CONCORDIA AI
安远AI

info@concordia-ai.com
www.concordia-ai.com

China's AI Safety Evaluations Ecosystem

As concerns over the safety of frontier AI systems have grown, governments, developers and scholars around the world have increasingly focused on developing mechanisms to evaluate AI systems for dangerous risks and societal impacts. AI safety evaluations provide an early warning that models may possess excessively dangerous capabilities. While evaluations fall short of providing complete safety assurance, they are an important tool for risk mitigation. As the science of AI safety evaluations is still nascent, the global community has a stake in improving scientific rigor and sharing best practices so that countries around the world can institute appropriate and sufficiently robust evaluation measures. China's Global AI Governance Initiative called for "a testing and assessment system based on AI risk levels" and the Bletchley Declaration articulated international support for safety testing and evaluation of frontier AI systems.

China already possesses advanced AI capabilities and substantial AI evaluations projects, so it has an important role to play in these conversations. We believe that this report and our new Chinese AI Safety Evaluations Database provide the first comprehensive analysis of these evaluations in English. We hope to facilitate mutual learning and engagement on AI safety evaluation best practices among leading Chinese and international institutions. We welcome engagement and outreach with other organizations interested in fostering internationally interoperable AI safety evaluation practices and standards.

In this paper, we first describe requirements around AI safety evaluations in Chinese AI governance. Next, we share our methodology for creating the Chinese AI Safety Evaluations Database. The database covers a range of safety and societal risks from advanced AI systems, but our analysis below focuses primarily on "frontier AI risks" given the greater need and potential for international cooperation on transnational and catastrophic threats. Then, we describe notable trends from the database, including which risks were mainly tested for, the type or methodology of evaluations, languages used, and modality. Lastly, we provide detailed descriptions of key government-supported, academic, and private research groups for AI safety evaluations.

Key takeaways:

- The Chinese government currently requires developers to conduct pre-deployment testing and evaluation of their AI systems for ideological orientation, discrimination, commercial violations,

violations of individual rights, and application in higher risk domains. There are signs that this could expand in the future to incorporate testing for frontier or catastrophic AI safety risks.

- The risk areas that received the most testing by Chinese AI safety benchmarks are bias, privacy, robustness to adversarial and jailbreaking attacks, machine ethics, and misuse for cyberattacks.
- Chinese evaluations tested for all categories defined as frontier AI risks, with misuse for cyberattacks as the most tested frontier risk.
- Chinese AI safety evaluations primarily comprise static benchmarks, with a small number of open-source evaluation toolkits, agent evaluations, and domain red teaming efforts. Chinese institutions do not appear to have conducted human uplift evaluations.
- Shanghai AI Lab, Tianjin University NLP Lab, and Microsoft Research Asia Societal AI team are the only research groups that have published two or more frontier AI safety evaluations in China. However, many other government-backed, academic, and private industry research groups have also published evaluations covering a broad spectrum of AI safety and social impacts concerns.

Chinese policy requirements for AI safety evaluations

China's July 2023 Interim Measures for the Management of Generative Artificial Intelligence Services require providers of generative AI service that can affect public opinion to file their service with regulators and undergo safety/security self-assessments, which in turn require government approval before being publicly rolled out.[1] Similar requirements were already in place for recommendation algorithms and deepfakes respectively in 2021 and 2022.

Chinese "self-assessment" evaluations for generative AI largely follow a technical document (Ch, En) issued in February 2024 by TC260, one of China's official technology standardization committees. The document suggests that pre-deployment tests cover 31 types of safety risks, split up into five main categories:

1. Violating socialist core values, such as content endangering national security, harming the image of the state, promoting terrorism, or false information.
2. Discriminatory content, such as discrimination based on ethnicity, gender, beliefs, nationality, etc.
3. Commercial violations, such as IP infringement or violation of business ethics.

[1] The Chinese word 安全 can be translated both as safety and security depending on the context. In ambiguous cases, we translate it as safety/security. See our State of AI Safety in China report pages 4-5 for a full description.

info@concordia-ai.com
www.concordia-ai.com

4. Violating rights and interests of others, such as privacy violations, defamation, and endangering others' health.
5. Generation of inaccurate or unsafe information in specific fields, such as automated control systems, critical information infrastructure, and medical information services.

While the February 2024 technical document does not list any frontier AI safety risks among the 31 main safety risks, it makes a broad suggestion for providers to "pay close attention to long term risks" including deceptiveness (欺骗人类), self-replication (自我复制), self-modification (自我改造), and misuse in cyber (编写恶意软件), biological (制造生物武器), or chemical (化学武器) domains. However, the lack of concrete testing requirements for these risks suggests that frontier risks have not yet reached a threshold for triggering government action. The February technical document has since entered the process of being adapted into a more authoritative national voluntary standard.[2] The first draft of that national voluntary standard, published in May 2024, did not maintain references to "long term risks."

In addition to central level policies, local governments in jurisdictions key to AI development issued policies in mid to late 2023 calling for AI safety testing and evaluation. Anhui, Beijing, Guangdong and Shanghai all called for AI safety, ethics, and/or robustness testing.[3] These measures enable policy experimentation at the local level, which can factor into future national-level policies.

China's current policy framework therefore incentivizes developing AI safety evaluations and benchmarks primarily for issues such as ideological content, discrimination, IP protection, and privacy. Nevertheless, interest in evaluating frontier AI safety could grow, particularly with a new commitment at the highest levels of the Chinese leadership for "establishing AI safety oversight systems" at the Third Plenum of the Communist Party of China (CPC) in July 2024. Subsequently, in September, TC260 published a new AI Safety/Security and Governance Framework (Ch, En) during the Cyberspace Administration of China's Cybersecurity Publicity Week Forum to "implement" China's Global AI Governance Initiative. The framework incorporates severe misuse risks and loss of control risks such as "designing cyber weapons," "reducing capability requirements for non-experts to design, synthesize, acquire, and use nuclear, biological, and chemical weapons and missiles," "self-replication," and "seeking for external power."

[2] For more on TC260 and voluntary standards in the AI domain, see Concordia AI's The State of AI Safety in China Spring 2024 Report slide 67 and State of AI Safety in China report pages 13-15.
[3] For more information, see Concordia AI's The State of AI Safety in China Spring 2024 Report slide 70.

info@concordia-ai.com
www.concordia-ai.com

Methodology

We have collected a database of AI safety-relevant evaluations published by Chinese research groups. The dataset builds upon our pre-existing Chinese Technical AI Safety Database and was created by searching arXiv for papers with keywords such as "AI safety," "adversarial AND AI," "AI safety AND benchmark," "AI safety AND evaluation," "biological AND AI," "cyber AND AI," "AI AND bias AND benchmark," and "AI AND ethics AND benchmark." These search terms map onto 14 risk categories relevant to general purpose AI models, of which seven are commonly grouped under the umbrella of "frontier AI safety." Our definition and categorization of these risk categories draw upon taxonomies from the International Scientific Report on the Safety of Advanced AI: Interim Report, UK AI Safety Institute, and Center for AI Safety et al. The eight frontier AI safety risks reflect the UK AI Safety Summit's focus on misuse and loss of control risks, which are bolded in the below table.

The database only includes papers by authors at Chinese institutions primarily evaluating large models (e.g. medical imaging models would not be included). Evaluations are categorized by type into static benchmarks, automated platforms or leaderboards, AI agent evaluations, domain expert red teaming, human uplift experiments, and open-source testing toolkits, inspired by several existing taxonomies. Benchmarks that evaluate only AI capabilities, including AI R&D capability, were not included. This dataset is limited by publicly available information; in particular, evaluations performed by private companies and government-affiliated bodies may be disproportionately underrepresented. Evaluations for non-frontier risks, such as bias, and privacy, may be slightly underrepresented as they were not the primary focus of this report, but the search terms were designed to find such tests. See the database for the full explanation of our methodology and definitions of key terms.

info@concordia-ai.com
www.concordia-ai.com

Testing Dimensions

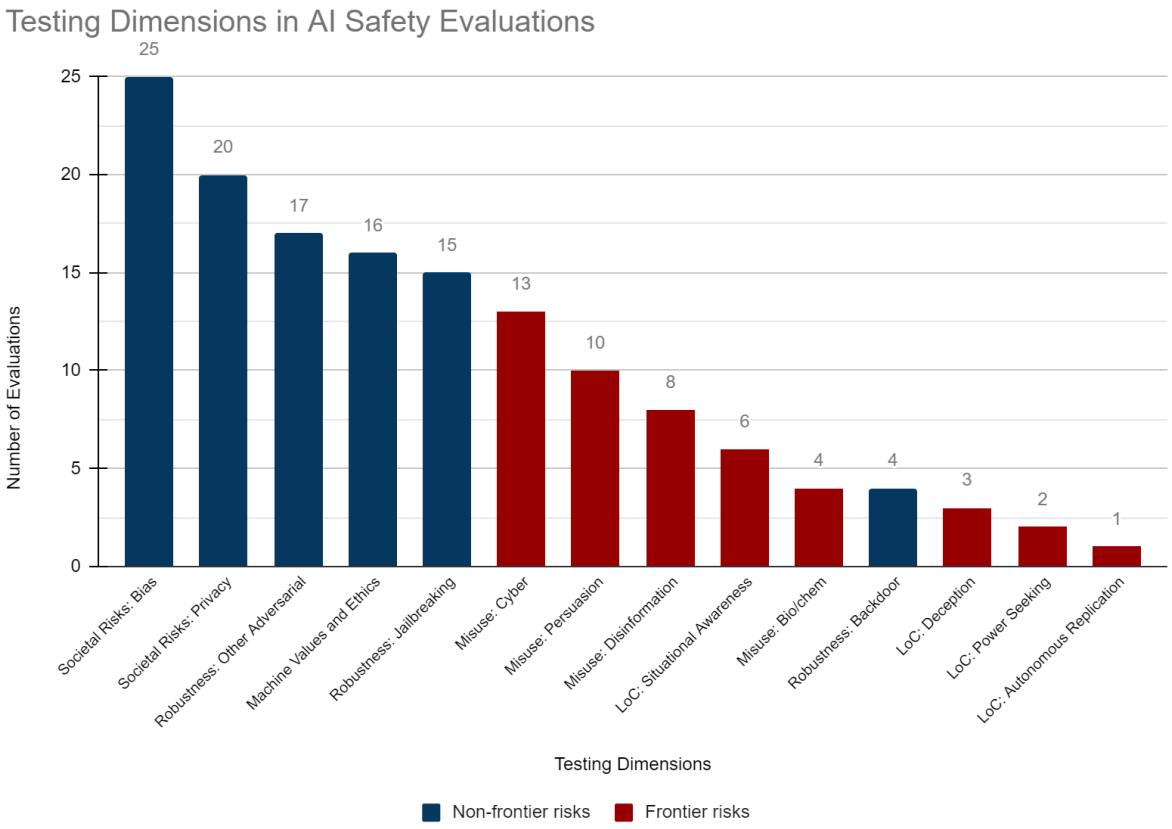
Misuse	**Cyberattacks**
	Biological and Chemical Weapons
	Disinformation
	Persuasion
Loss of Control	**Autonomous Replication or Adaptation**
	Deception
	Situational Awareness
	Power Seeking
Robustness to Attacks	Jailbreaking
	Other Adversarial Attacks
	Backdoor Attacks
Societal Risks	Bias
	Privacy
Machine Values and Ethics	Machine Values and Ethics

Notable Trends

Testing Dimensions

Evaluations have thus far concentrated on societal risks, with privacy and bias as the most-tested risk categories. Adversarial and jailbreak robustness have also received substantial attention, reflecting China's historically strong AI robustness research. Machine ethics was the fourth most common evaluation subject, indicating interest in value alignment and moral conduct. Overall, Chinese institutions have produced evaluations for all of the risks in our taxonomy, demonstrating a broad foundation that is strongest on testing societal safety and robustness.

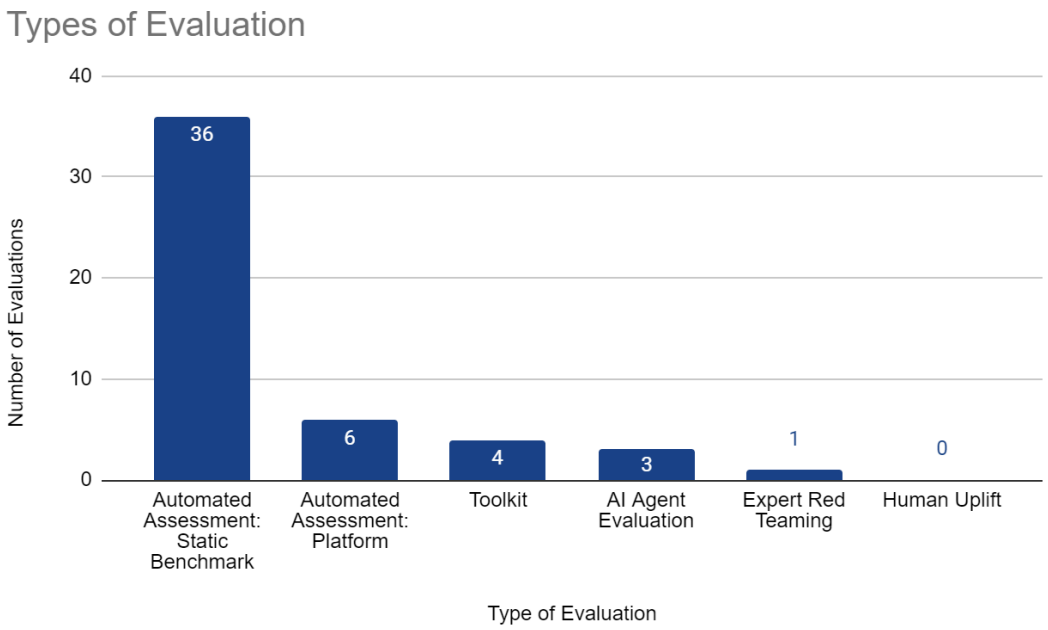
info@concordia-ai.com
www.concordia-ai.com



Cyberattack misuse received the most testing among frontier AI safety domains (Misuse: Cyberattacks, Misuse: Biological and Chemical Weapons, Misuse: Disinformation, Misuse: Persuasion, and Loss of Control categories), while evaluations of other frontier safety areas lagged behind. Cyberattacks were the sixth most-tested risk overall, with 13 evaluations, whereas loss of control domains, disinformation, persuasion, and biological or chemical weapons were each evaluated ten or fewer times. In all, 21 out of 41 evaluations tested for at least one frontier risk. The institutions conducting frontier safety evaluations range from government-backed independent institutions to academic research groups and private companies. Shanghai AI Lab (SHLAB) researchers were anchor authors on six such evaluations, compared to two for the next most prolific research groups, and further elaboration will be provided in the Key Frontier Safety Evaluations Groups section.

Types of Evaluations

The vast majority of AI safety evaluations in China are static benchmarks, some of which also include platforms that rank performance across a wider range of publicly-available benchmarks.[4] Despite the limitations of static benchmarks, they remain an important starting point for probing model safety and allow public comparison or open access to datasets. Our database also documents six open-source AI evaluation toolkits, including SHLAB's CompassKit and toolkits for specialized topics such as trustworthiness by Tsinghua University researchers. These toolkits can be adapted by researchers around the world for different evaluation use cases.



AI agent evaluations are an emerging area of research globally due expectations of increasingly agentic capabilities, and Chinese institutions are beginning to explore this domain with three evaluations for

[4] Automated Assessment: Static Benchmark refers to a dataset (usually multiple choice or question and answer), evaluation methodology, and scoring rubric that can be used to assess models. Automated Assessment: Platform refers to a platform that aggregates results from a number of preexisting static benchmarks to rank models against each other. AI Agent Evaluation refers to testing models for the ability to autonomously perform tasks, usually in a simulated environment. Red Teaming refers to dedicated adversarial testing by domain experts for topics such as cyberattack misuse and bias. Human Uplift refers to comparing performance of AI systems on a given task relative to the counterfactual dangerous case, e.g. biological experts with access to the internet. Toolkits refers to an open-source testing tool or library that other institutions can utilize to conduct evaluations.

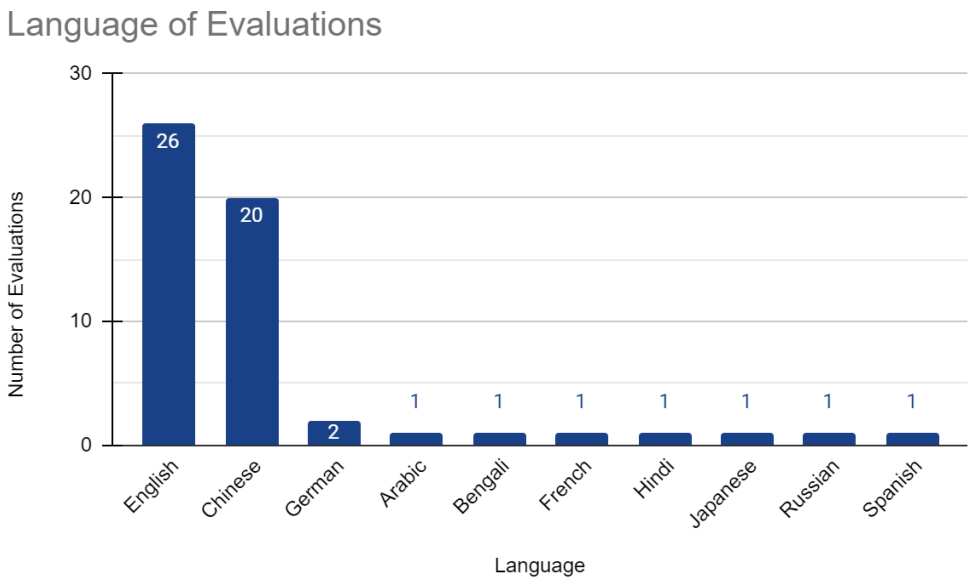
evaluating agents in virtual environments. Beijing Institute of General AI's (BIGAI) Tong Test tackles capabilities and value alignment of a virtual AI child. Shanghai Jiao Tong University's R-Judge examines agent risks in environments such as web search and financial transactions. Meanwhile, SHLAB's PsySafe evaluates safety of multi-agent systems based on agent psychology.

Domain expert red teaming is rare among Chinese evaluations projects, with only one coded in our dataset. However, leading Chinese large model startups Zhipu AI and Baichuan Intelligence conducted internal red teaming prior to model deployment, though this was likely red teaming of general safeguards rather than for domain-specific risks. Baidu's safety/security team has also written on red teaming for large model content generation. Meanwhile, at least four AI safety and security red teaming competitions were held in China between April and June 2024 by organizations including the Beijing and Shanghai Municipal Governments, Chinese Academy of Sciences, China Computer Federation, Tsinghua University, and Alibaba. These competitions focused mainly on adversarial attacks, jailbreaking, and generating unsafe content. Automating red teaming through LLMs is also a live research direction, such as in a recent paper by Tianjin University NLP Lab (TJUNLP). Overall, Chinese institutions show substantial interest in red teaming, but implementation in dangerous domains such as cyberattacks and biological or chemical misuse remains a gap.

There have not yet been any documented human uplift evaluations by Chinese safety evaluators.

Language

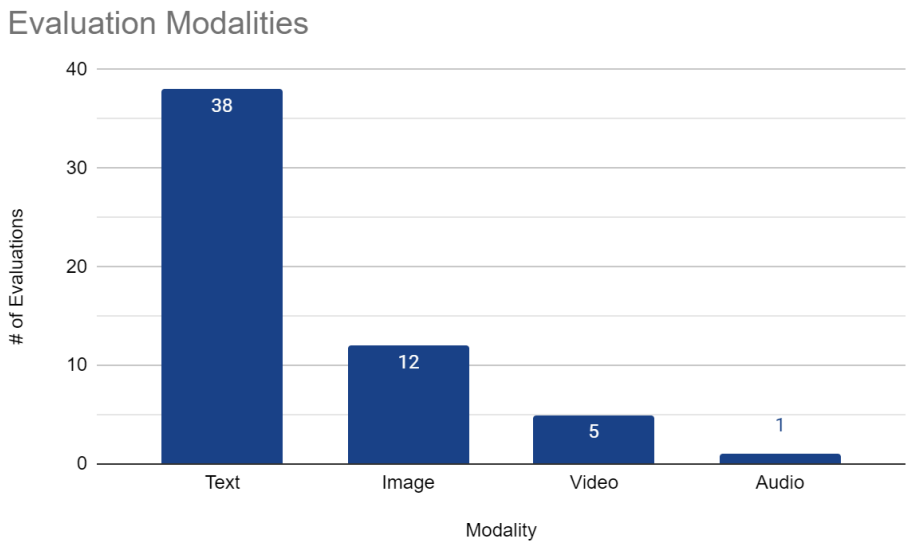
Surprisingly, English is the most common language used in evaluations by Chinese institutions, even more so than Chinese. This may reflect the interlinkage between AI safety evaluations research in China and the rest of the world, as Chinese benchmarks may often test non-Chinese AI models or utilize foreign benchmarks for inspiration.



Ten evaluations appear to be multilingual out of the 36 evaluations with information on languages tested. Multilingual benchmarks from China primarily combine Chinese and English, but the XSafety evaluation tests safety across ten languages including English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Japanese, and German. There is likely need for more work globally to improve model robustness to attacks in lower-resource languages.

Modality

Nearly all of the benchmarks documented include text as a testing modality. Almost one third of the evaluations also included some multimodal aspect, with image as the most common additional modality. While the multimodal benchmarks do not primarily test for frontier AI risks, several test for cyberattack misuse, biological or chemical misuse, and situational awareness. As safety of multimodal modals becomes an increasingly important topic, the substantial amount of Chinese research on this topic offers opportunities for mutual learning.



Key Frontier Safety Evaluations Groups

This section provides a detailed description of selected Chinese organizations with frontier AI safety evaluations, divided into three categories: government-backed independent institutions, academic research groups, and private companies. Institutions were coded as key evaluations groups if they had one senior researcher who was an anchor author on at least two frontier AI safety evaluations. All other government-backed institutions with AI safety evaluations were also coded as key evaluations groups, regardless of their number of frontier safety evaluations, due to their potential influence over government policy.

Government-backed research institutions

At least four government-backed institutions in China are pursuing AI safety evaluation projects.[5] Shanghai AI Lab (SHLAB) has issued the largest number of academic benchmarks, while the Beijing Academy of AI (BAAI) and China Academy of Information and Communications Technology (CAICT) have developed evaluation platforms that involve non-public safety datasets. Beijing Institute of General

[5] While one paper has an author from Peng Cheng Laboratory, another government-backed institution, the author appears to be primarily affiliated with Beijing Jiaotong University and no other authors have a Peng Cheng affiliation, so this is not coded as a Peng Cheng evaluation project.

AI (BIGAI) has an unorthodox test focused on evaluating capability and alignment of an AI child in a simulated environment.

These institutions all possess ties to policymaking, including through involvement in standards processes and participation in government-supported research projects. Therefore, their research and platforms could play an outsized role in influencing future government policy on AI safety evaluations, though academic researchers and companies also can undertake government projects and participate in standards. *Note: These institutions are ordered based on the number of frontier safety evaluations in our dataset, then if tied, on the number of frontier risks covered by their evaluations, then if tied, alphabetically.*

Evaluations by government-backed independent institutions

	Institution	Evaluation	Cyberattacks	Biological and Chemical Weapons	Disinformation	Persuasion	Autonomous Replication or Adaptation	Deception	Situational Awareness	Power-Seeking
1	SHLAB	SALAD-Bench	Yes	Yes	Yes	Yes		Yes		
2	SHLAB	Flames				Yes			Yes	
3	SHLAB	MM-SafetyBench	Yes		Yes	Yes				
4	SHLAB	OpenCompass	Yes			Yes			Yes	
5	SHLAB	PsySafe	Yes					Yes		
6	SHLAB	BeHonest						Yes		
7	BAAI	FlagEval	Yes			Yes				
8	CAICT	AI Safety Benchmark		Yes					Yes	
9	BIGAI	Tong Test								

Shanghai AI Lab (SHLAB)

- *Background:* SHLAB is a national-level research institution founded in 2020 and currently led by former JD.com Senior Vice President and Tsinghua professor ZHOU Bowen (周伯文). Director Zhou recently articulated a policy of balancing AI development with AI safety as a "45-degree law" at the World AI Conference (WAIC), and SHLAB has teams focused on large model safety and governance research. SHLAB participates in AI standards as leader of a large model-focused group under China's primary national AI standards body, with standards in the drafting and feedback phase. SHLAB is also participating in the drafting of an industry standard on generative AI safety/security assessments as part of a project under the Cyber Security Association of China.
- *Frontier AI safety evaluations projects:*

- SALAD-Bench. This LLM safety benchmark provides 200+ questions in many safety categories, including biological and chemical harm, cyber attack, and malware generation.
- Flames. This LLM benchmark tests five dimensions of value alignment, including on traditional Chinese values. It implements attack methods including disguise, reverse induction, and unsafe inquiries.
- MM-SafetyBench. This benchmark tests multimodal LLMs (MLLMs) against image-text pairs across 13 scenarios.
- OpenCompass. OpenCompass is a platform ranking large models across capabilities and safety dimensions using a variety of publicly available benchmarks. It also contains a chatbot arena-style performance evaluation and is accompanied by evaluation toolkits for LLMs and vision-language models.
- PsySafe. This is a framework for evaluating the safety of multi-agent systems for jailbreaking, misuse, and deception.
- BeHonest. This benchmark assesses three dimensions of honesty in LLMs to address risks of misinformation, misleading users, and even escaping human control.

Beijing Academy of AI (BAAI)

- *Background:* BAAI was founded in 2018 with the support of the Ministry of Science and Technology (MOST) and Beijing Municipal Government. BAAI stated that it is developing large model evaluation methods and tools as part of two different projects on AI platforms and evaluations by MOST and the Ministry of Industry and Information Technology (MIIT). BAAI co-organized the International Dialogues on AI Safety (IDAIS)-Beijing in March 2024, and three of its leaders – Founding Chairman ZHANG Hongjiang (张宏江), Chairman HUANG Tiejun (黄铁军), and Director WANG Zhongyuan (王仲远) – signed the ensuing joint statement among top Chinese and global experts on red lines for ensuring AI safety.
- *Frontier AI safety evaluations projects:*
  - FlagEval 天秤平台. BAAI created the FlagEval platform in June of 2023, and it provides automated evaluations of AI models using a range of private and public benchmarks. In May 2024, BAAI announced that FlagEval tested 140 Chinese and foreign large models across seven capabilities and safety categories.

China Academy of Information and Communications Technology (CAICT) / AI Industry Alliance of China (AIIA)

- *Background:* CAICT is a public institution overseen by MIIT. CAICT provides advice to the government on ICT policy, publishing reports on topics including large model governance. It also works closely with AIIA, a key industry association in the sector, to develop industry standards and best practices.[6] CAICT/AIIA provide evaluations as a paid service, which private companies can use to certify compliance with industry best practices and improve prospects for winning contracts with certain vendors. CAICT AI Research Institute director WEI Kai (魏凯) gave a presentation on CAICT's AI safety evaluations work at WAIC 2024.
- *Frontier AI safety evaluations projects:*
    - AI Safety Benchmark. This benchmark was announced in April 2024 by CAICT and AIIA; it has since been updated twice in Q2 and Q3 2024. The first version included 400,000 questions in three main categories: S&T ethics, data security, and content security. The Q2 version added attack methods and tested red lines for socialist values and illegal actions, while the Q3 version will increase focus on multimodal models. This benchmark is also the safety component of a comprehensive benchmarking project named "Fangsheng" in collaboration with the BAAI, iFlytek, and Tianjin University.

Beijing Institute of General AI (BIGAI)

- *Background:* The BIGAI was established in 2020 under the support of the Beijing Municipal Government, MOST, and Ministry of Education. It is pursuing an unusual and less-proven "small data, big tasks" approach to AGI, involving training an intelligent agent in a virtual environment, rather than seeking to scale large models. BIGAI Director ZHU Songchun (朱松纯) discussed AGI development, loss of control, and aligning AI with human values in a speech to representatives of China's top national political advisory body in January.
- *Frontier AI safety evaluations projects:*
    - Tong Test. BIGAI originally published this test in 2023. While it does not test for frontier AI safety risks as defined in this paper, it does test for value alignment. Notably, this evaluation involves use of a virtual environment for dynamic embodied physical and social interactions with AGI agents.

[6] For more information on CAICT and AIIA, see slides 63, 74, and 75 of Concordia AI's The State of AI Safety in China Spring 2024 Report.

Academic Research Groups and Private Research Labs

Many academic research groups have published benchmarks for societal and frontier AI risks. However, the only academic professor who was the anchor author of more than one frontier AI safety evaluation in our database is the head of Tianjin University NLP Lab. This shows that there is widespread interest among Chinese academics for researching AI safety evaluations, but few labs have decided upon evaluations as a main line of work.

There are two main types of private entities that have publicly announced AI safety evaluations work in China. The first type is China-based corporate labs of major technology or AI companies. The second is Chinese companies providing AI safety or security evaluations as a service. The only company with a researcher who has published multiple evaluations on frontier AI safety is Microsoft Research Asia's Societal AI team. Due to the paucity of public information on testing-as-a-service companies, we did not attempt to identify all such companies; nevertheless, we have included three notable companies in the sector – Ant Group, RealAI, and BotSmart – in the database to ensure some level of representation.

Evaluations by academic and private company research groups

Evaluation	Cyberattacks	Biological and Chemical Weapons	Disinformation	Persuasion	Autonomous Replication or Adaptation	Deception	Situational Awareness	Power-Seeking
CRiskEval					Yes	Yes	Yes	Yes
OpenEval							Yes	Yes
BIPIA	Yes		Yes	Yes				
SciMT-Safety		Yes						
SuperCLUE-Safety (SC-Safety)	Yes			Yes				
SafetyBench	Yes							
Bot AIGC Security Lab Test	Yes		Yes	Yes			Yes	
Trustworthy LLMs	Yes		Yes					
BabyBLUE	Yes	Yes	Yes	Yes				
S-Eval	Yes							
DeepfakeBench			Yes					
T2VSafetyBench			Yes					
CValues				Yes				

Tianjin University NLP Lab (TJUNLP)

- *Background:* TJUNLP conducts research on large language models, AI alignment & safety, multilinguality & machine translation, and AI for science. TJUNLP director XIONG Deyi (熊德意) participated in a panel on AI safety evaluations at WAIC.
- *Frontier AI safety evaluations projects:*
    - CRiskEval. This benchmark is fully focused on LLM frontier risks, covering 21 types of risks with four risk levels.
    - OpenEval. This platform assesses Chinese LLMs across capability, alignment, and safety, with the safety section based on translating questions in Discovering Language Model Behaviors with Model-Written Evaluations to Chinese.

Microsoft Research Asia Societal AI Team (MSRA)

- *Background:* MSRA's "Societal AI" team focuses on large model fairness, reliability, safety, privacy, assurance, tolerance, transparency, and responsibility. MSRA Societal AI Team leader XIE Xing (谢幸) participated in a panel on frontier AI safety evaluations, funding, and governance at the 2023 International AI Cooperation and Governance Forum.
- *Frontier AI safety evaluations projects:*
    - BIPIA. This benchmark seeks to protect LLMs from indirect prompt injection attacks across a range of real-world scenarios, including email, web, table, summarization, and code.
    - SciMT-Safety. This benchmark contains red-teaming queries focused on risk of AI misuse in scientific contexts, including around dangerous chemicals and biological toxins.