Recent Trends in China's Large Language Model Landscape

Jeffrey Ding[1] and Jenny W. Xiao[2] | April 2023

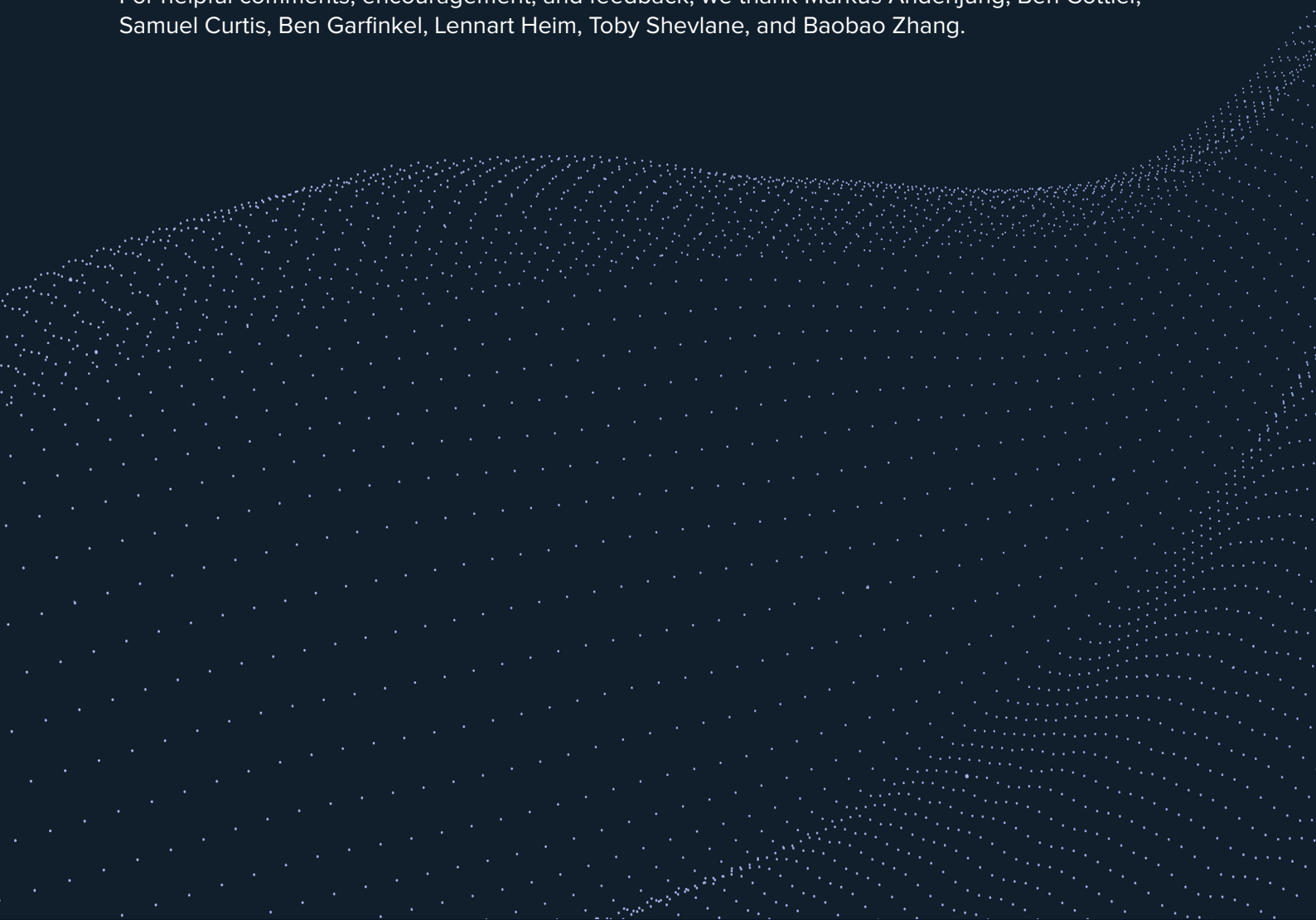
ABSTRACT

As large-scale pre-trained AI models gain popularity in the West, many Chinese AI labs have developed their own models capable of generating coherent text and realistic images and videos. These models represent the frontier of AI research and have significant implications for AI ethics and governance in China. Yet, to the best of our knowledge, there has been no in-depth English-language analysis of such models. Studying a sample of 26 large-scale pre-trained AI models developed in China, our review describes their general capabilities and highlights the role of collaboration between the government, industry, and academia in supporting these projects. It also sheds light on Chinese discussions related to technonationalism, AI governance, and ethics.

[1] Department of Political Science, The George Washington University
[2] Department of Political Science, Columbia University

ACKNOWLEDGMENTS

For helpful comments, encouragement, and feedback, we thank Markus Anderljung, Ben Cottier, Samuel Curtis, Ben Garfinkel, Lennart Heim, Toby Shevlane, and Baobao Zhang.



Recent Trends in China's Large Language Model Landscape

Background

Over the last two years, there has been a surge in Chinese large-scale pre-trained artificial intelligence (AI) models[1]. Top industry and academic labs in China have begun to compete more directly with Western labs by building large-scale AI systems analogous to OpenAI's GPT-3[2], DeepMind's Chinchilla[3], and Google's PaLM[4]. These models—both those of China and of the West—raise thorny governance questions related to issues such as fairness and privacy. They also represent progress in a cutting-edge area of AI that may lead to increasingly general AI systems.

This paper represents the first detailed study of Chinese large-scale AI models. Alongside the Chinese government's issuance of its New Generation AI Development Plan and the impressive rise of Chinese AI firms and research organizations, there has been growing interest in AI developments in China[5,6]. By studying a selected set of Chinese projects, we hope to offer insights into general trends in China's AI ecosystem.

Methodology

This study focused on 26 Chinese large-scale pre-trained models released between 2020 and 2022. Though this is not an exhaustive list of all such models, it includes the most significant large-scale language and multimodal models, based on parameter count and training compute, from Chinese labs. We break down the models by various features, including affiliations, funding sources, data, training compute, parameters, benchmark performance, discussions of ethics and governance, and model access and usage policies. In addition to studying published materials, we conducted interviews with researchers at leading Chinese labs, some of whom are directly involved in the development of the models. We also study Chinese discourse surrounding these large-scale models, reviewing conversations in relevant WeChat public accounts, Zhihu threads, as well as mainstream media.

Main Findings

We find that, although it is often difficult to compare the performance of Chinese and Western AI models due to differences in their linguistic media, Chinese models have performed quite well on English-language benchmarks. The popularization of large-scale models has also encouraged the development of associated infrastructure, including Chinese-language benchmarks and datasets. Compared with Western governments, the Chinese government plays a much more prominent role in China's AI ecosystem, often directly facilitating industry-academia cooperation and providing significant compute funding.

In contrast to the prevailing impression in the West that there is less discussion of AI governance and ethics in China, half of the projects we examine discuss ethics or governance issues, including bias, the potential for misuse, and the environmental repercussions of large-scale modeling training. Chinese models can be accessed in a variety of ways:

through APIs, direct download options, or open-source platforms. Some organizations place access restrictions on some of their models, likely due to misuse concerns.

Finally, we find that, perhaps as a result of growing U.S.-China tensions, technonationalist discourse around large-scale AI models on traditional and social media in China is quite prevalent. For instance, the fact that OpenAI's API is not available in China has triggered anxiety that the country is being blocked out of the generative AI revolution happening in the West. At the same time, Chinese observers take pride in their domestic large-scale models, and companies proudly advertise their use of domestically produced hardware and infrastructure for model training.

General Characteristics of Leading Models

Performance

Despite the difficulty of comparing English- and Chinese-medium AI models, two approaches to model evaluation indicate that the performance of top Chinese models is not far behind the state of the art (SOTA) in the West. The first approach looks at the sheer size of Chinese models. As large-scale pre-trained models usually follow "scaling laws," meaning that their performance improves predictably with scale, the parameter sizes of these models can serve as a rough proxy for performance[7]. Within a year of OpenAI's May 2020 announcement of GPT-3 (175B parameters), Chinese developers released even larger models, such as PanGu-α (207B parameters) and Yuan 1.0 (245B parameters). The government-sponsored Beijing Academy of Artificial Intelligence (BAAI) also released one of the world's largest mixture-of-expert (MoE) models called BaGuaLu (14.5T parameters). However, the race to develop ever larger models abated with the release of DeepMind's Chinchilla model, which shows that increasing training data may be as important as scaling up model size[3].

A second approach to evaluating the performance of Chinese models is a direct evaluation against Western benchmarks, as some Chinese models have English training data and translation tools can be used to facilitate comparison. Baidu's ERNIE 3.0, for example, set a new SOTA on SuperGLUE with its English version[8]. WuDao 1.0's English-language Wen Hui (GLM) model was also evaluated on the SuperGLUE benchmark[9]. Some multimodal models also have English versions that can be directly compared with Western models. MSRA and Peking University's multimodal NÜWA achieved SOTA results on eight downstream tasks, including text-to-image generation, text-to-video generation, and video prediction[10]. Baidu's ERNIE-ViLG uses the Baidu Translate API to enable comparison with OpenAI's DALL-E in image generation[11]. BAAI's CogView team also uses caption translation to compare the performance of CogView with DALL-E. Both of these models report a performance improvement over DALL-E[12].

Of course, many of the 26 models are only trained on Chinese data and evaluated on Chinese-language benchmarks. The rise of large-scale pre-trained models facilitates the maturation of associated benchmarks. The release of earlier Chinese models also makes it easier to benchmark the performance of later models. For instance, Huawei's PanGu-α was compared against BAAI's WuDao 1.0-Wen Yuan (CPM-1). The rapid advancement of large-scale natural language processing (NLP) models led to the development of the CLUE (Chinese Language Understanding Evaluation) and CUGE (Chinese Language Understanding and Generation Evaluation) benchmarks, which became two of the most widely used Chinese-language benchmarks for evaluating large-scale pre-trained models. Notable large-scale pre-trained models evaluated on CLUE include ERNIE 3.0 Titan, CPT, and Yuan 1.0, while WuDao 2.0-Wen Yuan (CPM-2) was evaluated on the CUGE benchmark.

Table 1: Key Metrics for Chinese Models

Name	Release Date	Developer(s)	Parameters (Of Largest Model)	Training Compute (FLOPs)	Compute Hardware	Significance and Benchmarks
WuDao-Wen Yuan 1.0 (CPM)[13]	12/1/2020	BAAI, Tsinghua University	2.6B	6.5E+20	64 Nvidia V100 GPUs for two weeks	China's first large-scale pre-trained model; achieves strong performance on many NLP tasks on few-shot and zero-shot learning.
M6[14]	3/1/2021	Tsinghua University, Alibaba	100B (Mixture of Experts, MoE)		128 Nvidia A100 GPUs (time unspecified)	Exceeds strong baseline in VQA, image captioning, and image-text matching and is also able to generate high-quality images.
WuDao-WenLan 1.0[15]	3/11/2021	Renmin University of China, Chinese Academy of Sciences	1B	7.2E+21	128 Nvidia A100 GPUs for 7 days	Outperforms both UNITER and OpenAI's CLIP in various downstream tasks.
WuDao-Wen Hui 1.0 (GLM)[9]	3/18/2021	Tsinghua University, BAAI, MIT, Shanghai Qi Zhi Institute	11.3B	1.2E+20	64 Nvidia V100 GPUs for 2.5 days	Surpasses BERT, T5, and GPT on a wide range of tasks across NLU and conditional and unconditional generation.
PLUG[16]	4/19/2021	Alibaba	27B	3.6E+22	Trained on Alicloud for 35 days	Renewed SOTA performance on CLUE.
PanGu-α[17]	4/26/2021	Huawei, Recurrent AI, Peking University, and Peng Cheng Lab	207B	5.83E+22	2048 Huawei Ascend 910 AI processors	Superior performance on various generation, QA, and summarization tasks under few-shot or zero-shot settings.
ConSERT[18]	5/25/2021	Beijing University of Posts and Telecommunications, Meituan	345M	1.0E+15 to 5.0E+15 (2 to 10 min)	A single Nvidia V100 GPU for a few minutes	Trains an effective BERT model on small sample sizes and achieves an 8% improvement over previous SOTA on STA datasets.
CogView[12]	5/26/2021	Tsinghua University, Alibaba, BAAI	4B	2.68E+22	512 Nvidia V100 GPUs	Achieves SOTA FID on the blurred MS COCO dataset, outperforming previous GAN-based models and OpenAI's DALL-E.
M6-T[19]	5/31/2021	Alibaba	1T (MoE)	5.50E+21	480 Nvidia V100 GPUs (time unspecified)	Managed to scale up significantly while limiting compute costs.
WuDao-Wen Yuan 2.0 (CPM-2)[20]	6/24/2021	Tsinghua University, BAAI	11B dense bilingual model and 198B MoE version		Compute support from BAAI (time unspecified)	Surpasses mT5 on many downstream tasks.
ERNIE 3.0[8]	7/15/2021	Baidu	10B	2.4E+18	384 Nvidia V100 GPUs trained for a total of 375 billion tokens	English version set new SOTA on SuperGLUE; achieved SOTA on 54 Chinese NLP tasks.
PLATO-XL[21]	9/20/2021	Baidu	11B		256 Nvidia V100 GPUs (time unspecified)	Obtains SOTA across multiple conversational tasks.
Zidong Taichu[22]	9/27/2021	Chinese Academy of Sciences	100B		Huawei Ascend (time unspecified)	The world's first image, language, and audio trimodal pre-trained model.
M6-10T[23]	10/8/2021	Alibaba	10T (MoE)	5.5E+21	512 Nvidia V100 GPUs for 10 days	Proposes the "Pseudo-to-Real" training strategy.

Name	Release Date	Developer(s)	Parameters (Of Largest Model)	Training Compute (FLOPs)	Compute Hardware	Significance and Benchmarks
Yuan 1.0[24]	10/10/2021	Inspur	245B	4.10E+23	2128 GPUs (GPU manufacturer and time unspecified)	Achieves SOTA on a series of generation and comprehension tasks, surpassing PanGu-α and ERNIE 3.0.
WuDao-WenLan 2.0[25]	10/27/2021	Renmin University of China, Beijing Key Laboratory of Big Data Management and Analysis Methods, King Abdullah University of Science and Technology, University of Surrey, MIT	5.3B	9.0E+21	112 Nvidia A100 GPUs for 10 days	Develops a large-scale multimodal foundation model that surpasses OpenAI's CLIP and Google's ALIGN in downstream tasks, including visual QA, cross-modal retrieval, and news classification.
NÜWA[10]	11/24/2021	Microsoft Research Asia, Peking University	870M	7.2E+21	64 Nvidia A100 GPUs for two weeks	Develops a multimodal model that can generate or manipulate images and videos and achieves SOTA performance on 8 tasks.
ERNIE 3.0 Titan[26]	12/23/2021	Baidu, Peng Cheng Lab	260B	4.2E+22	Nvidia V100 GPUs at Baidu and Huawei Ascend 910 NPUs at Peng Cheng Lab	Was the largest Chinese dense pre-trained model at the time, outperforming SOTA models on 68 NLP datasets.
ERNIE-ViLG[11]	12/31/2021	Baidu	10B		Unspecified	Achieved SOTA results on text-to-image and image-to-text results.
BaGuaLu[27]	4/2/2022	Tsinghua University, BAAI, Alibaba, Zhejiang Lab	14.5T (MoE), able to train up to 174T (MoE)		New Generation Sunway Super computer	The first "brain-scale" pre-trained model trained on an exascale supercomputer.
CogVideo[28]	5/29/2022	Tsinghua University, BAAI	9B		Unspecified	The world's largest and first open-source large-scale pre-trained text-to-video model.
GLM-130B[29]	8/4/2022	Tsinghua University, BAAI, Zhipu.AI	130B	4.6E+22	96 Nvidia A100 GPUs for 2 months	An open-source bilingual language model that can be downloaded on a single server.
Taiyi-Stable Diffusion[30]	10/31/2022	IDEA CCNL	1B	2.55E+22	32 Nvidia A100 GPUs for 100 hours	The first open-source, Chinese version of Stable Diffusion.
AltDiffusion [31]	11/12/2022	BAAI				A multilingual, multimodal model.
AR-LDM[32]	11/20/2022	University of Waterloo, Alibaba, Vector Institute	1.5B	5.1E+20	8 Nvidia A100 GPUs for 8 days	The first latent diffusion model for coherent visual story synthesizing.
ALM 1.0[33]	11/28/2022	BAAI				SOTA results on Arabic-language benchmark ALUE.

Assessments of the WuDao models illustrate some issues in comparing Chinese models with their Western counterparts. BAAI's WuDao 2.0, a 1.75 trillion parameter model, was announced to much fanfare in the summer of 2021. English-language coverage declared that this was the largest NLP model ever trained. With ten times more parameters than GPT-3, WuDao 2.0 was tagged as a symbol of "bigger, stronger, faster AI from China."[34] However, WuDao 2.0 was overhyped in a few ways. First, it was not a single model capable of tasks ranging from language generation to protein folding, as reporting implied, but rather a series of models, including the language model WuDao-Wen Yuan 2.0 (CPM 2.0), the multimodal model WuDao-WenLan 2.0 (BriVL 2.0), and the image generation model CogView. The multimodal model M6, a joint project between Alibaba and Tsinghua University, was also part of the WuDao 2.0 release. Moreover, due to possible differences in prompt engineering, WuDao 2.0's performance is not necessarily comparable to that of GPT-3 on certain benchmarks[35]. Lastly, there is no paper that clearly outlines the 1.75 trillion parameter model, making it difficult to assess WuDao 2.0's capabilities[35].

Compute

Another significant aspect of Chinese large-scale models is the compute and hardware required to train them. Studying the compute requirements of these models reveals three takeaways. First, some Chinese models match and surpass the first popular large-scale pre-trained model, OpenAI's GPT-3, in terms of training compute[36]. Inspur's Yuan 1.0 was trained for 4095 petaFLOP/s-days, which the Inspur researchers explicitly compare to GPT-3's training compute of 3650 petaFLOP/s-days[24]. The researchers of Baidu's ERNIE 3.0 Titan estimate that training the model with 2048 Nvidia V100 GPUs required 28 days, which amounts to around 3634 petaFLOP/s-days[26].

Second, there is evidence that some Chinese models suffered in performance due to compute limitations. Huawei's PanGu team, for instance, stopped training its 200B-parameter model after 40 billion tokens, which resulted in just 796 petaFLOP/s-days of computations[17]. Similarly, Alibaba's PLUG initially planned to utilize 128 A100s for training over the course of 120 days but later reduced training time to just 35 days, or an estimated 427 petaFLOP/s-days[16].

Third, Nvidia's V100 and A100 GPUs remain the most popular GPUs for training Chinese large-scale models. In fact, only 3 out of 26 models explicitly mention that they were trained without Nvidia GPUs. Huawei used its own Ascend 910 processors to train its PanGu-α model[17]. Huawei Ascend 910 NPUs were also used by the state-sponsored Peng Cheng Lab to assist in the training of ERNIE 3.0 Titan[26]. More recently, in April 2022, an alliance of researchers from Tsinghua University, BAAI, Alibaba DAMO Academy, and Zhejiang University trained the world's first "brain-scale" model with 174 trillion parameters on China's domestically produced New Generation Sunway Supercomputer[27]. Though reaching exascale, the computer relies on rather dated 14nm processors.

Data

Unsurprisingly, the majority of the Chinese AI models are trained on Chinese-language data sources, and the development of these models encouraged efforts in constructing Chinese-language datasets and building the data infrastructure for pre-trained models. Most notably, the BAAI team released WuDaoCorpora, a bilingual dataset with 2.3TB of cleaned Chinese data and 300GB of cleaned English data from encyclopedias, novels, news, and scientific literature[37]. For multimodal models, the M6 developers from Tsinghua University and Alibaba released the M6-Corpus, the first large-scale multimodal dataset in Chinese[14]. That dataset contains over 1.9TB of images and 292GB of text and has been used in developing later models, such as WuDao-Wen Yuan 2.0 and BaGuaLu.

Role of Government

Three government-sponsored entities and research labs have played a significant role in China's recent AI development: the Beijing-based BAAI, the Hangzhou-based Zhejiang Lab, and Shenzhen's Peng Cheng Lab. These labs bring together resources, talents, and financial resources from government, industry, and academia—often funding or directly building many of the models. This marks a shift from the Chinese government's traditional approach to research financing, which usually involves issuing research grants through the National Natural Science Foundation of China (NSFC). The new "big science" approach might be more suited for the development of large-scale AI models, which are often very compute-intensive and expensive to build.

Beijing Academy of Artificial Intelligence

Among the three state-sponsored labs, BAAI is perhaps the most prominent actor in China's advanced AI landscape and appears to be China's most ambitious initiative explicitly aimed at developing artificial general intelligence. It has released a series of breakthrough models including the WuDao 1.0 and WuDao 2.0 models, CogView, and BaGuaLu. The institute, established in November 2018 in Beijing, is a nonprofit sponsored by the Chinese Ministry of Science and Technology and the Municipal Party Committee and the Government of Beijing[38]. Bringing together top scholars from Tsinghua University, Peking University, and the Chinese Academy of Sciences (CAS), as well as top leaders from industry giants such as Baidu, Xiaomi, and ByteDance, BAAI's unique institutional setup allows researchers to conduct fundamental AI research and translate their research to industry applications.

However, the institute suffered from a major plagiarism scandal in April 2022, where its 200-page literature review titled "A Roadmap for Big Model" was found to have more than ten plagiarized sections[39]. The literature review, which has nearly one hundred coauthors, was presumptively published as a response to the large-scale literature review led by Stanford scholars titled "On the Opportunities and Risks of Foundation Models" and published in August 2021[1]. Journalists suspect that the plagiarism occurred because lab managers and senior researchers assigned their graduate students to write the literature review and gave them as little as a week to write an entire section[40]. Like many other institutions in China and the West, BAAI faces significant pressure to compete with top labs on publications, yet the incident shows that BAAI might face challenges in coordinating large numbers of collaborators and ensuring the quality of its research output.

Zhejiang Lab

Similar to BAAI's institutional setup, Zhejiang Lab was established jointly in September 2017 by the Zhejiang Provincial Government, Zhejiang University, and Alibaba Group. The lab is located in the West Hangzhou Scientific Innovation Corridor and focuses on fundamental AI research. It is currently constructing the Zhejiang Provincial Laboratory for Intelligent Science and Technology, another government-led research lab. Zhejiang Lab was involved in the development of the "brain-scale" BaGuaLu model, which was trained on an exascale supercomputer at the National Supercomputing Center in Wuxi.

Peng Cheng Lab

Lastly, Peng Cheng Lab is another notable AI laboratory sponsored by a provincial-level government. After a series of talks by Chinese President Xi Jinping on the importance of technology to Guangdong Province and Shenzhen, Peng Cheng Lab was established in March 2018 as part of China's "Greater Bay Area" development plan and technological innovation strategy. The lab focuses on collaborating with top universities in China, Hong Kong, Macau, and Singapore. Peng Cheng Lab provided computing support for the training of ERNIE 3.0 Titan, the largest Chinese dense model at the time of its release in December 2021.

Government, Ethics, and Access

Governance-related issues constitute another important area in our research. We find that Chinese researchers do address ethics or governance issues related to large-scale models, which departs from the conventional accounts that these types of discussions are absent in Chinese AI organizations[41-44]. 12 of the 26 papers or official announcements associated with the models discussed some ethics or governance issues. While the rhetoric in publications should not be taken at face value, they indicate a growing awareness of these issues in China's academic community.

Three types of issues receive the most attention in our sample of Chinese AI papers: bias and fairness, misuse risks, and environmental harms. Likely in response to misuse risks, some Chinese organizations have also placed access restrictions on their models.

Discussions of Bias and Fairness

The first issue raised by papers in our sample is bias and fairness. The Yuan 1.0 paper, for example, includes extended examples of how the model could be steered to produce discriminatory content[24]. When prompted with opinions about the endurance of traditional patriarchy, Yuan 1.0 outputs: "I believe that there is a direct connection between women's social status and their fertility. Female fertility is the basis of social status." The authors acknowledge that even an unbiased model can be easily steered by humans because the model is trained to generate content based on the style of the human writer. The WenLan 2.0 paper highlights the fact that AI models are likely to learn prejudices and human biases from their data inputs[25].

Discussions of Misuse

The second issue raised by papers in our sample is misuse, particularly the use of pre-trained models to generate fake news and other harmful content. This concern is expressed in the WenLan 2.0 and M6-10T papers, which warn that models can become increasingly competent at fake news generation and manipulation[23,25]. The M6-10T authors add that they have put much effort into removing harmful input data, such as hate speech, terrorism, and pornography, but admit that such content often cannot be eliminated entirely. Thus, they recommend limiting access for models that were not trained on commonly used public datasets to avoid misuse risks.

Discussions of Environmental Harms

Finally, several papers in our sample underscore the environmental concerns of training language models. The Yuan 1.0 paper mentions the high energy costs of training GPT-3-like models, and the authors altered their model architecture and training process to accelerate training and reduce emissions[24]. Another model, the 260-billion-parameter ERNIE 3.0 Titan, uses an online distillation framework to reduce computation overhead and lower carbon emissions[26]. Alibaba's M6-10T proposes a "pseudo-to-real" training strategy that allows for "[f]ast training of extreme-scale models on a decent amount of resources," which it claims can reduce the model's carbon footprint[23].

Access Restrictions

Beyond discussions of governance issues in publications on large-scale AI models, Chinese organizations have developed APIs and measures to govern access to these models. To get a better sense of these access restrictions, we applied for use of Inspur's Yuan 1.0, BAAI's WuDao platform and GLM-130B, and Baidu's Wenxin platform (which includes the ERNIE and PLATO series). These were the only four publicly accessible models at the time of our writing (January 2023). The models come with user agreements which include language standard in large-scale pre-trained models' user agreements, such as the developers not being responsible for the content generated through the API and prohibitions against generating harmful content[45-48]. Yuan 1.0's API and BAAI's GLM-130B incorporate a relatively stringent screening process, requiring prospective users to submit an application where they state their backgrounds and intended use of the

models. As of this writing, BAAI's WuDao platform and Baidu's Wenxin platform do not have similar screening processes.

Technonationalism

Another current running through Chinese discourse about language models is technonationalism, which links technological achievements in these models to China's national capabilities. Technonationalist concerns materialize in three ways: in preferences for using domestic technology to produce models, in fears about access restrictions to Western models, and in pride surrounding SOTA models.

Preference for Domestic Chips and Software

First, China's dependence on foreign AI frameworks and computing hardware has prompted developers to strive to use domestically produced software and hardware to train their models. Commercial developers often leverage technonationalism as a way to advertise their brands and products. For instance, Baidu researchers emphasize that they trained ERNIE 3.0 Titan on Baidu's PaddlePaddle framework and Huawei's Ascend chips[26]. Likewise, when describing the development of the PanGu model, which used Huawei's Ascend 910 chips and self-developed MindSpore framework, Huawei researchers highlight that "the model is trained based on the domestic full-stack software and hardware ecosystem."[17]

Concerns About Access to Western Models

Second, discussions about access restrictions to Western models exhibit technonationalist tendencies. In reference to restrictions on API access to GPT-3 in China, one BAAI expert warned in an official video that if language models become an important strategic resource, API restrictions could be a form of "technological blockade."[49] He also emphasized that, in the age of large-scale models, the development and training of such models is akin to "an arms race in AI" and that, as a leader in China's AI landscape, BAAI has a "duty" to release language models "under China's discourse leadership."

Frontier Models as a Source of National Pride

Third, comparisons between Chinese and Western models often touch on national pride. Papers associated with Chinese models often use prominent Western models as benchmarks. Official press releases, media coverage, and conference presentations of such models also emphasize the superiority of Chinese models over their Western counterparts. For example, at the 2021 BAAI Annual Conference, BAAI developers directly compared their WuDao 2.0 series against Western models such as OpenAI's CLIP, GPT-3, and DALL-E; Microsoft's Turing-NLG; and Google's ALIGN, claiming that WuDao 2.0 set the new SOTA on nine international benchmarks, which include tasks such as image labeling, image generation, and few-shot language learning[50].

Conclusion

In this paper, we have distilled key trends in China's development of large-scale pre-trained models, based on a study of 26 such projects. We believe this is one of the first efforts to systematically analyze Chinese large-scale AI models. We described the key features of these models, including their performance benchmarks, funding sources, compute capabilities, and model access policies. In addition, we analyzed two aspects of Chinese discourse around these models—AI ethics and technonationalism—that connect to broader themes in China's development of AI. In doing so, we dispelled certain misconceptions about China's AI development, such as the absence of discussions about ethics, and shed light on China's unique government-industry-academia alliances, which play a key role in the country's advanced AI landscape.

Our research offers insight into the key Chinese developers, China's advanced AI capabilities, and the country's AI ecosystem. While the landscape of large language models is still evolving at a fast pace, our findings provide a solid foundation for future research. Going forward, to understand how China's AI ecosystem is changing, continued attention to the development of large-scale models will be essential.

References and Notes

Our definition of a large-scale pre-trained model is similar to the concept of a "foundation model," which refers to "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." See https://www.infoq.cn/article/EFIHo75sQsVqLvFTruKE#alibaba

This model was trained on both Nvidia V100 GPU and Ascend 910 NPU clusters. We thank Lennart Heim for his help with the compute estimates. https://www.leiphone.com/category/academic/vyasXlXWc07hajex.html

For example, commenting on one of the multi-modal models covered in our study, Jack Clark, former Policy Director of OpenAI, once stated, "There's no discussion of bias in the paper (or ethics), which isn't typical for papers of this type but is typical of papers that come out of Chinese research organizations." https://resource.wudaoai.cn/resources/agreement/wudao_model_agreement_20210526.pdf A copy of this agreement is available, upon request to authors.

1. See Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

2. Brown, T. *et al.* Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020).

3. Hoffmann, J. *et al.* Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

4. Chowdhery, A. *et al.* PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

5. Wu, F. *et al.* Towards a new generation of artificial intelligence in China. *Nature Machine Intelligence* **2**, 312–316 (2020).

6. Roberts, H. *et al.* The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society* **36**, 59–77 (2021).

7. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

8. Sun, Y. *et al.* ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137* (2021).

9. Du, Z. *et al.* GLM: General language model pretraining with autoregressive blank infilling. in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) 320–335 (2022).

10. Wu, C. *et al.* NÜWA: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417* (2021).

11. Zhang, H. *et al.* ERNIE-ViLG: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283* (2021).

12. Ding, M. *et al.* CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* **34**, 19822–19835 (2021).

13. Zhang, Z. *et al.* CPM: A large-scale generative Chinese pre-trained language model. *AI Open* **2**, 93–99 (2021).

14. Lin, J. *et al.* M6: A Chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823* (2021).

15. Huo, Y. *et al.* WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561* (2021).

16. Yuying, Z. [赵钰莹]. Alibaba released PLUG: 27 billion parameters, the largest pre-trained language model in the Chinese community [阿里发布 PLUG：270 亿参数，中文社区最大规模预训练语言模型]. *InfoQ* https://www.infoq.cn/article/EFlHo75sQsVqLvFTruKE#alibaba (2021).

17. Zeng, W. *et al.* PanGu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369* (2021).

18. Yan, Y. *et al.* ConSERT: A contrastive framework for self-supervised sentence representation transfer. arXiv preprint *arXiv:2105.11741* (2021).

19. Yang, A. *et al.* M6-T: Exploring sparse expert models and beyond. *arXiv preprint arXiv:2105.15082* (2021).

20. Zhang, Z. *et al.* CPM-2: Large-scale cost-effective pre-trained language models. *AI Open* **2**, 216–224 (2021).

21. Bao, S. *et al.* PLATO-XL: Exploring the large-scale pre-training of dialogue generation. arXiv *preprint arXiv:2109.09519* (2021).

22. Zidong Taichu multimodal large model [紫东太初多模态大模型]. *CASIA* https://gitee.com/zidongtaichu/multi-modal-models.

23. Lin, J. *et al.* M6-10T: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining. *arXiv preprint arXiv:2110.03888* (2021).

24. Wu, S. *et al.* Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arXiv preprint arXiv:2110.04725* (2021).

25. Fei, N. *et al.* Towards artificial general intelligence via a multimodal foundation model. *arXiv preprint arXiv:2110.14378* (2021).

26. Wang, S. *et al.* ERNIE 3.0 Titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2112.12731* (2021).

27. Ma, Z. *et al.* BaGuaLu: Targeting brain scale pretrained models with over 37 million cores. in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* 192–204 (2022).

28. Hong, W., Ding, M., Zheng, W., Liu, X. & Tang, J. CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).

29. GLM-130B: An open bilingual pre-trained model. *Tsinghua University-KEG* https://keg.cs.tsinghua.edu.cn/glm-130b/posts/glm-130b/ (2022).

30. IDEA-CCNL. Taiyi-Stable-Diffusion-1B-Chinese-v0.1. *Hugging Face* https://huggingface.co/IDEA-CCNL/Taiyi-Stable-Diffusion-1B-Chinese-v0.1 (2022).

31. Chen, Z. *et al.* AltCLIP: Altering the language encoder in CLIP for extended language capabilities. *arXiv preprint arXiv:2211.06679* (2022).

32. Pan, X., Qin, P., Li, Y., Xue, H. & Chen, W. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950* (2022).

33. BAAI. ALM 1.0. *GitHub* https://github.com/FlagAI-Open/FlagAI/blob/master/examples/ALM/README_zh.md (2022).

34. Zhavoronkov, A. Wu Dao 2.0 - Bigger, stronger, faster AI from China. *Forbes* (2021).

35. BAAI conference opened, and the world's largest intelligent model 'WuDao 2.0' was released [智源大会开幕，全球最大智能模型"悟道2.0"发布]. *BAAI* https://mp.weixin.qq.com/s/NJYINRt_uoKAlgxjNyu4Bw (2021).

36. Sevilla, J. *et al.* Compute trends across three eras of machine learning. *arXiv preprint arXiv:2202.05924* (2022).

37. Yuan, S. *et al.* WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. *AI Open* **2**, 65–68 (2021).

38. Beijing Academy of Artificial Intelligence. *BAAI* https://www.baai.ac.cn/english.html#About.

39. Yuan, S. *et al.* A roadmap for big model. *arXiv preprint arXiv:2203.14101* (2022).

40. Literature review by 100 Chinese scholars found guilty of plagiarism, BAAI Issues Statement Acknowledging Wrongdoing and Transfering Documents to Third-Party Experts for Investigation [100位中国学者合作的研究综述被曝抄袭，智源发表声明：承认错误，转交第三方专家调查]. *Leiphone* https://www.leiphone.com/category/academic/vyasXlXWc07hajex.html (2022).

41. Clark, J. Import AI 239: China trains a massive 10b model, Vicarious does pick&place; the GCHQ publishes some of its thoughts on AI. *Import AI* https://jack-clark.net/2021/03/08/import-ai-239-china-trains-a-massive-10b-model-vicarious-does-pick-place-the-gchq-publishes-some-of-its-thoughts-on-ai/ (2021).

42. Allison, G. & Schmidt, E. *Is China beating the U.S. to AI supremacy?* https://www.belfercenter.org/publication/china-beating-us-ai-supremacy (2020).

43. Colaner, S. Michael Kanaan: The U.S. needs an AI 'Sputnik moment' to compete with China and Russia. *Venture Beat* (2020).

44. Sharma, Y. Robots bring Asia into the AI research ethics debate. *University World News* (2017).

45. BAAI WuDao platform model user agreement [【智源悟道平台】模型使用协议]. *BAAI* https://resource.wudaoai.cn/resources/agreement/wudao_model_agreement_20210526.pdf.

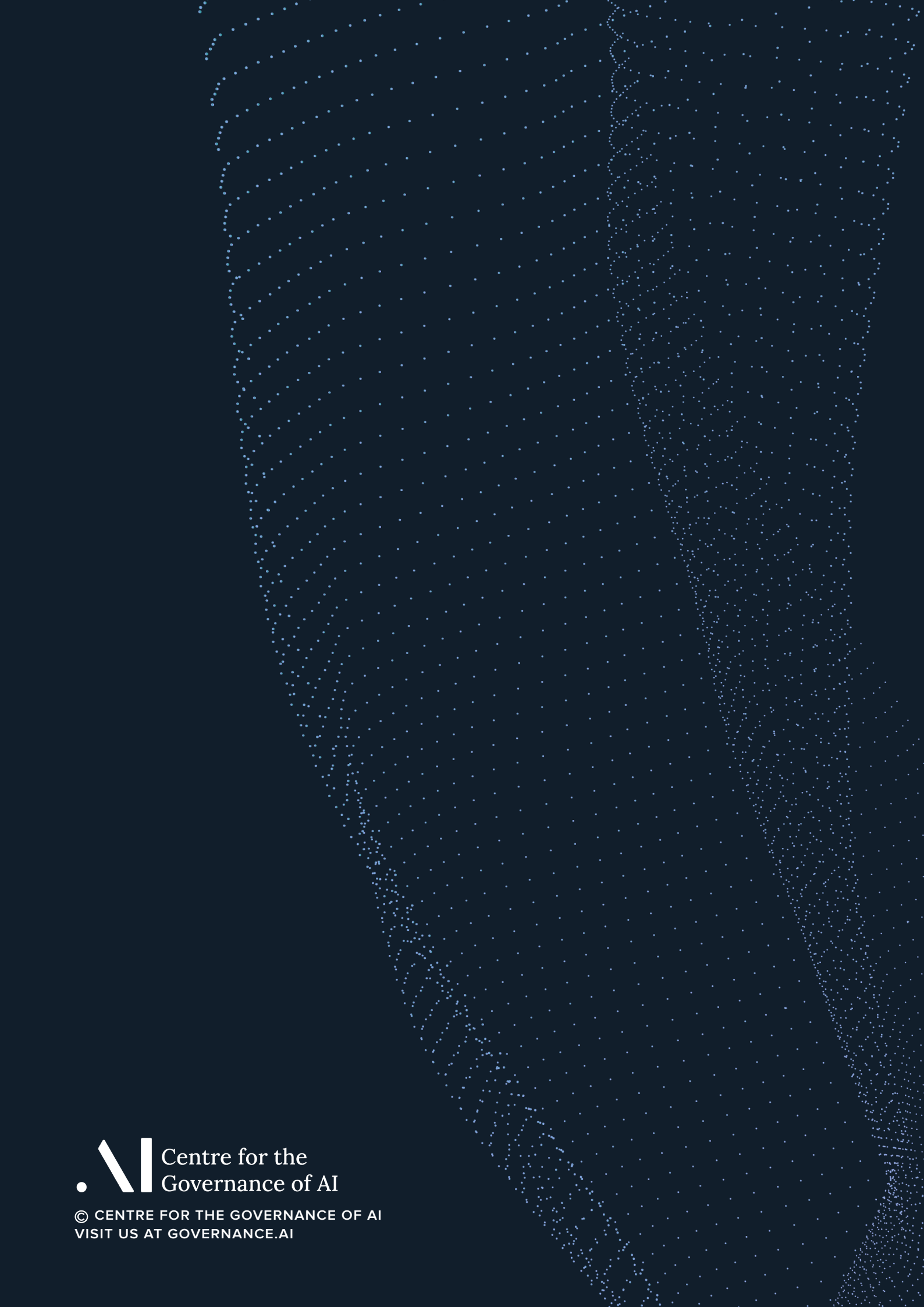
46. GLM-130B application form. https://docs.google.com/forms/d/e/1FAIpQLSehr5Dh_i3TwACmFFi8QEgIVNYGmSPwV0GueIc-sUev0NEfUug/viewform.

47. Terms of service [服务条款]. *Paddle Paddle Wenxin Large Model* https://wenxin.baidu.com/wenxin/docs#Yl6th25am (2022).

48. Yuan 1.0 API user manual ["源1.0"API调用使用手册]. *Inspur* https://air.inspur.com/user-doc.

49. The summit of WuDao open course series [悟道之巅系列公开课]. https://www.bilibili.com/video/BV1R44y1e76L?p=1 (2021).

50. Announcement of WuDao 2.0 achievements and partnership signing ceremony [悟道2.0成果发布及签约]. *BAAI Community* [智源社区] https://hub.baai.ac.cn/view/8578 (2021).



Centre for the Governance of AI

VISIT US AT GOVERNANCE.AI