i.e., a labeler provides outputs directly or ranks multiple options. Labelers are carefully chosen according to a variety of criteria, e.g., they should be sensitve to preferences of diverse ethnic groups. This mechanism of adjustment has been employed in recent models such as InstructGPTL. Ouyang et al. (2022) and GPT-4OpenAI (2023). However, the idea to fine tune models on instructions ("instruction tuning") was introduced priorlyWei et al. (2021).

Two characteristics of foundation models are to be emphasized: in-context learning and homogenization.

In-context Learning and Prompting

Importantly, foundation models can be used to address tasks not explicitly trained for through in-context learning. The model extracts a rich set of patterns and broad skills from the diverse training data. In turn, they can perform downstream tasks simply by providing a *prompt*, i.e., a description of the task in natural language or through a visual representation and, possibly, a few examples. For instance, language models such as GPT-3 and later versions can translate text as illustrated in Figure 3 or add numbers by prompting, e.g., "4+5 = " – although they were not trained on any mathematics task. Furthermore, performance can be improved by crafting adequate prompts. For example, Kojima, Gu, Reid, Matsuo, and Iwasawa (2022) showed that by simply adding "Let's think step by step." to an input, the model performed significantly better on various benchmarks. This *emergent behavior* has been an unexpected phenomenon even for AI experts. It has not been observed in smaller models trained with fewer data. For language model, one can distinguish two types of prompts: (i) cloze prompts, which predict the blanks of a textual string (e.g., "I love this car, it is a [X] car"), and prefix prompts, which expand a given string (e.g.,"I love this car. What's the sentiment of the statement? [X]"). Furthermore, a catalogue of prompting patterns(White et al., 2023) has been proposed to address common problems using language models. For example, the intention of the "flipped interaction" pattern is for the LLM to enquire the needed information to perform a task, e.g., a user might prompt "ChatGPT ask me questions to achieve [X]." This makes it easier for users to anticipate needed or impactful information to state a prompt for a task. The "Fact Check List" Pattern aims to ensure that users understand the assumptions (and facts) on which a response is based. A user might add to a query something like "The set of facts should be the fundamental facts that could undermine the veracity of the output if any of them are incorrect". However, there is no guarantee that the prompts fulfill their intended purpose. Prompting has also been studied for text-to-image models V. Liu and Chilton (2022).

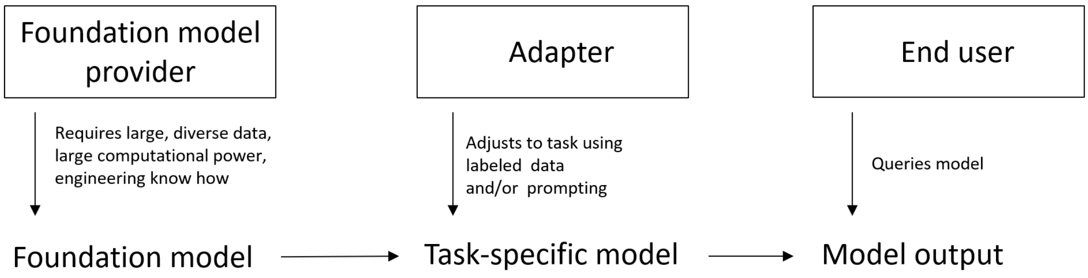

Figure 2: Actors and outcomes in AI development involving foundation models.

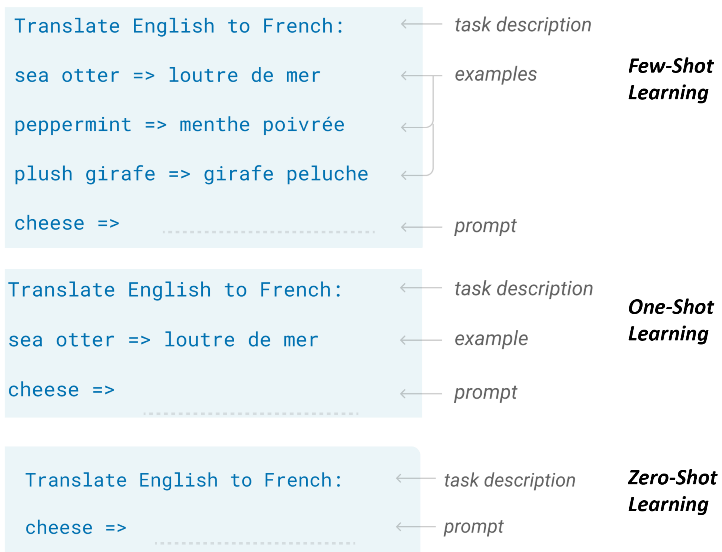

Figure 3: The emergent phenomenon of *in-context learning*: It enables solving tasks the model was not explicitly trained for using a textual description and possibly examples ("shots"). Adjusted from (Brown et al., 2020)

Homogenization

Foundation models might transcend all areas of AI through *homogenization*Bommasani et al. (2021) and lead to a concentration of power. Homogenization can occur in three ways, new AI models are adjustments of (i) a few foundational models, (ii) trained on a few datasets, or (iii) by a few organizations. Foundation models might reduce deep learning model diversity because there is less need to encode task-specific information in a model by model developers. Instead, general models can be employed due to the increased richness of training data. Foundation models might be trained on all globally available data, e.g., the entire Internet or large proprietary datasets such as images on social media platforms used by billions of people. This is costly, technically demanding, and requires access to large data, implying that only a few organizations are capable of doing so. For example, the training of GPT-3 is estimated to cost millions of dollars[3]. However, foundation models can be adapted to specific tasks using little (labeled) data or prompting. The actors involved in AI development might change. Some could be foundation model providers, others will act as foundation model adapters based on fine-tuning with labeled data or prompting, as illustrated in Figure 2. The end user might explicitly or implicitly query a model, e.g., if a model is embedded in a tool to detect grammar or spelling errors the interaction is implicit. The choice of adaptation mechanism depends mostly on three factors: compute budget, data availability, and access to the foundation model. Adjusting the model itself is only feasible if one can modify model internals either directly or indirectly, e.g., a foundation model provider might provide the possibility to upload data and specify training. This is called transfer learning, i.e. the model is adjusted to a specific task using a (small) labeled data set. In-context learning allows using a foundation model without any further additions or changes to the model for novel tasks through designing inputs ("prompting") and it is therefore always possible except if model providers forbid or block certain prompts. For example, while in February 2023 ChatGPT would generate jokes specifically about men, e.g., if prompted "Tell a joke about men.", later versions (March 2023) refused to do so.

1 Challenges and Opportunities for IS Research

Foundation models can be thought of as what economists refer to as a general-purpose technology(Bresnahan & Trajtenberg, 1995). They might increase productivity and innovation and boost the adoption of AI. Prior to wide-spread accessibility to models such as ChatGPT and Dall-E, research was most exclusively done by computer scientists. However, foundation models could lead to far-reaching implications on people, technology, and organizations. Research questions are by no means limited to these areas. For example, regulations and philosophical questions are also important to study. More concretely, it is unclear to what extent such models will or can possess human traits and, in turn, to what extent they even deserve special rights, e.g., is "speech by AI" covered under the First Amendment protecting freedom of speech under US law? Given the broad range of research questions, a multi-disciplinary perspective is needed to understand the implications, challenges and provide solutions to problems. Information systems researchers are therefore well-equipped to make important contributions.

Some opportunities and challenges overlap with deep learning, e.g., challenges in XAI(Meske, Bunde,

[3] https://lambdalabs.com/blog/demystifying-gpt-3

Schneider, & Gersch, 2022), opportunities of AI in business (Nguyen, Sidorova, & Torres, 2022), and ethical concerns such as fairness(Feuerriegel, Dolata, & Schwabe, 2020). Challenges and opportunities can be derived from the foundational character of these models, which stems from (i) performance boosts on existing tasks, (ii) their adaptability to many tasks, (iii) homogenization ("access, power and control of these models"), (iv) novel emergent behavior ("in-context learning"), and (v) resource needs. Each of these characteristics can be investigated from various perspectives such as people, technology, and organizations as illustrated next through a few important areas and questions worth further investigation.

1.1 People

Due to being trained on large amounts of data and their generative capability, deep learning and, more so, foundation models can be used for many new applications by people of all age groups and social backgrounds. They have also been reported to be used for productivity, entertainment, and social interaction Brandtzaeg and Følstad (2017); Skjuve (n.d.). That is, people have used them to discuss personal problems. Schoolchildren use them to do their homework. Architects can get ideas for the design of houses. However, overall it is not yet well understood what applications are suitable and how to integrate foundation models into these applications.

Foundation models can foster interactivity such as needed in chatbots(Adamopoulou & Moussiades, 2020). They possess the ability of in-context learning encouraging to alter or design prompts, i.e., a user can query a model, investigate the output, adjust the query by providing an example, etc. However, prompt engineering is not yet wel–understood though research is growing at a rapid pace(P. Liu et al., 2021; W. X. Zhao et al., 2023).

Misuse can arise due to increased content quality, leading to content that is often indistinguishable from those of humans(Clark et al., 2021) at low costs reducing the barrier for harmful attacks(Brundage et al., 2018). Understanding risks for misuse, prevention and detection are important areas that require training of society as a whole aside from technological and regulartory measures.

Homogenization can amplify bias and arbitrary exclusion (Creel & Hellman, 2022). It can also lead to "culture homogenization", i.e., spreading one implicit perspective across multiple domains. Thus, it is essential to better understand the limitations of these models such as biases and hallucination (fabrication of facts and self-contradict)Ji et al. (2023). Since outputs of many current models lack proper attribution (to the training data), i.e., sources are missing that could be checked by an end-user, verifying outputs is difficult. In turn, this leads to many (research) objectives: understanding issues, educating people, deriving norms for foundation model development.

1.2 Technology

Mitigating inequities such as biases on the technical level aims at interventing at different steps of the model development pipeline (e.g., data (Lu, Mardziel, Wu, Amancharla, & Datta, 2020), model objectives (J. Zhao, Zhou, Li, Wang, & Chang, 2018), and adaptation). It is challenging since the possible uses of foundational models are unknown, and issues might stem from various sources along the pipeline. In turn, this makes it more difficult to assign responsibilities, and it makes a proactive intervention difficult.

Foundation models posit multiple risks beyond biases. The risk of misalignment between the training objective and the wanted behavior is considerable, which poses safety risks. For example, a common objective for general-purpose language models is to predict the next word within a document. But the desired goal can be to output only true or helpful text(Tamkin, Brundage, Clark, & Ganguli, 2021). Defining proper objectives and evaluation measures and tests for these models taking into account the impact on people is a further area of study.

Foundation models constitute a single point of failure. That is, if a model has memorized private data of an organization or an individual, all adjusted models could potentially leak this sensitive information. Data poisoning, i.e., the intentional introduction of data samples into training data to manipulate model behavior is a key concern since data might be scrapped from public sources such as the Internet.

Due to their large model size and data size, foundation models cause high energy consumption during training and operation. Models like GPT-3 or Gopher produce more than CO_2eq than a roundtrip flight with 300 passengers from New York to London, i.e., more than 300 tCO_2eq (Rae et al., 2021). Some techniques to reduce energy consumption used in classical deep learning(Schneider, Seidel, Basalla, & vom Brocke, 2022) might be deployed.

Furthermore, these models are still based on the transformer model from 2017. Fresh ideas outside computer science might also help to significantly improve models or and maybe lead to novel models marking another breakthrough in AI.

1.3 Organization

Many applications leveraging foundation models in a commercial and non-commercial setting primarily rely on usage of foundation models without model adjustments. However, possible gains in performance and their adaptability to diverse tasks with few labeled data might foster the adoption of AI and enable new applications for organizations. Companies might fine-tune models for their specific needs, e.g., a company might adjust a lanuage model like ChatGPT based on data of customer call centers to make a novel chatbot. Dependencies on suppliers of AI-related technology might change. That is, a company might rely on a foundation model provider with proprietary models (see Figure 2) rather than training models from scratch or using publicly pre-trained models. With increased adaptation to a foundation model dependencies can increase, while at the same time, the fact that adaptation tends to require fewer data and training effort might also make it easier to switch between foundation model providers. The actors in delivering AI services might change, e.g., AI products might rely on foundation models provided by one company that are further adapted by another company. In contrast, today commonly one company develops an AI model. As of now, it is not yet clear, for which applications it makes sense to adapt AI models by fine-tuning with additional labeled data. It might be more difficult to assess model risk and behavior. For once, model access might be limited, e.g., if only outputs are accessible, certain explainability methods that require internal model information cannot be used. Data access might also be limited, making it difficult to assess data quality and data biases.

Investigating applications in an organizational context leads to many subquestions, e.g., how to adjust business processes? How to govern (training) data, foundation models, and adapted models? Is it economical to collect data and fine-tune such models? How to ensure accountability? What are business models and services emerging from new applications enabled by foundation models? Existing data and AI governance frameworks might be leveraged, e.g., (Schneider, Abraham, Meske, & Vom Brocke, 2022). Existing works, e.g., on AI as a service (Lins et al., 2021), might provide a starting point.

The use of large, uncurated data increases the risk of privacy and copyright issues. For outputs, liability issues can arise as well as legal protections for outputs touching on ownership. For example, systems like Dall-E generate illustrations of unpreceded quality based on textual descriptions. However, these models rely on data possibly scraped from the Internet. In turn, this raises questions such as: How is intellectual property handled and attributed? Can outputs by an AI based on public data by owned by the AI company or should it be given to the data creators? Can a person having discovered prompts to achieve striking outcomes claim a "copyright" to these prompts, similar to program code in ordinary software?

2 Outlook

Already today foundation models have started to revolutionize information retrieval as witnessed by the recent integration of GPT-4 into Microsoft's search engine BingBingBlogs (2023). They might also reduce the need for labeled data, which is an important cost factor. Novel phenomena such as in-context learning can enhance the possibilities how one can interact wiht these models, while homogenization leads to new power and control structures. AI might shift from a technology characterized by openness, where pre-trained models are shared even by research institutions of industrial companies, to a technology that becomes inaccessible to researchers and controlled by a few organizations. However, it is unclear, how long power will remain concentrated. Judging from a computational perspective, given Moore's law within one, two or maybe three decades such models could become our personal assistants running on smartphones, similarly to how computers owned by large corporations in the 60ies costing millions of dollars became personal computers in the 80ies. However, access to large datasets would also be needed. Multiple decades might seem like a long time into the future, thus, one might ask "What comes after today's foundation models?". Our historical analysis hints at an obvious trend: "Larger, more compute-intensive models trained on more data of larger diversity". Even larger models might be more versatile through adaptation and prompting, leading to even broader adoption in practice. Thus, research should tackle pressing issues related to foundation models as soon as possible.

Acknowledgements: Special thanks goes to X and Y for fruitful discussions on the manuscript.

References

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 373–383).

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Basalla, M., Schneider, J., & vom Brocke, J. (2022). Creativity of deep learning: Conceptualization and assessment. In *Proceedings of the 14th international conference on agents and artificial intelligence, ICAART* (pp. 99–109).

Bay, H., Tuytelaars, T., & Gool, L. V. (2006). Surf: Speeded up robust features. In *European conference on computer vision* (pp. 404–417).

BingBlogs. (2023). Confirmed: the new bing runs on openai's gpt-4. $https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAIs-GPT-4.(Accessed2023-04-20)$

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. Part I: The causes. MIS quarterly*, 17–32.*

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In International conference on internet science *(pp. 377–392).*

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies 'engines of growth'? Journal of econometrics*,* 65*(1), 83–108.*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. Advances in neural information processing systems*,* 33*, 1877–1901.*

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... others (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Chen, L., Sun, L., & Han, J. (2023). A Comparison Study of Human and Machine Generated Creativity. Journal of Computing and Information Science in Engineering*, 1–37.*

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. arXiv preprint arXiv:2107.00061.

Creel, K., & Hellman, D. (2022). The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. Canadian Journal of Philosophy*, 1–18.*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 ieee conference on computer vision and pattern recognition *(pp. 248–255).*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... others (2023). "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. International Journal of Information Management, 71*, 102642.*

Feigenbaum, E. A. (1977). The art of artificial intelligence: Themes and case studies of knowledge engineering. In Proceedings of the fifth international joint conference on artificial intelligence *(Vol. 2).*

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair ai. Business & information systems engineering, 62*(4), 379–384.*

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT press.*

Google. (2023). Palm 2 technical report. https://ai.google/static/documents/palm2techreport.pdf.

Jackson, P. (1986). Introduction to expert systems.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55*(12), 1–38.*

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916.

Lins, S., Pandl, K. D., Teigeler, H., Thiebes, S., Bayer, C., & Sunyaev, A. (2021). Artificial intelligence as a service. Business & Information Systems Engineering*, 63(4), 441–456.*

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586.

Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 chi conference on human factors in computing systems *(pp. 1–23).*

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In Proceedings of the seventh ieee international conference on computer vision *(Vol. 2, pp. 1150–1157).*

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender bias in neural natural language processing. In Logic, language, and security *(pp. 189–202). Springer.*

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics*, 5(4), 115–133.*

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Information Systems Management*, 39(1), 53–63.*

Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). Artificial intelligence in business: A literature review and research agenda. Communications of the Association for Information Systems*, 50(1), 7.*

OpenAI. (2023). Gpt-4 technical report.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., … others (2020). Video-based ai for beat-to-beat assessment of cardiac function. Nature*, 580(7802), 252–256.*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … others (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems*, 35, 27730–27744.*

Paris, C. L., Swartout, W. R., & Mann, W. C. (2013). Natural language generation in artificial intelligence and computational linguistics *(Vol. 119). Springer Science & Business Media.*

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... others (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). Zero-shot text-to-image generation. In International conference on machine learning *(pp. 8821–8831).*

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine*,* 4*(1), 1–13.*

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... others (2022). A generalist agent. arXiv preprint arXiv:2205.06175.

Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance. MIS quarterly*,* 43*(3), 695–720.*

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks*,* 61*, 85–117.*

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2022). Artificial intelligence governance for businesses. Information Systems Management*, 1–21.*

Schneider, J., Seidel, S., Basalla, M., & vom Brocke, J. (2022). Reuse, reduce, support: Design principles for green data mining. Business & Information Systems Engineering*, 1–19.*

Skjuve, M. (n.d.). Why people use chatgpt. Available at SSRN 4376834.

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. arXiv preprint arXiv:2301.08653.

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503.

Thrun, S. (1998). Lifelong learning algorithms. In Learning to learn *(pp. 181–209). Springer.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems*,* 30*.*

Wang, H., & Raj, B. (2017). On the origin of deep learning. arXiv preprint arXiv:1702.07800.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., … Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning gender-neutral word embeddings. arXiv preprint arXiv:1809.01496.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … others (2023). A Survey of Large Language Models. arXiv preprint arXiv:2303.18223.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack obama's books of 2019. *Profile books.*