Registration Form for Representative Achievements

Applicant: Yanghua Xiao **College/School: College of Computer Science and Artificial Intelligence**

Research Category: [ ] Basic Innovation [√] Applied Innovation

Achievement Title:

Theory and Methods of Data Science for Large Models

List of Supporting Materials for the Achievement (itemize the high-level papers, monographs, authorized patents, policy advisory reports, etc. that constitute this achievement):

I. High-Level Papers

1. Nianqi Li, Siyu Yuan, Jiangjie Chen, Jiaqing Liang, Feng Wei, Zujie Liang, Deqing Yang, **Yanghua Xiao** (corresponding author). Past Meets Present: Creating Historical Analogy with Large Language Models. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025): 3942-3957 (**ACL 2025 Outstanding Paper Award**)

2. Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, **Yanghua Xiao** (corresponding author). Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024): 18099-18107 (this work was rated by AINLP as one of the most valuable benchmarks for evaluating Chinese large models)

3. Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, **Yanghua Xiao** (corresponding author). InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): 1840-1873 (this work is one of the most highly cited benchmarks in the area of personality evaluation for role-playing agents)

II. Authorized Patents

1. Method and device for constructing a multi-constraint fused six-tuple dataset for government-affairs Text2SQL, Patent No.: 2026106445834

2. Method and apparatus for generating knowledge-rich questions for the robustness of question answering systems, Patent No.: 2022112396663

3. Method for detecting the reliability of answers generated by large-scale language models, Patent No.: 2023111954525

III. Policy Advisory Reports

1. Generative Language Models and Artificial General Intelligence: Connotation, Paths, and Implications, People's Tribune - Academic Frontiers, July 2023 (Part 2)

2. New Trends in the Development of Artificial Intelligence Large Models and Reflections Thereon, People's Tribune - Academic Frontiers, 2024, No. 13

3. Data Bottlenecks in AI Development and Paths to Breakthrough, People's Tribune - Academic Frontiers, 2026, No. 11

Brief Description of the Achievement (200-300 characters): Data science for large models is an important foundation for the development of artificial general intelligence, and the development of large models faces three challenges: knowledge is difficult to evaluate, reasoning is prone to distortion, and personalized interaction is

difficult to measure. To address these issues, the project proposed a historical event analogy reasoning method that integrates fact verification and self-reflection, suppressing reasoning hallucinations and stereotypical tendencies, and won the ACL 2025 Outstanding Paper Award; built Xiezhi, an all-discipline, continuously updated knowledge evaluation benchmark for large models, which has been recognized by a number of institutions and used in the evaluation of Chinese large models; and proposed InCharacter, a personality fidelity evaluation benchmark for role-playing agents based on psychological interviews, which has attracted industry attention and has been cited and reviewed in a paper in a Nature portfolio journal. In response to industry needs, the project put forward policy advisory recommendations for the development of large models and artificial general intelligence, established a system of training data and core capability labels, and formed a data curriculum optimization scheme that won the Huawei Spark Award. The related technologies have been used in the training of multiple foundation models, including Huawei Pangu and Alibaba Qwen, improving data governance and model capability development, and have been applied to fields such as the overseas expansion of large models and shipping-industry large models, promoting high-quality development of the industry.

Notes:

I. Scientific Contributions

The development of large models is confronted with the practical bottleneck of insufficient supply of high-quality data and low efficiency of data utilization, which in turn constrains the iterative improvement of model capabilities; merely expanding data scale and relying on empirical trial and error makes it difficult to achieve efficient training and higher-order intelligence. Developing the data science of large models aims to reveal the laws by which data-driven intelligence is formed, to move model training from empirical engineering toward a quantifiable, interpretable, and controllable scientific system, and to lay the foundation for the industrialization of large models. At present, research on large models still faces three fundamental problems. First, the knowledge covered by models is broad and evolves rapidly, and traditional static benchmarks can hardly reveal their cross-disciplinary knowledge structure and capability boundaries in a comprehensive way. Second, although models have accumulated massive amounts of historical knowledge, they struggle to break through surface-level similarity and reliably transfer historical experience to real-world problems. Third, role-playing agents can imitate a character's language and experiences, but whether their deeper personality remains consistent with the target role still lacks effective measurement. Addressing these problems, this project conducts research along three directions: evaluation of domain knowledge, historical analogy reasoning, and personalized intelligent interaction, yielding scientific findings on the following three themes:

1. A method for the dynamic evaluation and boundary characterization of the full-domain knowledge capabilities of large models was established. To address the limited coverage of traditional evaluation, the tendency of data to become outdated, and the difficulty of distinguishing differences in model capabilities, a continuously updated full-domain knowledge evaluation system, Xiezhi, was constructed, revealing the intrinsic relationship between the growth of model scale and

knowledge capabilities across domains.

2. A historical analogy reasoning mechanism for the cross-temporal transfer of knowledge in large models was established. To address the problem that large models “possess historical knowledge but cannot effectively draw on history,” the project proposed a historical analogy acquisition task, systematically evaluated the analogy discovery capability of large models, and improved their ability to transfer historical knowledge.

3. Methods for measuring personality consistency and evaluating trustworthy interaction in the role intelligence of large models were established. To address the problem that the evaluation of role-playing agents remains at the level of knowledge reproduction and language imitation and can hardly reach inner cognition and behavioral tendencies, the project proposed a role personality evaluation method based on psychological interviews, advancing role intelligence research from “whether the language sounds alike” to the deeper evaluation of “consistent personality and trustworthy behavior.”

These related contributions have resulted in 110 papers published in Class A journals or conferences recognized by the China Computer Federation (CCF), 11 authorized patents, and 3 policy advisory papers. The results have repeatedly been presented as keynote speeches at the invitation of industry. The course-based learning scheme for large models won the Huawei Spark Award, and the analogy reasoning evaluation research received the ACL 2025 Outstanding Paper Award (see Figure 1). The first Advanced Technology Research Center within Huawei's global operations was approved, focusing on exploratory natural research in the direction of data science for large models. With support from Huawei's openPangu program, the applicant was the first internationally to offer the course Large Model Data Engineering, and, as the person in charge, undertook the construction of the Ministry of Education Virtual Teaching and Research Section for openPangu Models and Applications. The results have been applied to the training of Huawei's Pangu foundation models and to the construction of multimodal training data for Alibaba's Qwen models, applied to the training of the shipping and trade large model at Shanghai Yitong International, and have supported the cross-cultural and cross-lingual corpus construction of the Huawei Translation Center as well as the construction of professional corpora for business departments such as Huawei Data Communication.

II. Evaluation of Representative Publications

The work on analogy reasoning for historical events proposed by the applicant won the ACL 2025 Outstanding Paper Award (Paper 1). The organizing committee held that the work

“innovatively proposed a self-reflection mechanism to suppress hallucinations and stereotypical tendencies in the reasoning process of large models.” Nitesh V. Chawla, Fellow of ACM, IEEE, AAAS, and AAAI, has cited this paper many times and devoted a separate paragraph to introducing its work, highly recognizing the self-reflection framework proposed in the paper and the strong consistency between the automatic multi-dimensional similarity metric and human evaluations (“Their automatic multi-dimensional similarity metric demonstrates high alignment with human evaluations”).

The applicant constructed Xiezhi (Paper 2), an all-discipline language model evaluation benchmark dataset containing 516 domain dimensions and 250,000 pieces of domain evaluation data, and also proposed a one-out-of-fifty testing mechanism, which verifies relatively well whether a language model truly masters knowledge (see Figure 2). This evaluation dataset was rated by AINLP as one of the most valuable benchmarks for evaluating Chinese large models; it is a representative of “subject knowledge evaluation” alongside C-Eval and has been indexed by platforms such as OpenDataLab. In a paper, Xing Xie, Fellow of ACM and IEEE, held that Xiezhi can effectively evaluate the knowledge level of large language models in different subject areas, and that evaluation through Xiezhi enables researchers to recognize the inherent limitations of these models and deepen their understanding of their capabilities in diverse fields (“Xiezhi enables researchers to comprehend the notable limitations inherent in these models and facilitates a deeper comprehension of their capabilities in diverse fields”).

InCharacter (Paper 3), the benchmark for evaluating the personality fidelity of large-model role-playing constructed by the applicant, is one of the most cited benchmarks in the direction of personality evaluation for role-playing agents (about 240 citations on Semantic Scholar). In an article published in Nature Machine Intelligence, Maja J. Matarić, Fellow of IEEE, ACM, and AAAS and Member of the US National Academy of Engineering, cited this paper alongside CharacterEval as a representative “psychometric framework for role personality.” In their paper, the related researchers held that InCharacter evaluates role personality fidelity through interview-style probing grounded in psychological traits, providing a rigorous psychometric basis for measuring whether a role-playing agent truly “becomes” the role it plays. (InCharacter evaluates personality fidelity through interview-style probing grounded in psychological traits.)

The applicant was the first internationally to offer the course Large Model Data Engineering, conducting systematic research on key steps such as training data governance, cognitive capability evaluation, and domain model optimization, and forming a large-model data engineering system that covers the whole process of corpus grading and classification, data synthesis, mix-ratio

optimization, quality assessment, and capability evaluation. The course is funded by Huawei's openPangu special fund and Fudan University's AI-BEST curriculum reform fund, and was approved as a Ministry of Education “Intelligent Foundations” virtual teaching and research section (openPangu Models and Applications) (see Figure 3).

All three policy advisory articles were published in the “Research on Twenty-Four Major Issues” column of People's Tribune - Academic Frontiers, a high-end academic journal supervised by People's Daily and sponsored by the People's Tribune Magazine. The journal is a source journal of the Chinese Social Sciences Citation Index (CSSCI), a national Chinese core journal, and a core journal in the A-list of the AMI Comprehensive Evaluation of Chinese Humanities and Social Sciences Journals, and is also a designated publication for releasing National Social Science Fund results (see Figure 4).

III. Application and Impact of the Research Results

In terms of constructing and optimizing domain knowledge datasets, the applicant proposed an approach to building high-quality datasets. Centering on improving the cognitive capabilities of large models and proceeding from perspectives such as basic evaluation, reasoning cognition, emotional cognition, and cultural cognition, more than 40 large-scale, high-quality datasets with more reasonable data distributions were constructed, promoting theoretical and methodological research in academia and the selection of algorithm applications in industry. The benchmark research work for travel planning, “TravelPlanner: A Benchmark for Real-World Planning with Language Agents,” was reposted on social networks by Turing Award laureate Yann LeCun (more than 100,000 reposts) and was selected as an ICML 2024 highlight paper. Another work on large language models and human cognition, “Why Did Apple Fall: Evaluating Curiosity in Large Language Models,” went viral on the overseas social platform X and was proactively reposted and recommended by leading overseas AI technology news platforms, generating enormous influence (see Figure 5).

To address the core difficulties of large-model training data, namely being miscellaneous, hard to manage, chaotically mixed, and inefficient to train on, the applicant, proceeding from theory, proposed a set of ideas for building a system of training data and core capability labels as well as a data curriculum optimization scheme for large-model data governance, and won the Huawei Spark Award (see Figure 6). In addition, the large-scale multimodal data science classification system and the scientific evaluation system for concept understanding that were developed have been applied to the training of Alibaba's large models on multimodal tasks, achieving notable results in practical

application and providing Alibaba with efficient support for multimodal training corpus data governance (see Figure 7); the shipping and trade domain large model that was trained has effectively empowered shipping and trade data products and achieved good results on tasks such as HS code classification (see Figure 8).

The applicant received funding for three consecutive research and development projects from the Huawei Translation Center (see Figure 9) and established a complete research system of evaluation and data augmentation oriented to the social cognition and cross-cultural interaction of large models. The research has successively constructed several authoritative datasets: a cultural competence evaluation dataset based on cognitive psychology (AAAI 2026); cross-cultural safety evaluation datasets, including the CulShield cultural safety benchmark covering 77 countries and regions with more than 2,020 taboo samples, and a dataset containing 16 languages and 36,800 expert-refined samples (ACL 2026); and a dataset for evaluating the cross-cultural situational competence of large models (EMNLP 2026). The related results supported Huawei in building GaoYao (ACL 2026), currently the largest cross-cultural and cross-lingual evaluation dataset, which covers 26 languages and 51 countries and regions with 182,000 samples in total. The above research effectively alleviates pain points of large models such as the disconnect between cultural cognition and actual behavior, uneven cross-lingual capabilities, and insufficient localization of translation data; through technical paths such as model capability diagnosis, data optimization, and filtering, it has improved the multilingual and cross-cultural adaptability of large models, providing an evaluation basis and data foundation for the safe and reliable global expansion of domestically developed large models.

The applicant was the first in the world to propose a “students test AI” method for constructing training data. Based on the idea of “reverse evaluation,” students set the questions to examine large language models, with a scoring mechanism positively correlated with the model failure rate. Practice shows that most students can effectively identify the reasoning blind spots of models, and some test cases achieved the breakthrough of models scoring zero. This course assessment reform focuses on guiding students to develop a critical understanding of the capability boundaries of AI systems and on shifting the evaluation paradigm from the assessment of knowledge memorization to the assessment of human-machine collaborative decision-making capability. The “students test AI” evaluation model has attracted widespread attention at home and abroad: official WeChat public accounts such as People's Daily, Xinhua News Agency, Global Times, and China Youth Daily all published reports with more than 100,000 views, and the German newspaper Frankfurter Post also published a special commentary, making it a hot topic of public discussion (see Figure 10).

Ute Schmid, an expert in cognitive systems at the University of Bamberg, commented on the reform: “It not only helps students understand the limitations of AI models, but also helps cultivate students' ability to critically judge the quality of content.” At present, this teaching reform has received support from Alibaba, and a “students test AI” competition covering 100 courses is being organized to collect high-quality human-versus-machine test questions at scale.



Figure 1. Knowledge-enhanced analogy reasoning model won the ACL 2025 Outstanding Paper Award



Figure 2. All-discipline language model evaluation benchmark Xiezhi



Figure 3. The openPangu Models and Applications virtual teaching and research section was selected as one of the fourth batch of pilot virtual teaching and research sections announced by the Department of Higher Education of the Ministry of Education



Figure 4. People's Tribune - Academic Frontiers published 3 influential policy advisory articles

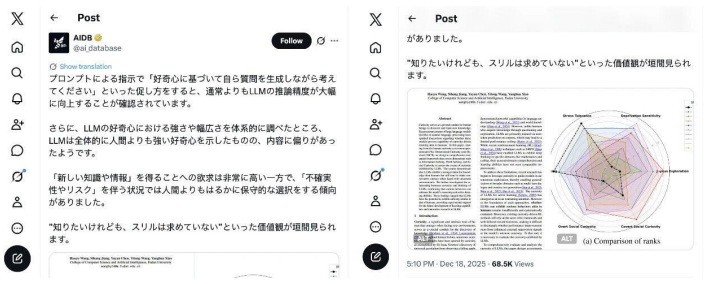
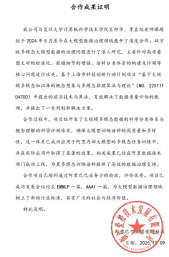
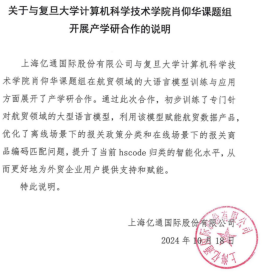
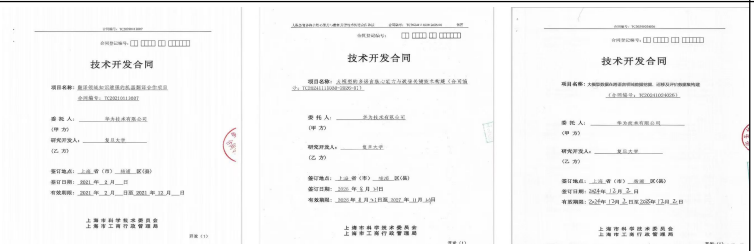
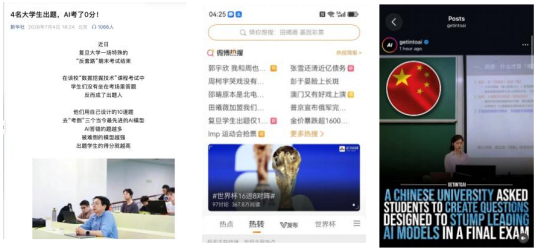


Figure 5. A special report by the overseas technology



Figure 6. The building of the training

media outlet AIDB on the paper “Why Did Apple Fall: Evaluating Curiosity in Large Language Models,” with 60,000 views	data and core capability label system and the data curriculum optimization scheme won the Huawei Spark Award
	
Figure 7. Proof of collaborative results with Alibaba: large-scale multimodal data science classification system and scientific evaluation system for concept understanding	Figure 8. Proof of collaboration with Shanghai Yitong International: training and application of a large language model in the shipping and trade domain
	
Figure 9. Three consecutive research and development project grants from the Huawei Translation Center: social cognition and cross-cultural interaction of large models	Figure 10. The “students test AI” teaching reform in the data mining course: trending on Weibo and reported by multiple authoritative media at home and abroad