Issue Brief

China's AI Workforce

Assessing Demand for AI Talent

Authors
Diana Gehlhaus
Joanne Boisson
Sara Abdulla
Jacob Feldgoise
Luke Koslosky
Dahlia Peterson

CSET CENTER for SECURITY and EMERGING TECHNOLOGY

November 2022

AMPLYFI

Executive Summary

For the United States to build a globally competitive artificial intelligence (AI) workforce, U.S. AI education and workforce policies must grow, cultivate, attract, and retain the world's best and brightest. The AI workforce is global and in high demand, and a large share of top-tier technical talent in the United States is foreign-born. Given China is a major producer of AI-skilled talent, understanding its AI workforce could provide U.S. policymakers with important insight.

While a better understanding of China's AI workforce would be valuable, surprisingly little data is available. To date, most current knowledge stems from sporadic Chinese ministry, university, and company reports, or from news media outlets. A review of these reports suggests China is heavily investing in AI education and workforce development, in addition to recruiting native AI talent working abroad. But are these policies and investments producing the desired AI workforce?

To bridge the Chinese AI workforce data gap, we provide an analysis of the state of AI workforce demand in China. We assess AI workforce demand using job postings data, widely used in labor economics for this purpose. We do this using a novel sample of 6.8 million unique job postings in China, built in partnership with AMPLYFI, a UK-based machine learning company. The AMPLYFI team classified these postings as AI or non-AI related, using the same two technical and two non-technical occupational categories as defined in previous CSET research on the U.S. AI workforce.

We find that a sizable share—more than 30 percent—of the 6.8 million job postings analyzed could be considered AI or AI-related. Of these, 14 percent, or 955,000, are part of the technical AI workforce, defined here in two teams (Technical Teams 1 and 2) as talent that *could be* involved in the design, development, and deployment of AI.[1] The majority of technical AI job postings—about 53 percent—require a bachelor's degree, while most non-technical AI positions do not. Moreover, about 60 percent of technical postings are located in Guangdong Province, Shanghai Municipality, and Jiangsu Province, likely due to their economic importance as locations for major Chinese AI companies and universities. The wide range of median salaries across Chinese cities and provinces further points to areas where AI-related technical talent commands relatively high earnings, such as Shanghai and Beijing.

The majority of AI-related positions were also junior in terms of desired experience; this suggests three things. First, other informal avenues outside of traditional job

boards, such as social networks, university partnerships, and internal referrals, may be more often used to recruit more skilled or experienced hires. Second, given that many technical positions require at least a college degree, and given the rising number of college graduates, it suggests four-year degrees could be a gatekeeper market signal for technical positions similar to the United States. Finally, among the AI-related job postings CSET reviewed, many called for "fresh graduates." This could be a reference to the pervasiveness of the expected "9-9-6" (nine to nine, six days a week) work culture in China's tech community.

Moreover, we find a sizable share of technical job postings are for roles *that are* actually involved in AI. Scanning both technical teams' job postings using a list of AI keywords, we find that 10 percent, or 36,000, of Technical Team 1 postings (e.g., most computer and mathematical science occupations) were specific to AI, with "machine learning" and "robotics" as leading keywords. A larger share of postings containing an AI keyword also required graduate degrees; 17 percent of Technical Team 1 postings with an AI keyword did, compared 4 percent of all Technical Team 1 postings. A review of about 100 postings requiring a PhD showed a wide range of AI applications being recruited for including robotics, image recognition, blockchain, bioinformatics, autonomous vehicles, and smart cities.

Overall, our findings highlight the complexity of China's AI ecosystem and the need to better understand the state of China's AI workforce. Investing in more data on China's AI workforce could assist the United States in its own national AI strategic planning, particularly in developing targeted U.S. national security, education, and workforce policies.

This paper is the first in a series on AI workforce demand in China. It provides the foundation for future reports by detailing the dataset, methodology, and descriptive statistics for key fields of interest, such as desired education, experience, and salary. Future work will analyze these job postings in greater detail by providing insights into the geographic distribution of jobs, AI applications for which companies are hiring, skills and fields of study associated with AI job postings, and details of the companies hiring these workers. We hope this series will provide a holistic assessment of the nature of China's AI workforce from a demand-side perspective.

Table of Contents

Introduction

Not long ago, the United States was unmatched in its ability to recruit and retain the world's best AI talent. But current U.S. policies risk being outdated with the growing size of and demand for the AI workforce, and there are concerns that current limits on the number of foreign-born researchers able to stay are inadequate.[2] Meanwhile, China is actively targeting the many Chinese-born AI researchers living in the United States to return, along with incentives for non-Chinese scholars to move. To counter this, some in the United States are proposing analogous talent attraction programs.[3]

Having better data on China's AI workforce—supply and demand—could help inform the United States' AI education and workforce policies with the goal of staying globally competitive. This includes understanding what types of positions Chinese enterprises are hiring for, where they are located, and what types of education, skills, and experience are being sought. This is in complement to understanding China's AI education ecosystem, discussed in previous CSET research.[4]

Yet far too little data is currently available. This series of reports bridges the information gap, providing new data on the state of China's AI workforce as it relates to talent demand. We approach this through a demand-side assessment, looking at job postings for AI and AI-related talent. We analyze job postings as it is a widely used indicator in economics to measure employer demand for talent.[5]

Using a novel dataset, we analyzed 6.8 million unique job postings in China as posted in the first half of 2021. Our data collection methodology was designed to capture a representative sample of four Chinese job boards. Moreover, due to the strong representation of one predominant job board for technical talent, 51Job (前程无忧), we believe our sample allows for some generalization of our findings for technical talent.[6] However, this data should not be used for other direct U.S.-China comparisons because the sample frame of job boards is not representative in the same way as available U.S. job postings data aggregators such as Lightcast. Planned improvements will instead allow for more direct comparisons in future work. As such, we refrain from direct comparisons with the exception of desired education for technical talent job postings.

This overview paper is the first in a series of reports assessing China's AI workforce demand. This first report provides the big picture assessment, along with an explanation of the methodology for curating the dataset for analysis. For this series,

CSET partnered with AMPLYFI, a machine learning firm specializing in unstructured data analysis.[7] To analyze Chinese AI workforce labor demand, AMPLYFI curated, cleaned, and normalized Chinese job postings across four major national job boards.

We begin with an overview of China's AI education and workforce policies, along with current available data on the state of China's AI workforce as context for the country's investment in cultivating a globally competitive cadre of AI talent. We then provide our overview of demand for AI talent in China, starting with our definition of China's AI workforce and the process by which our data was curated and compiled. We then provide a high-level analysis of this data. The paper concludes with implications and a road map to the future reports in this series that will build on this novel dataset.

China's AI Education and Workforce Policy

Global ambitions, significant funding, a growing workforce, and a data-rich environment are positioning China for a new era of AI innovation. If the country is able to take advantage of AI's transformative potential with wide-scale AI adoption across industries, it could give China an important competitive edge.

Among China's stated AI priorities is to build an AI workforce through domestic education and training. Since 2017, China has released several strategic plans and initiatives designed to enhance AI education at each level of education, with strong emphasis on enhancing industry and academic partnerships. The jumpstarting effort for China's AI education drive was the seminal State Council New Generation AI Development Plan of July 2017, which called for systemic provision of AI education. China's Ministry of Education (MOE) released another important strategic document in April 2018, the "AI Innovation Action Plan for Institutions of Higher Education."[8] Among other goals, it called for opening "50 AI schools, research institutes, or interdisciplinary research centers by 2020."[9] Forthcoming CSET research shows that China has met and exceeded that goal.[10]

China's latest policy push to boost AI workforce development comes via the 14th Five Year Plan (2021–2025) for National Informatization, released by the Central Commission for Cybersecurity and Informatization. Its 19 references to talent include calls to consolidate categorizations of technical job titles, improve talent evaluation systems, and create big data job certifications.[11] The Plan also highlights measuring AI's impacts on education and training, and it supports certain third parties to provide on-the-job training and certifications. Leading companies are also encouraged to provide training to support higher education institutes, particularly microelectronics and software institutes.[12] Still, the Plan offers few details on the strategic execution.

The implementation of these plans at the postsecondary education level is apparent. An undergraduate AI major is now offered at 440 universities, all approved by China's MOE. In both 2020 and 2021, the AI major was the most popular new addition to universities' curricula.[13] In addition, AI education is provided through at least 50 AI Institutes, which are split into 17 research institutes and 33 AI colleges.[14] China has also called for increased supply and training for AI graduate students at the doctoral level,[15] and has stated that AI will be incorporated into the "Special Enrollment Plan for the Cultivation of High-level Talents in Key Fields Urgently Needed by the State."[16] Talents are called upon to apply AI to industrial innovation, social governance, national

security, and other fields, and to support the mission and needs of major national projects and development plans.[17]

At all levels of AI education, China's AI companies are asked to partner with universities to train both teachers and students. At the postgraduate level, the intention is to have a revolving door between industry and universities. First, at the teaching level, companies are encouraged to train university instructors in the latest cutting-edge methods.[18] Leading researchers at AI companies can also conduct studies through "double employment" at universities.[19] There is also an emphasis on increased and coordinated funding through university-industry partnerships.[20]

Overall, prior CSET research shows China appears to be following its strategic plans for AI education. However, the risk remains that a significant proportion of lower-tier universities have insufficient resources to properly train students in AI, while schools at the K–12 level in rural areas continue to struggle with adequate teaching capacity. This could ultimately affect consistency in providing Chinese youth with AI skills.[21]

Current Assessment of China's AI Workforce

While there is good information on AI education in China, compiling an assessment of the country's AI workforce is quite difficult. Reliable data is extremely limited, based mainly on anecdotal or sporadic reporting, with no publicly available official statistics on the supply of Chinese AI talent. For example, the Chinese government publishes only highly aggregated employment statistics in its annual Statistical Yearbook.[22] Most discussion instead comes from a range of reports from Chinese ministries, research institutes, and news media assessing the demand for AI and AI-adjacent talent.[23] What is available is presented here to provide context.

In terms of demand, Chinese ministries, universities, companies, and other media routinely report talent shortages in critical AI and AI-adjacent areas such as AI R&D, semiconductors, and cybersecurity.[24] Articles and reports cite employer surveys and the large numbers of unfilled vacancies in these fields, similar to in the United States.[25]

For example, in 2020 China's Ministry of Human Resources and Social Security released a report titled "Analytical Report on Current Employment Prospects for AI Engineers and Technicians." This report quantified an AI talent gap of more than 5 million workers, with a domestic supply/demand ratio of 1:10 (0.1).[26] This report further states: "If talent development is not strengthened, the talent gap will exceed 10 million by 2025."[27] Another 2020 report from China's Ministry of Industry and Information Technology's (MIIT) Talent Exchange Center, using internally collected data and data from the Chinese jobs platform BOSS (直聘 or Zhipin in Chinese), estimated the supply/demand ratios for algorithm research positions and application development positions to be 0.13 and 0.17, respectively. For AI-specific positions, it defined critical shortages as anything less than 0.4, estimating ratios of 0.37 for artificial intelligence chip engineers, 0.23 for machine learning engineers, 0.2 for natural language processing engineers, 0.09 for computer vision engineers, and 0.08 for intelligent speech engineers.[28]

The high demand for AI talent stems from the rise of China's AI community. This ranges from mega-companies such as Alibaba, Baidu, and Tencent to a robust ecosystem of AI-related start-ups and state-backed R&D institutes and research labs.[29]

That said, some reports suggest demand may dampen going forward. China's regulatory crackdown of its tech sector in the past year has led to some anecdotal

reporting that tech companies are downsizing their workforces, reducing the number of new tech-skilled hires, and lowering starting salaries.[30] China's "zero-COVID" policies resulting in long and repeated shutdowns of key industrial zones alongside slower economic growth, associated supply chain issues, and the looming threat of a global economic slowdown from high inflation could also affect AI-related employment demand.[31]

In terms of supply, there appears to be a gap between the specialized technical fields in demand and technical backgrounds of recent graduates. News reporting of data from Maimai, China's LinkedIn, shows a skills mismatch between the rapidly growing internet sector of years prior and the new emerging technologies seen as the focus of China's strategic investment.[32] This is also evidenced by recent articles continuing to suggest that new college graduates, half of whom studied science, technology, engineering, or mathematics (STEM),[33] often struggle to find good jobs that pay well.[34] One recent official statement from the Chinese government even suggests new graduates find work for local governments and organizations located in rural provinces. At the same time, a record number of Chinese youth are earning bachelor's and graduate degrees.[35] Previous CSET research also shows that China continues to increase the number of graduates holding advanced STEM degrees,[36] with doctorates projected to grow rapidly over the coming years.[37]

Moreover, evidence suggests that China's AI talent pool is largely domestic. A 2019 report from Chinese job board Liepin states that more than 90 percent of the country's AI workforce is native-born.[38] This has implications for the active discussion surrounding Chinese nationals returning from abroad. While CSET research shows high stay rates for Chinese nationals pursuing STEM doctoral degrees in the United States,[39] other research suggests more graduates may be returning to China in recent years.[40] Still, China is continuing efforts to attract top researchers globally to teach in its universities,[41] in addition to actively marketing to native Chinese talent currently abroad.[42] If it turns out that more STEM graduates are returning to China, this will further increase the available supply of technical talent.[43]

Measuring China's AI Workforce Demand

Gaps in available data limit our ability to describe China's AI workforce. As a result, we embarked on curating a novel dataset from which better assessments can be made of the demand for AI talent in China. This more data-driven assessment sheds light on the education, experience, salary, and location requirements for a range of AI job postings.

Defining China's AI Workforce

In order to analyze China's AI workforce demand, it is first necessary to define the set of workers that comprise this cadre of talent. Here the approach mirrors previous CSET research on the U.S. AI workforce.

Previous CSET research defines the AI workforce as "the set of occupations that include people who are qualified to work in AI or on an AI development team, or have the requisite knowledge, skills, and abilities (KSAs) such that they could work on an AI product or application with minor training."[44] This definition purposefully includes technical and non-technical occupations that may not currently be working on AI applications—but have the KSAs to do so—in order to capture the entire pool of AI talent. The 54 selected occupations, taken from the 2018 Standard Occupational Classification (SOC) System, were binned into four categories:

> (1) Technical Team 1: occupations that are or could be actively working in AI, are needed to provide technical inputs into AI applications, or could laterally move into an AI development role. (Examples: computer scientist, software developer, network and database administrator.)
>
> (2) Technical Team 2: occupations that have the related knowledge, skills, and abilities to perform technical roles on an AI team, either as is or with minimal additional training. (Examples: electrical engineer, web developer, IT support.)
>
> (3) Product Team: occupations that complement AI technical occupations in product development. (Examples: product managers, legal compliance officers.)
>
> (4) Commercial Team: occupations that provide support for the scaling, marketing, or acquisition of AI at the organizational level. (Examples: sales engineers, purchasing agents.)

Defining China's AI workforce in this way also allows for comparison with the U.S. AI workforce. Future reports in this series will make such comparisons as applicable.

Measuring Demand

We analyze job postings to measure employer demand for talent.[45] Data on job openings is important for measuring and evaluating workforce dynamics, because a high number of openings in a given occupation or field suggests a strong demand in which to anchor decisions on building supply.[46] Coupled with other indicators such as low unemployment and rising real earnings, persistent high demand could also indicate the presence of a talent shortage.[47] For example, the vitality of the cybersecurity workforce is so consequential to national security that the U.S. government, in partnership with Lightcast and CompTIA, created a tool known as CyberSeek to measure cybersecurity job postings as a way to identify potential talent shortfalls.[48]

Using cutting-edge machine learning techniques, in partnership with AMPLYFI, we compiled a dataset of 6.8 million unique job postings in China in the spring of 2021. Of these, about 2.1 million, or 31 percent, fell into one of the four categories of AI occupations. The sample came from four job boards: 51Job (前程无忧), CJOL, LinkedInCN, and Liepin (猎聘).[49]

The process of finding and collecting AI job postings involved three general tasks: (1) acquiring data by harvesting job postings from Chinese job boards and removing duplicates; (2) classifying each relevant job posting into an AI or non-AI category; and (3) normalizing the information pulled from those postings.

The steps involved in each of these tasks included several rounds of collaborative experimentation, iteration, and testing. For example, when developing the AI job classification model, AMPLYFI experimented with varying training dataset sizes, which CSET reviewed and provided feedback on at each stage.

Importantly, the tasks of harvesting and building a job classifier model were not completed chronologically. Each was worked on in parallel, and with iteration based on the success of attempted methods to incrementally improve results. The result is a pipeline that begins with the input of harvested job postings and ends with the output of job postings classified by occupational category with normalized information fields.

This data collection methodology was designed to capture a representative sample of four major job boards in China. Moreover, due to the strong representation of one predominant job board for technical talent, 51Job, the sample enables some broader generalization of this study's findings.[50] The development process is explained at greater length later in this report. Appendix A contains a complete list of data fields extracted, and Appendix B contains a full methodology.

China's AI Workforce Demand

The total number of job postings harvested for our dataset is 10.1 million. Of these, 3.3 million were identified as duplicates, resulting in a final dataset of 6.8 million unique Chinese job postings.

The final dataset contains approximately 2.1 million job postings that were classified as "AI" using the definition above, or about 31 percent of all 6.8 million extracted job postings.[51] The breakdown of job postings by category is provided in Table 1.

Table 1. Chinese Job Postings by Category

Category	Number Job Postings	Share
Technical Team 1	364,358	5.4%
Technical Team 2	591,074	8.7%
Product Team	453,230	6.6%
Commercial Team	692,420	10.2%
Not AI	4,707,216	69.1%
Total Job Postings	6,808,298	100.0%

Source: CSET and AMPLYFI Chinese job postings dataset.

For comparison, we analyze each AI job category alongside job postings deemed to be not AI-related. Using the data, we describe summary measures of basic characteristics. While we provide summary data for all categories of the AI workforce, much of our analysis focuses on Technical Team 1 and Technical Team 2 job postings due to higher policy interest. The following sections highlight the distribution of jobs by desired education, desired experience, average salary ranges, and geographic location.

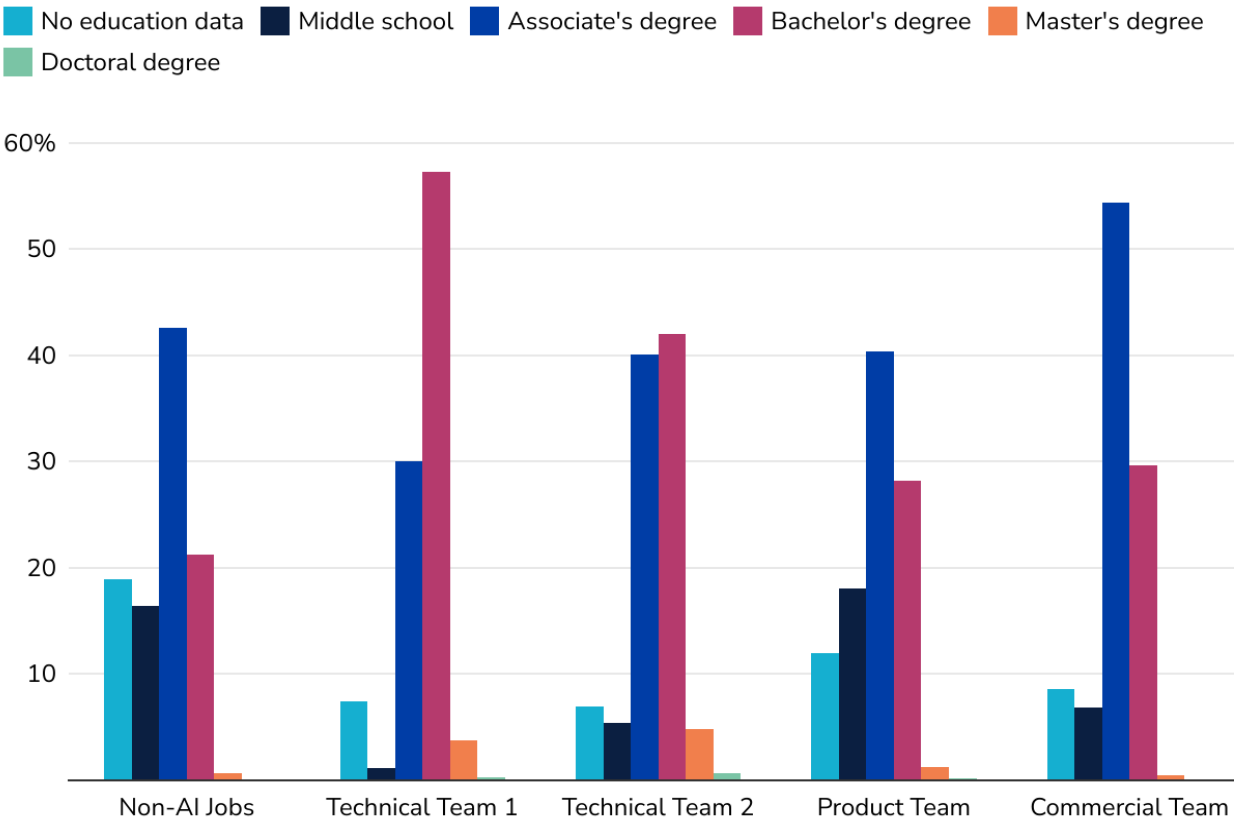
Desired Education

Desired education varied across AI workforce categories, although the overwhelming majority of job postings required at least some postsecondary education. However, of these, more technical positions desired at least a bachelor's degree, while non-technical Product and Commercial Team roles were more likely to ask for the

equivalent of an associate's degree.[52] (Detail on the educational categories used for this analysis is provided in Appendix C.[53])

Figure 1 shows the distribution of job postings by desired educational attainment for each of the workforce categories. Of all the categories, Technical Team 1 had the most job postings requiring at least a bachelor's degree, more than 60 percent. About one-third of the job postings in Technical Team 1 did allow for less than a bachelor's degree, but this compares to almost half of Technical Team 2 and about 60 percent of both non-technical AI job categories. Relatively few job postings required a graduate degree, although those that did were mostly for technical roles.[54]

Figure 1. Most Technical Team 1 Job Postings Require a Bachelor's Degree



Source: CSET and AMPLYFI Chinese job postings dataset.

Table 2 provides desired education requirements for technical talent in China compared with technical talent job postings in the United States. For the United States, we used Lightcast job postings that matched CSET's definition of technical talent,[55] and analyzed the desired educational attainment for those postings.

Table 2. Most Technical Team 1 postings in China and the United States Seek a Bachelor's Degree

	Educational Level	Technical Team 1	Technical Team 2
Chinese AI Workforce Job Postings	Less than a bachelor's degree	31.2%	45.5%
	Bachelor's degree or higher	61.3%	47.5%
	No data	7.5%	7.0%
U.S. AI Workforce Job Postings	Less than a bachelor's degree	7.2%	10.1%
	Bachelor's degree or higher	60.0%	57.4%
	No data	32.8%	32.5%

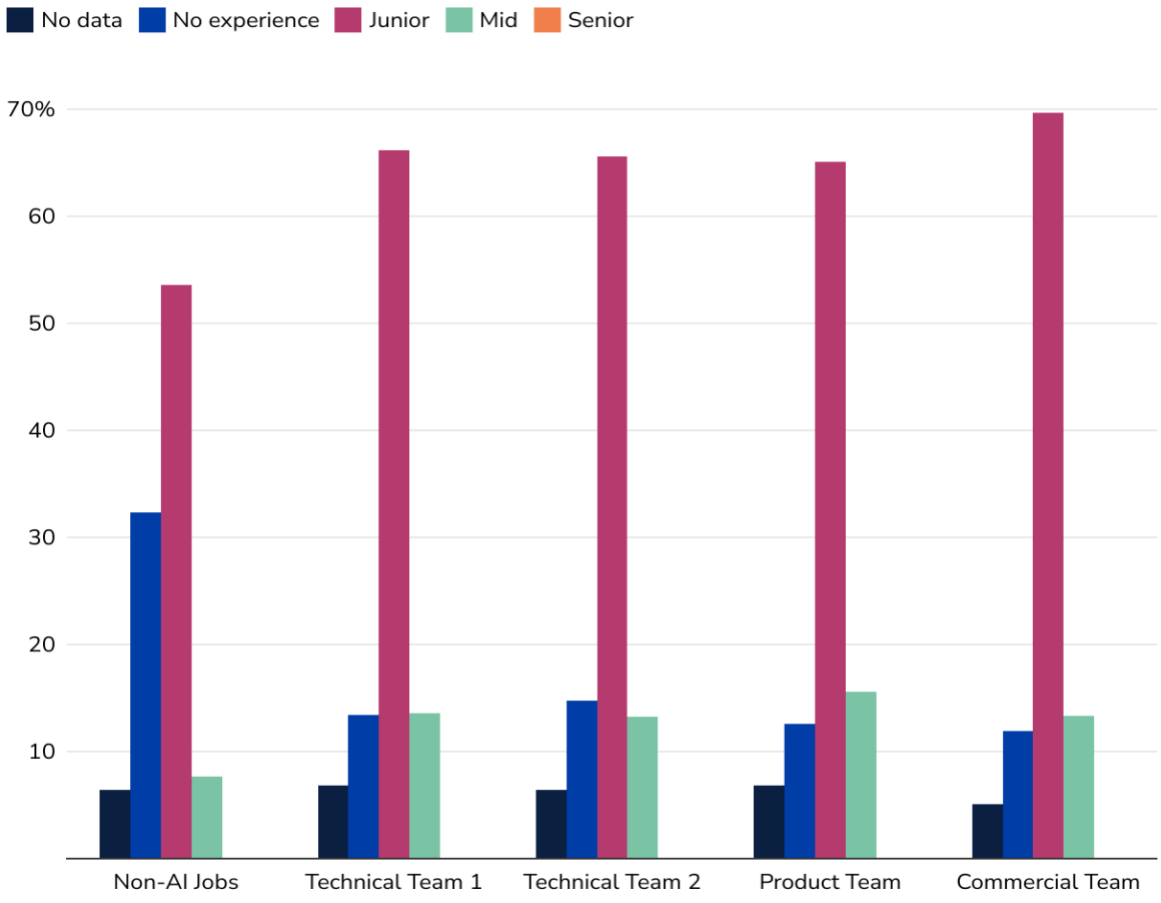
Source: CSET and AMPLYFI Chinese job postings dataset; CSET calculations using Lightcast data of minimum degree requirements for postings in 2021.

When it comes to technical talent, a large share of employers in both China and the United States are looking for a bachelor's degree or higher. The implication is that in both countries, many employers likely view a bachelor's degree as a gateway credential, especially for Technical Team 1 positions. That said, shares in the Chinese job postings analyzed suggest more appreciation for associate's equivalent degrees than in the United States. This is especially true for Technical Team 2 job positions, where the share specifying a bachelor's degree is almost on par with the share specifying an associate's degree (47.5 percent vs. 45.5 percent).

Desired Experience

Across all 6.8 million job postings analyzed, nearly 80 percent were for entry- or junior-level positions requiring five years or less of experience.[56] The remaining postings were mostly for mid-level positions requiring 5–15 years of experience, with very few executive or senior-level positions. Experience requirements across the four categories of AI jobs were similar, even between technical and non-technical roles, with about two-thirds of postings requiring 1–5 years of experience. Figure 2 shows the distribution of required experience across each AI workforce category.

Figure 2. The Majority of Job Postings Analyzed are More Junior



Source: CSET and AMPLYFI Chinese job postings dataset.

The difference between AI workforce job postings and job postings classified as not in the AI workforce is apparent in the number of postings that require no experience. Almost one-third of non-AI workforce postings did not require experience, more than

twice as many as AI workforce postings.[57] The difference likely exists because non-AI workforce job postings encompass numerous industries such as retail, administrative, and service sector jobs that are less likely to require experience.

There could be several explanations for why such a large number of AI workforce job postings are for more junior positions. Mid- and senior-level positions may be more likely to be filled through internal referrals, recruiting firms, or social and personal networks. Specialized job posting sites, including WeChat, official government websites, and company web pages might also be more often used for jobs requiring highly specialized skills or experience.

The importance of personal connections is evident in the literature. In China, the personal and business connections that make up an individual's social network are referred to as *guanxi* (关系). Guanxi strongly implies mutual obligation between parties, and it plays a large role in Chinese business culture. Workers with strong guanxi can leverage those relationships when seeking a promotion or pursuing a new job.[58] As positions increase in seniority, it is reasonable to assume that guanxi would play an increasing role in filling positions—much as "networking" does for experienced U.S. positions—potentially explaining the few mid- and senior-level job postings.[59]

It could also be the case that there are more entry-level positions in the Chinese labor market because of the occupational and demographic composition of the workforce. In terms of occupations, generally there are fewer mid- and senior-level positions to hire for within organizations. Moreover, although many technical job postings may be more junior in terms of experience, most instead cite a college degree or higher for educational attainment. Another potential implication relates to the many postings CSET reviewed, including ones for technical talent, that called for "fresh graduates." This could reflect a preference for talent willing to practice the arduous "9-9-6" (nine to nine, six days a week) informal but expected work schedule pervasive in the Chinese tech community.

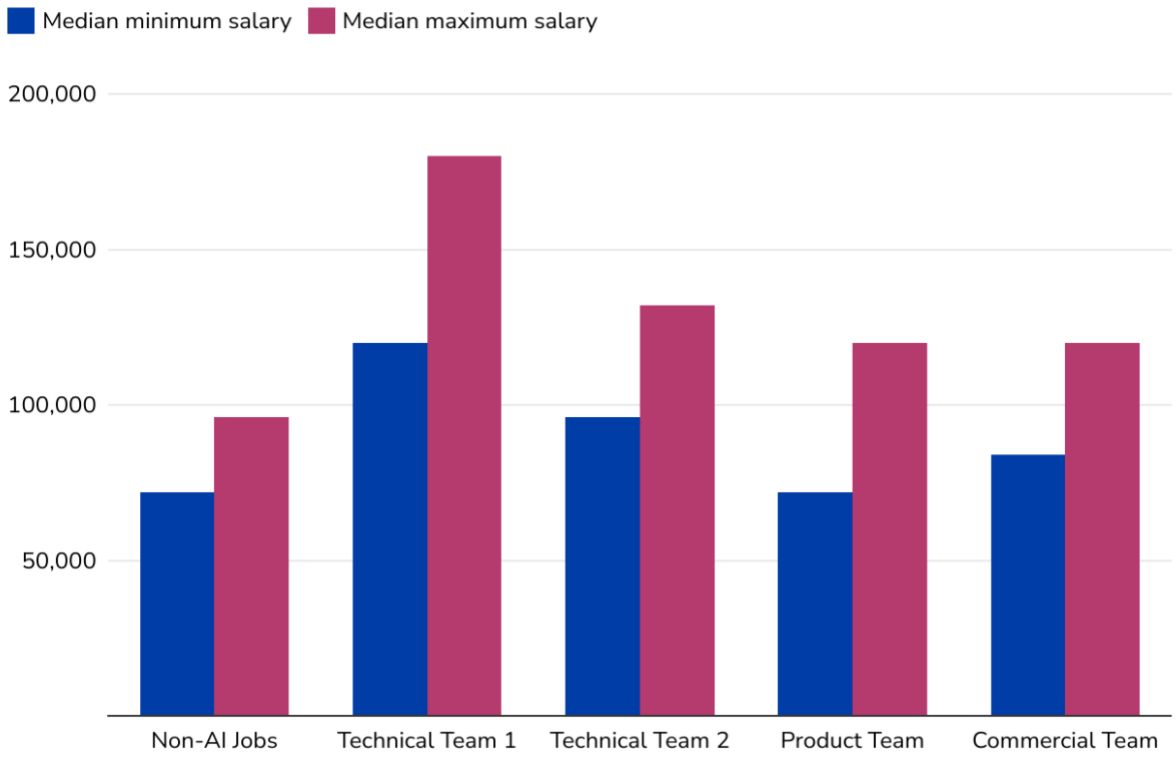
In terms of demographics, it could be that the sharp rise in college enrollments and graduates is moving the qualification emphasis toward education requirements. This mirrors the United States, particularly for more junior positions, where employers have a strong reliance on four-year degrees as a gateway market signal. This is further supported by the fact that news media reporting suggests placement "tri-contracts" between graduates, universities, and companies are common.[60]

Median Salary

Most salary data provided in the job postings were listed as a range. Therefore, this analysis uses the median minimum and maximum salaries. These fields were present for 94 percent of the job postings.[61]

Figure 3 shows the salary distribution across AI workforce categories. Job postings across AI workforce categories cited higher median salaries than non-AI workforce job postings, and Technical Team 1 job postings offered the highest median salaries.[62] The fact that both technical teams offered higher median salaries than non-technical AI teams and non-AI workforce job postings aligns with the listed desired education.

Figure 3. Median Salaries are Highest for Technical Team 1 and 2 Job Postings



Note: Approximately 6.5 percent of non-AI workforce postings were missing salary information, while between 5.4 percent and 8.4 percent of AI workforce postings in each category were missing salary information.

Source: CSET and AMPLYFI Chinese job postings dataset.

We also considered the purchasing power parity (PPP) salaries to enable a more appropriate value comparison for readers. PPP exchange rates account for the difference in price levels of goods and services in each country, allowing for a more direct comparison of the relative value of income. The median minimum and maximum salary by job category is provided in Table 3, along with the PPP conversion for 2021.[63]

Table 3: Median Salaries in CNY and USD Using the 2021 PPP Exchange Rate

	Annual Salary in Yuan		Purchasing Power Parity in USD	
	Median Minimum Salary	Median Maximum Salary	Median Minimum Salary	Median Maximum Salary
Non-AI Jobs	72,000	96,000	17,196	22,928
Technical Team 1	120,000	180,000	28,660	42,990
Technical Team 2	96,000	132,000	22,928	31,526
Product Team	72,000	120,000	17,196	28,660
Commercial Team	84,000	120,000	20,062	28,660

Source: CSET and AMPLYFI Chinese job postings dataset.

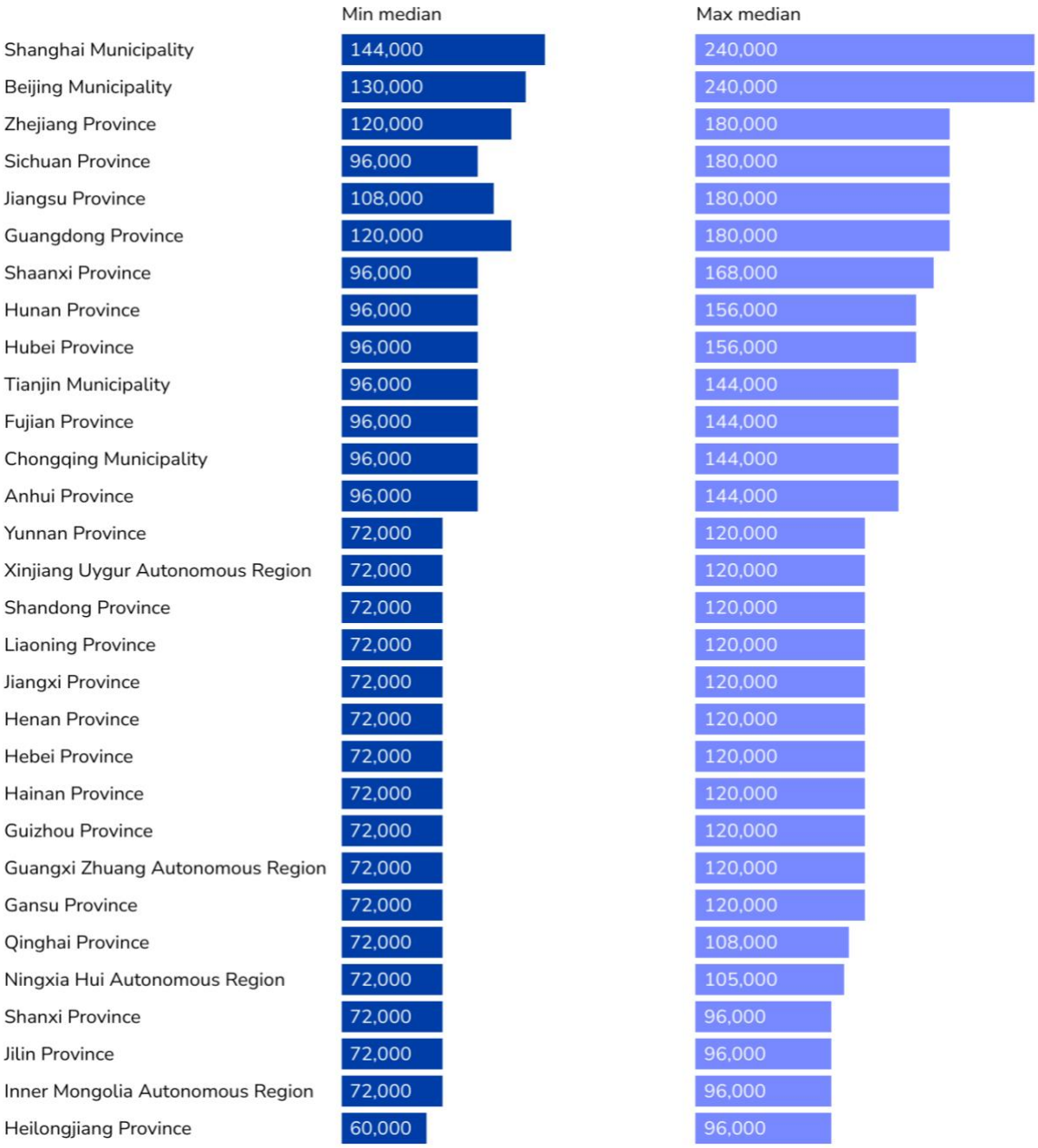
After converting the salary data using PPP, the median value of each job category ranged between $28,660 and $42,990 for Technical Team 1 postings to $17,196 and $28,660 for Product Team postings.[64] To ensure validity and increase confidence in these ranges reflecting actual starting salaries offered for these roles, several checks were conducted on the salary values from the job postings.[65]

A likely factor behind these salary ranges is the high percentage of entry- and junior-level postings within the corpus. It follows that these positions may have lower salaries than jobs requiring more experience, so that actual median salaries paid in China for these positions may be higher.

Another potential factor is the variability of salary by location. Country-wide PPP calculations do not take into account regional purchasing power disparities within the country. As is this case in the United States, it is reasonable to expect a difference in China's average salaries by region.

Figure 4 shows median salary ranges by province for Technical Team 1 job postings. Notably, major population centers such as Beijing and Shanghai had higher median salaries than did smaller provinces. For comparison, the 2021 PPP conversion to USD for the median range of salaries in Shanghai, the highest-paying region, was $34,392 to $57,320. The PPP converted median salaries for Heilongjiang Province, the lowest-paying region, ranged from $14,330 to $22,928.

Figure 4: Provincial Median Salaries for Technical Team 1 Vary Widely (in CNY)



Note: This chart includes only provinces with at least 100 non-missing values; the Tibet Autonomous Region was excluded because it only had 80 non-missing values. The number of non-values by province ranged from 116 (Qinghai Province) to 96,877 (Guangdong Province).

Source: CSET and AMPLYFI Chinese job postings dataset.

Geographic Distribution

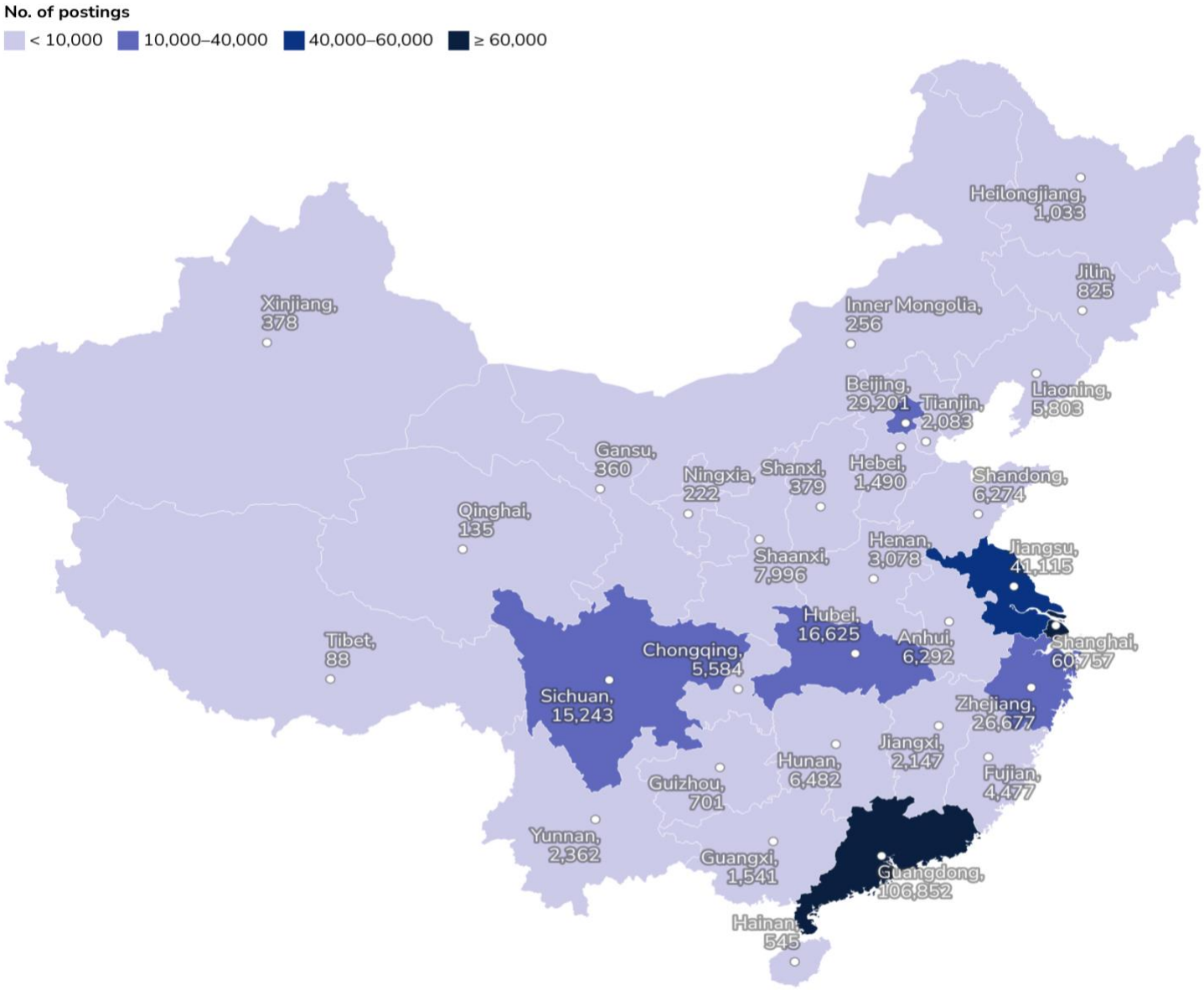
Job postings for technical talent in China demonstrate a notable geographic concentration. About 30 percent of the job postings for both Technical Team 1 and Technical Team 2 are in Guangdong Province alone, which resoundingly dominates job postings for both categories. Shanghai Municipality and Jiangsu Province also have large concentrations of job postings, and combined these top three regions account for almost 60 percent of postings in both categories.

Several factors likely explain the prominence of these areas. According to China's National Bureau of Statistics, in 2020 the populations for Guangdong Province, Shanghai Municipality, and Jiangsu Province were 126 million, 24.9 million, and 84.7 million respectively.[66] In addition to being China's largest province by population, Guangdong is also known for its advanced technology hub in Shenzhen, China's Silicon Valley, and is home to tech giants including genomics company BGI, drone maker DJI, Huawei, and Tencent. Shanghai has the Instagram-equivalent Xiaohongshu and facial recognition unicorn Yitu, while Jiangsu Province has a higher concentration of academic institutions in Nanjing and Suzhou.

These three regions are also economically important in China. Guangdong Province had the highest gross domestic product (GDP) of all provinces in 2020 at 11.1 trillion yuan, while Jiangsu ranked second at 10.3 trillion yuan. As a city, Shanghai was smaller economically, although still among top-performing regions with a GDP of 3.9 trillion yuan.[67] However, converting to GDP per capita using the population estimates above, Shanghai is an economic hub in terms of wealth concentration with a per capita GDP of over 156,700 yuan ($37,490). This compared to 121,300 yuan ($29,030) in Jiangsu Province and 88,200 yuan ($21,110) in Guangdong Province, with average per capita GDP for China at $10,360 in 2020.[68]

Figures 5 and 6 provide the geographic distribution of job postings classified as Technical Team 1 and Technical Team 2. As Figure 5 shows, postings for jobs classified as Technical Team 1 appear to be fairly concentrated in Guangdong Province, Shanghai Municipality, Jiangsu Province, and Zhejiang Province. Together these four provinces accounted for about 235,000, or 65 percent, of Technical Team 1 job postings.

Figure 5. Technical Team 1 Job Postings are Concentrated



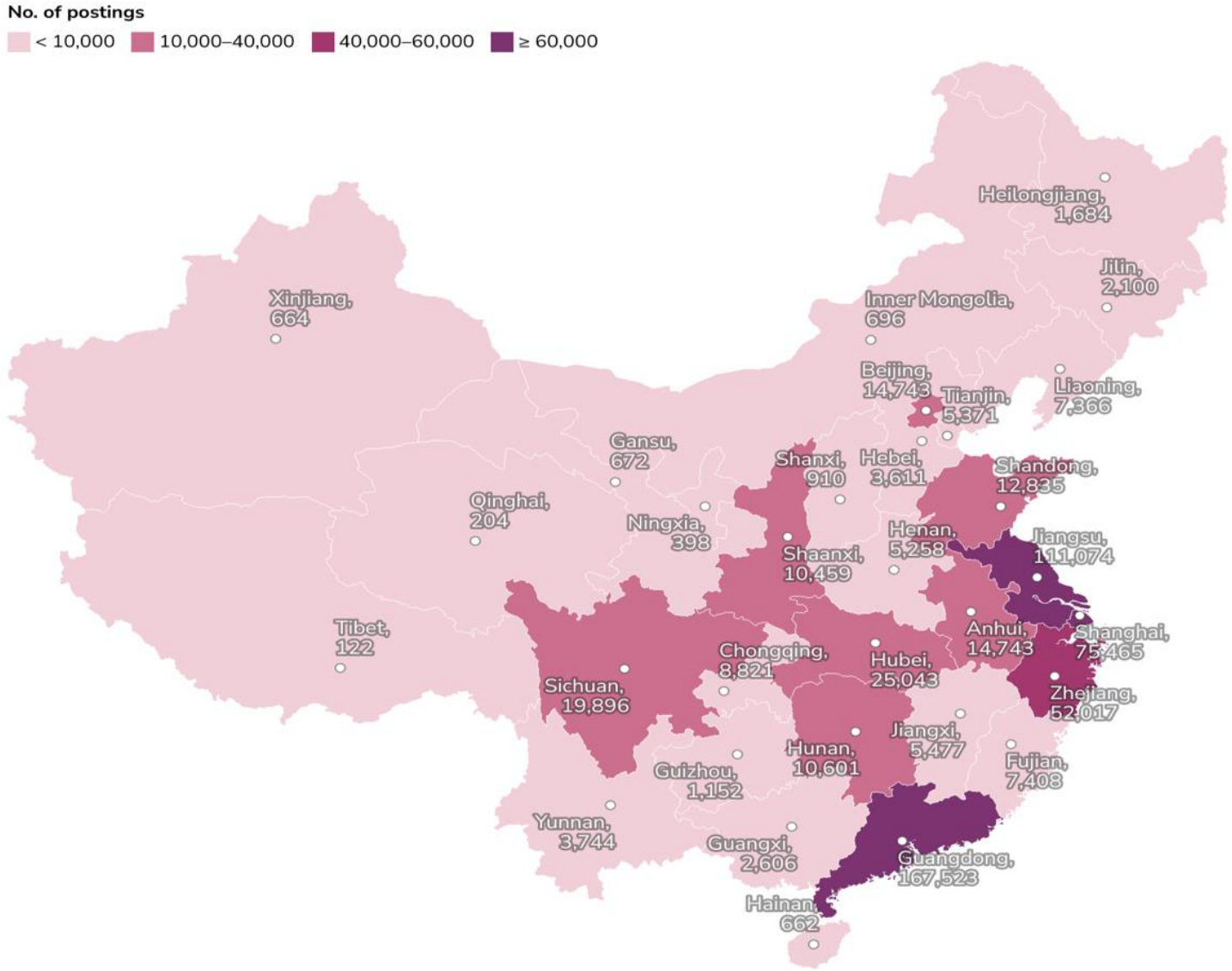
Note: Of these entries, 2.0 percent were missing location data. Taiwan and Hong Kong are not depicted because the preponderance of the job boards included in this study concern mainland China.

Source: CSET and AMPLYFI Chinese job postings dataset.

Job postings for Technical Team 2 positions suggest a similar degree of geographic concentration. Figure 6 shows that four regions—Guangdong Province, Jiangsu Province, Shanghai Municipality, and Zhejiang Province—accounted for about 406,000, or 69 percent of all Technical Team 2 job postings in the sample.

Subsequent research will provide a detailed analysis of geographic distribution and clustering as it relates to key industrial and investment hubs, top-ranked universities, university-affiliated AI institutes, and organizations affiliated with the People's Liberation Army (PLA). It will also explore the top companies hiring for these positions within each province and municipality.

Figure 6. Technical Team 2 Job Postings are also Concentrated



Note: Of these entries, 1.9 percent were missing data. Taiwan and Hong Kong are not depicted because the preponderance of the job boards included in this study concern mainland China.

Source: CSET and AMPLYFI Chinese job postings dataset.

Job Demand for AI-Specific Positions

Our definition of AI workforce is intentionally broad, including all technical and non-technical occupations with the knowledge, skills, and abilities that could work on an AI product team. While this is valuable, it is also interesting to understand job postings that are AI-specific in order to better capture AI-specific hiring trends. That is, job postings *that are* actively engaged in AI applications as opposed to the broader interpretation of postings *that could be* engaged in the design, development, and deployment of AI.

To capture the share of job postings that are explicitly AI, we conducted analysis on the subset of AI workforce postings where the job title or description contained an identified AI keyword (list provided in Appendix D). To minimize the potential mischaracterization of AI keywords within a job description, and again for policy relevance, we consider only job postings identified as Technical Team 1 or Technical Team 2.

A notable share—10 percent, or about 36,000—of all Technical Team 1 job postings contained at least one AI keyword. Table 4 shows the top ten keywords in descending order. “Artificial intelligence” and “machine learning” lead the keyword list for Technical Team 1, with “deep learning,” “robotics,” and “data mining” also having a sizable number of distinct postings with mentions. “Robotics” is also the leading keyword in Technical Team 2 job postings, potentially suggesting a stronger focus on hiring by China’s AI-engaged firms for applications related to machine learning and robotics. Still, demand for skills related to other AI applications such as computer vision and human-machine teaming were also visible on the list. Future research will provide the results of detailed analysis—topic modeling—to identify more granular themes.

Table 4. Technical Team Chinese Job Postings Containing an AI Keyword

	Technical Team 1		Technical Team 2	
Rank	Keyword	Number of Job Postings	Keyword	Number of Job Postings
1	artificial intelligence	11,904	robotics	12,352
2	machine learning	10,111	human machine	5,483
3	deep learning	8,322	artificial intelligence	1,811
4	robotics	7,762	machine learning	433
5	data mining	7,658	data mining	327
6	learning algorithm	4,021	deep learning	312
7	computer vision	3,500	autonomous	279
8	pattern recognition	2,685	facial recognition	222
9	human machine	2,619	pattern recognition	158
10	natural language processing	1,688	computer vision	141
	Top 10 keywords	34,406	Top 10 keywords	21,603
	All keywords	36,198	All keywords	22,031

Note: Totals for top 10 keywords and all keywords are the distinct number of postings containing at least one keyword to avoid double-counting.

Source: CSET and AMPLYFI Chinese job postings dataset.

Overall, a small share of Technical Team 2 job postings appear to be AI-specific. About 22,000, or 4 percent of identified Technical Team 2 job postings contain an AI keyword. This is likely due to the nature of occupations classified as Technical Team 2, many of which are for scientists and engineers, compared with the computer and mathematical occupations in Technical Team 1.

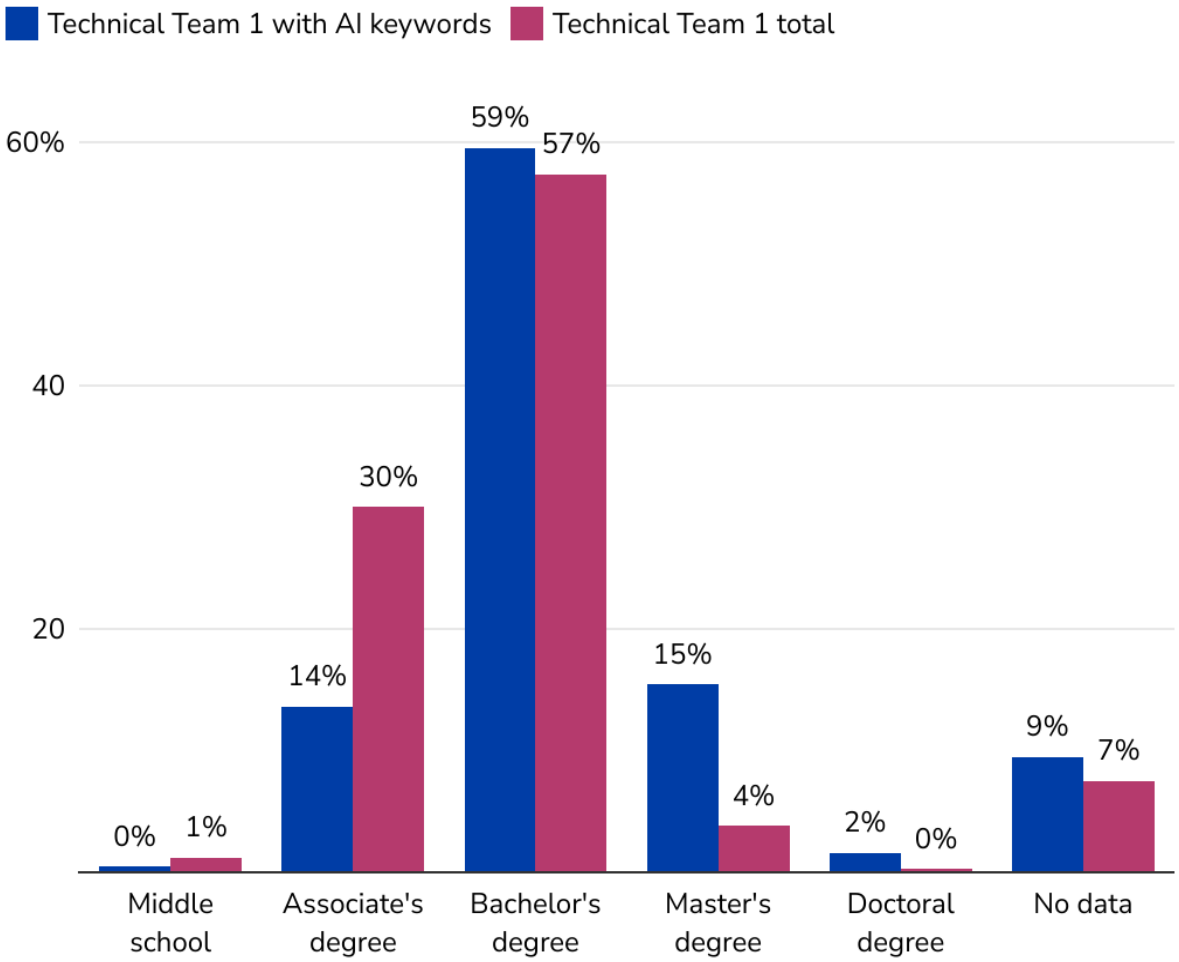
AI-Specific Desired Education

Far more Technical Team 1 job postings with an AI keyword than without required at least a bachelor's degree. More than three-quarters (76 percent) of AI-specific postings

within Technical Team 1 desired at least a bachelor's degree, compared to about three-fifths (61 percent) for all Technical Team 1 postings in the sample.

Moreover, more AI-specific jobs in this category required an advanced graduate degree. Whereas just 4 percent required a master's or doctoral degree across all Technical Team 1 job postings, 17 percent of Technical Team 1 postings containing an AI keyword did.[69] While future research will explore skills and field of study in more detail, computer science and electrical engineering were commonly specified. Figure 7 provides the distribution of education requirements for AI-specific job postings within those identified as Technical Team 1.

Figure 7. More Technical Team 1 Job Postings Containing an AI Keyword Desired an Advanced Degree

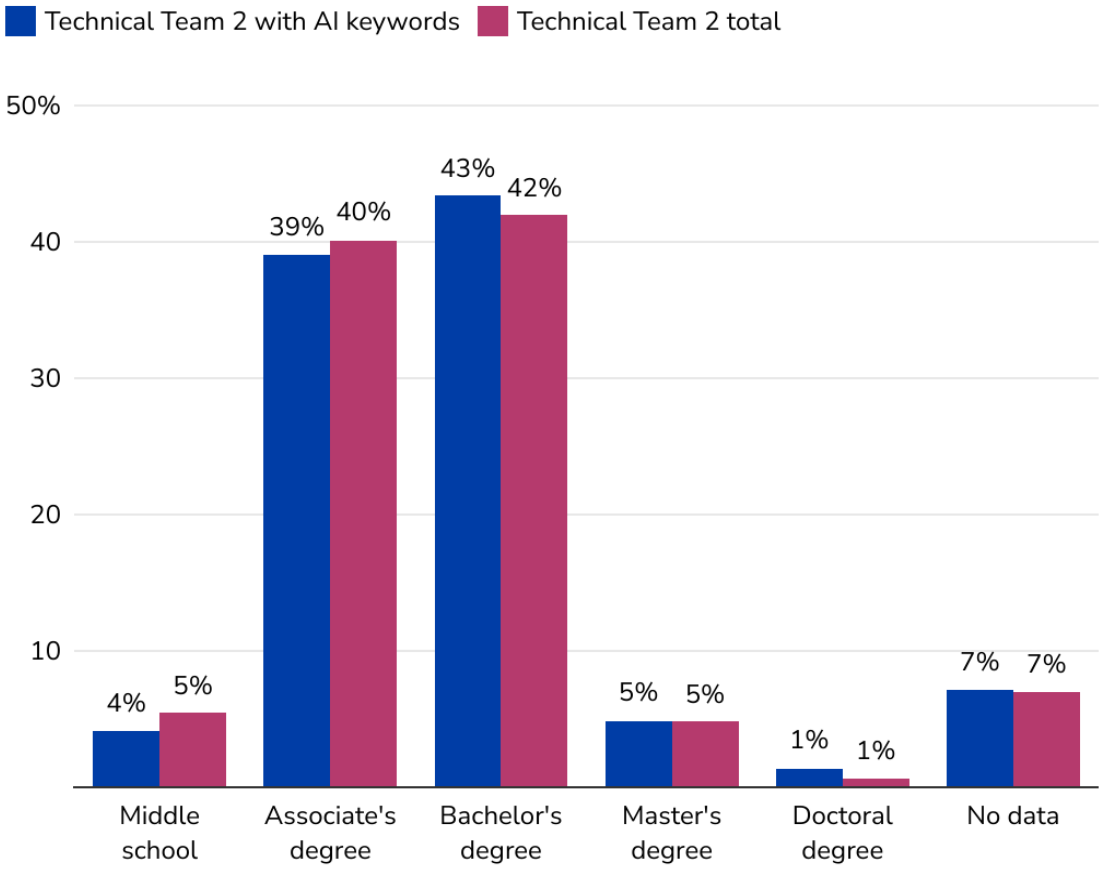


Source: CSET and AMPLYFI Chinese job postings dataset.

The number of postings that specifically mentioned a PhD was small, at about 800. The CSET team reviewed a sample of about 100 of these and found that a wide range of AI applications and hiring organizations were involved.[70] Applications being hired to develop included robotics, image recognition, blockchain, bioinformatics, autonomous vehicles, and smart cities, among others. Many postings asked for a record of publications in high-quality international journals and conference proceedings, and some explicitly preferred degree attainment from highly ranked institutions.[71]

In contrast, Technical Team 2 job postings with AI keywords were roughly aligned with the average for all Technical Team 2 job postings. About half of these postings desired at least a bachelor's degree, with a large share (about 40 percent) desiring the equivalent of an associate's degree. Figure 8 provides the distribution for AI-specific postings within Technical Team 2 by desired education.

Figure 8. Technical Team 2 Job Postings Containing an AI Keyword Mirror All Technical Team 2 Job Postings



Source: CSET and AMPLYFI Chinese job postings dataset.

Data Limitations

Compiling and analyzing demand for AI talent in China was challenging, as no such dataset previously existed. Building this with open-source data required considerable innovation in method and technique. With each step of the process came data limitations that could impact the results. The following section describes the main limitations by each step in turn.

Sample frame: The data collection methodology was designed to capture a representative sample of four major job boards in China. However, because it is not possible to be sure it is fully representative of all Chinese job boards, the generalizability of the findings presented here is limited. The limited ability to capture more job board postings with the approach used here was likely in part due to new, stricter privacy rules on Chinese job boards imposed in May 2021 by the Beijing Municipal Human Resources and Social Security Bureau after a large-scale data breach ("Regulations on the Administration of Online Recruitment Services"). While this harvest for job postings was conducted in 2021, the posting dates themselves ranged from September 2016 to July 2021. That is, all postings were active within a job board, but the date range of these postings was well beyond the date of extraction. Accordingly, no assertions are made or key findings stated that are specific to 2021.

U.S.-China Comparisons: The sample frame was not designed to be representative of all Chinese job postings, limiting the ability to make U.S.-China comparisons—with the noted exception for technical talent given the high prevalence of 51Job. This is because our data on U.S. job postings is representative of U.S.-based job boards, making the comparison not equivalent. Improved data incorporated into future research will allow for more direct comparisons, which will be the topic of future research.

Harvesting: A very high percentage of job postings—95 percent—came from just one platform, 51Job. Notwithstanding this study's assumption that this represents AI workforce demand in China due to the dominance of the platform in hiring for technical talent, it must be acknowledged that 95 percent is a very high share. We also note the potential for missing duplicate job postings in our approach to deduplication, along with issues related to inconsistent data availability across platforms. A notable limitation is that not all data fields were available, particularly for job postings outside of 51Job, and some initial errors in data extraction required a later re-extraction and repair.

Annotation: Regarding annotation, the categories of AI workforce used were complex, leaving open room for error. For example, very granular distinctions across categorical assignments within general occupation fields such as engineering, sales, and project management made it a challenge to fine-tune inter-rater reliability on annotations and model hyperparameter. Moreover, given that some annotation was conducted on an English translation, there is also the possibility that important context was lost in assignment.

Classifier Model: The key limitation is the performance of the model, particularly for Product Team and Commercial Team job postings. Despite best efforts to improve performance across a range of implemented best practices and testing, model accuracy scores in terms of precision and recall for these job categories remained lower than hoped. Moreover, although the classifier model performed better for Technical Team occupations, we acknowledge errors remain. Precision was optimized over recall, meaning that while the instance of job postings erroneously classified as Technical Team occupations was low (false positives), some postings classified as non–technical were technical (false negatives).

Field Normalization: Of all the normalized fields, salary had the most room for error. This is because of the large variation in how the data was entered by the hiring entity, including a range of contract terms and differences in whether pay was hourly, monthly, or annual. At times the terms were not clear and had to be imputed from similar postings. Moreover, while some salaries were exact, most were in a large range, limiting the power of analytic conclusions. Future work will include ongoing efforts to improve salary data quality. Separately, the high share of junior-level job postings meant that median salaries, when converting to purchasing power parity, were lower than might have been expected for technical talent. It is possible our sample is biased, but likely more job postings are inherently for junior talent.

Key Findings and Implications

This paper provides a first look at demand for AI talent in China. We find that of a sample of 6.8 million unique job postings across four popular Chinese job boards, over 30 percent could be counted as part of the AI workforce. Of these, 14 percent, or 955,000, could be counted as technical talent.

To account for differences across types of technical jobs, we split job postings for technical talent into two groups: Technical Team 1 and Technical Team 2. Technical Team 1 consists of computer and mathematical occupations that could be working in AI design, development, and deployment. Technical Team 2 consists of engineering, science, and other technical occupations that have the requisite knowledge, skills, and abilities to work in AI, but likely would first need some minor retraining or upskilling.

We find that most postings for Technical Team 1 positions require at least a bachelor's degree, similar to in the United States. This suggests that a four-year college degree acts as a gatekeeper to many of these positions, also similar to the United States.

Across all AI workforce categories, we find an overwhelming majority of identified Chinese AI job postings are junior or entry level. This suggests that hiring for more advanced talent (e.g., technical positions such as computer research scientists and machine learning engineers) may be more likely to happen outside of traditional online job boards, including via social networks, university partnerships, and internal referrals. Moreover, the high share of junior-level postings could point to a pervasiveness of an expected intensive work ethic targeted toward younger talent, and could reflect the large supply of college graduates entering the labor market each year for employers to choose from.

Moreover, there is evidence of geographic concentration for both Technical Team 1 and Technical Team 2 job locations. About 60 percent of all job postings for these teams were concentrated in Guangdong Province, Jiangsu Province, and Shanghai Municipality, with half of these (30 percent) in Guangdong Province. This is likely because these regions are important economically, containing many renowned AI companies and universities. Guangdong Province also has the largest population in China, likely a large factor in its dominant representation.

Using an AI keyword list to identify postings for jobs actively working in AI, we find that 10 percent (about 36,000) of job postings classified as Technical Team 1 contained an AI keyword. That is, these are job postings that *are* likely engaged in AI

applications as opposed to the broader interpretation of Technical Team 1. Machine learning and robotics were among the most frequently cited AI keywords within these positions. We also find that while a similar share of these jobs to all Technical Team 1 postings require a bachelor's degree (about 60 percent), a far larger share (4 versus 17 percent) require a graduate degree.

For the United States to maintain a competitive edge in AI talent attraction and retention, the data presented here underscores the need for, yet the complexity of, understanding the state of China's AI workforce. China's government and private sector are intentionally and strategically investing in developing and deploying AI capabilities, which could impact the global labor market for AI talent. Moreover, the geographic distribution of AI-related job postings also highlights the need to understand what is driving demand for AI talent beyond government strategic plans, and how changes in these factors could affect future supply and demand. For example, previous CSET research identified AI investment from public and private sources as driving AI advancement and adoption,[72] and other CSET research highlights the emphasis in the PLA on deploying various kinetic and non-kinetic AI operational capabilities.[73] Industrial development planning and execution is happening at the local government level, and a deeper understanding of those efforts could also contribute to broader understanding of China's demand for AI talent. The interplay of such factors on China's AI workforce dynamics, including Chinese local government AI strategic plans, will be the focus of a series of forthcoming reports.

Future Research

This is the first report in a series of research papers analyzing the state of AI workforce demand in China. Future work will analyze these job postings in greater detail. This will include providing more insights on the geographic distribution of these job locations, skills and fields of study associated with AI job postings, cluster analysis of job descriptions to understand the nature of AI applications companies are hiring for, and an analysis of the companies hiring these workers. An additional harvest of job postings will further enable year-over-year analysis for changes in trends and composition. Together, this series will provide a holistic assessment of the nature of AI job demand in China, and include more direct U.S.-China comparisons.

Regarding geographic location and associated hiring entities, future analysis will include a series of briefs specific to China's AI workforce demand based on geographic location and the hiring company or organization. These papers will: (1) identify where in China AI job clusters of U.S. national security interest exist, focusing on AI-specific jobs, surveillance jobs, and PLA-affiliated jobs, along with co-location analysis of known industrial park clusters, AI innovation pilot zones, and university AI institutes; (2) employ additional CSET-acquired investment data to discuss the relationship between public-private and private AI funding mechanisms and the previously identified job clusters; and (3) indicate where potential AI surveillance job clusters exist in China and the associated companies. Preliminary analysis indicates that most AI-related technical centers exist in traditional innovation hubs.

Appendix A: Full List of Job Posting Fields

Table A1 provides a complete list of all fields extracted for each job posting through the harvesting process.

Table A1. Job Posting Fields

Column name	Description
id	A unique identifier for the job posting
url_extracted	The url of the job posting at the time of harvesting
title_extracted	The title of the job posting extracted from the web page
body_extracted	The description of the job posting extracted from the web page
body_norm	The cleaned description for the job posting
company_name_extracted	The company name extracted from the job posting
company_information_extracted	The company information extracted from the job posting
release_date_extracted	The release date of the posting extracted from the job posting
location_extracted	The location of the job extracted from the posting
location_inferred	The location inferred from the description if the location_extracted field was empty
salary_extracted	The salary extracted from the posting
salary_inferred	The salary inferred from the description if the salary_extracted field was empty
salary_exact_monthly_value_pr	The exact monthly salary predicted by the salary prediction model
desired_experience_extracted	Desired experience extracted from the posting
desired_experience_inferred	Desired_experience inferred from the description if the desired_experience_extracted field was empty
desired_education_extracted	Desired education extracted from the posting
desired_education_inferred	Desired education inferred from the description if the desired_education_extracted field was empty
retrieve_date	The date this posting was retrieved by the harvester
source	The source of the job posting, e.g., "51Job"
pipeline_version	Useful in targeting documents for reprocessing should any elements of the pipeline be updated
job_category_pr	The job category predicted by the classifier for this posting
job_category0_distribution_pr	The probability of the job posting belonging to category 0

Column name	Description
job_category1_distribution_pr	The probability of the job posting belonging to category 1
job_category2_distribution_pr	The probability of the job posting belonging to category 2
job_category3_distribution_pr	The probability of the job posting belonging to category 3
job_category4_distribution_pr	The probability of the job posting belonging to category 4
salary_exact_value_norm	Exact value from the salary derived from either the inferred or extracted field, depending on which is present
salary_exact_annual_value_norm	Exact annual value from the salary derived from either the inferred or extracted field, depending on which is present
salary_exact_monthly_value_norm	Exact monthly value from the salary derived from either the inferred or extracted field, depending on which is present
salary_minimum_value_norm	Minimum value from the salary derived from either the inferred or extracted field, depending on which is present
salary_minimum_annual_value_norm	Minimum annual value from the salary derived from either the inferred or extracted field, depending on which is present
salary_minimum_monthly_value_norm	Minimum monthly value from the salary derived from either the inferred or extracted field, depending on which is present
salary_maximum_value_norm	Maximum value from the salary derived from either the inferred or extracted field, depending on which is present
salary_maximum_annual_value_norm	Maximum annual value from the salary derived from either the inferred or extracted field, depending on which is present
salary_maximum_monthly_value_norm	Maximum monthly value from the salary derived from either the inferred or extracted field, depending on which is present
salary_currency_norm	Currency value from the salary derived from either the inferred or extracted value, depending on which is present
salary_recurrence_norm	Recurrence derived from either the inferred or extracted value, e.g., "month," "year"
release_date_exact_date_norm	Exact date derived from the release_date_extracted field
release_date_minimum_date_norm	Minimum date derived from the release_date_extracted field
release_date_maximum_date_norm	Maximum date derived from the release_date_extracted field
release_date_status_norm	The status derived from the release_date_extracted field, e.g., release," "update"
desired_experience_exact_years_norm	The exact years of experience derived from either the inferred or extracted desired_experience fields
desired_experience_minimum_y	The normalized minimum required years of experience derived from either the

Column name	Description
ears_norm	inferred or extracted desired_experience fields
desired_experience_maximum_years_norm	The normalized minimum required years of experience derived from either the inferred or extracted desired_experience fields
desired_experience_exact_level_norm	The normalized exact required level of experience derived from either the inferred or extracted desired_experience fields
desired_experience_minimum_level_norm	The normalized minimum required level of experience derived from either the inferred or extracted desired_experience fields
desired_experience_maximum_level_norm	The normalized maximum required level of experience derived from either the inferred or extracted desired_experience fields
desired_education_exact_level_norm	The normalized exact required level of education derived from either the inferred or extracted desired_education fields
desired_education_minimum_level_norm	The normalized minimum required level of education derived from either the inferred or extracted desired_education fields
desired_education_maximum_level_norm	The normalized maximum required level of education derived from either the inferred or extracted desired_education fields
location_continent_en_norm	The continent derived from the extracted or inferred location in English
location_continent_zh_norm	The continent derived from the extracted or inferred location in Chinese
location_country_en_norm	The country derived from the extracted or inferred location in English
location_country_zh_norm	The country derived from the extracted or inferred location in Chinese
location_province_en_norm	The province-level location derived from the extracted or inferred location in English
location_province_zh_norm	The province-level location derived from the extracted or inferred location in Chinese
location_prefecture_en_norm	The prefecture-level location derived from the extracted or inferred location in English
location_prefecture_zh_norm	The prefecture-level location derived from the extracted or inferred location in Chinese
location_county_en_norm	The county-level location derived from the extracted or inferred location in English
location_county_zh_norm	The county-level location derived from the extracted or inferred location in Chinese
location_township_en_norm	The township-level location derived from the extracted or inferred location in English
location_township_zh_norm	The township-level location derived from the extracted or inferred location in Chinese
location_zone_en_norm	The industrial zone derived from the extracted or inferred location in English
location_zone_zh_norm	The industrial zone derived from the extracted or inferred location in Chinese

Column name	Description
location_details_en_norm	Populated for any lower level locations that could not be disambiguated
location_details_zh_norm	Populated for any lower level locations 'that could not be disambiguated
location_code_norm	A unique code for the location, ISO 3166 2 for continents and countries and administrative division code for lower levels.
location_level_norm	The level of the location, to identify the list in which to look for more information about this location.
body_norm	A cleaned version of the "body_extracted" field
is_primary	Whether the job posting was chosen as the primary representative in a set of duplicates
url_norm	The URL of the job posting with query parameters removed
pipeline_date_updated	When the posting was last run through the pipeline

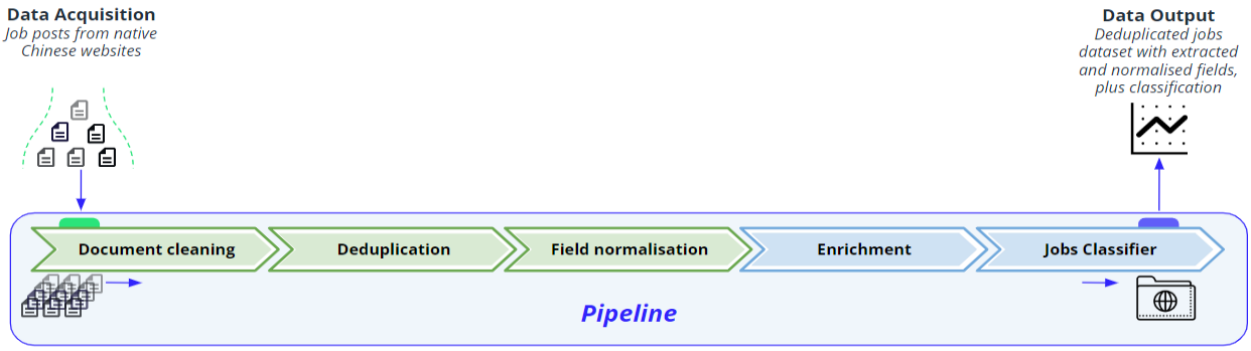
Source: AMPLYFI.

Appendix B: Data Collection and Model Classification

This appendix provides more detail on the process of compiling and curating the 6.8 million unique Chinese job postings, including model training and category assignment.

The process of finding and collecting AI job postings involved three general tasks: (1) harvesting job postings from Chinese job boards and removing duplicates; (2) classifying each relevant job posting into an AI or non-AI job category; and (3) normalizing the information pulled from those postings. This is illustrated in Figure B1.

Figure B1. Chinese Job Posting Data Acquisition Process



Note: Several tasks were developed simultaneously. This graphic shows the linear progression of a job posting through the completed pipeline from website harvest to data output.

Source: AMPLYFI.

Harvesting

The first step in the process was to find Chinese job postings and harvest them into a data lake. AMPLYFI deployed website crawlers to find and download a representative sample of web pages that displayed job vacancy announcements. The process of extracting job postings took place from February to July 2021.

Although four job boards were included, the vast majority, about 95 percent, of postings came from 51Job. This was due in part to the different measures each site had in place to detect and prevent web crawling and scraping, but also because 51Job was a leading site for technically oriented job postings in China. After the website was downloaded, the HTML pages were parsed to extract relevant information and populate the fields of the database.

Posting Deduplication

When harvesting job postings from Chinese websites, web crawlers sometimes downloaded the same posting more than once. To prevent double counting in the analysis, these duplicate postings needed to be captured and removed.

To address this, the AMPLYFI team conducted a two-step process. They first normalized URLs by isolating and removing suffixes, as the beginning of the string for each URL is identical in format but unique to each posting, leading to the extraction of a unique identifier for each unique posting.[74] They then compared these unique identifiers to identify duplicates and kept one copy. In case of content variations between the duplicates, the longest version of the posting was kept.[75]

The types of duplicates identified either were job postings listed several times, edited by the posting entity (employer) over time, or appeared under different sections of the websites. All of these resulted in variations to the end of the URLs that had to be isolated and removed to enable unique identifier comparison. All duplications discovered and removed were within the same job platform and not across platforms.[76]

Classifying AI Job Postings

In order to classify job postings into AI workforce categories, data is needed to train, test, and validate a classifier model. The two main components of the classification process are the creation of input data for the model and the creation of the model itself. These components, discussed briefly in the following sections, were constructed in parallel and with iteration.

Model Annotation

The process of generating training, testing, and validation data for the classifier model required creating a large set of labeled data. For this process, separate datasets of job postings for each category were created and annotated using different approaches.[77]

Three teams worked collaboratively throughout the annotation process: (1) the CSET team; (2) six hired AMPLYFI annotators; and (3) the AMPLYFI team. The AMPLYFI team facilitated communication between the first two teams, manually adjudicated particularly challenging job postings while working alongside the CSET team, and trained the six annotators.

Each dataset was prepared differently. The CSET team created what was ultimately used as the test set at the initial stage of annotation. To help train the AMPLYFI team, the CSET team began by annotating a small set of job postings that contained examples of positions from each category.[78] The AMPLYFI team then used this and the full list of CSET-defined AI occupations to have their annotators label another set of job postings in parallel with the CSET team. Working together, the teams compared the two labeled sets and resolved mislabeled postings, identifying job postings that had been labeled with high error rates and developing guidelines for labeling edge cases. The AMPLYFI team also adjudicated their annotation team's more challenging cases, at times with input from CSET.

For annotation, CSET and AMPLYFI team members reviewed each job posting's title, description, salary, company title and information, education requirement, and other relevant fields. Upon review of the available information and context, annotators assigned each posting to one of the four CSET-defined AI workforce categories, or to a fifth that was created to indicate that the posting did not fall into one of the other four categories. The work began shortly after initial harvesting and continued throughout the development of the classifier model.

The complexity and nuance of the AI-job categories added time to the annotation process. Each category comprised a wide range of occupations, and some occupations across categories shared key words or skills.[79] Further, some categories included certain occupations while omitting others that could appear similar.[80] And finally, some occupations were only included within relevant sectors (e.g., project managers were included only if the project was technical in nature). In some cases this was further complicated by imprecise translation (the CSET team worked primarily in English) or limited information within key fields.

Using CSET's test data and AMPLYFI-provided training, AMPLFYI's annotators labeled a dataset of about 5,500 job postings. The AMPLYFI team found that labeling job postings multiple times and conducting disagreement resolution exercises improved the performance of the classifier. Annotation stopped when the performance of the classifier stopped improving with additional labeled data.[81] The job postings selected for annotation were a mix of random and targeted samples presumed to be within each AI workforce category. This was necessary because of the uneven distribution of AI and non-AI job postings.

Of the 5,500 annotated job postings, about 4,800 were used for the final training set, and another 650 were used for model validation. The job postings used for the model test set were the 485 separately annotated by the CSET team. Measures of inter-annotator agreement (reliability) and agreement against the best-performing classifier model may be released in future documentation.

Critically, we note that areas with the most inter-annotator disagreement were also where the classifier ultimately had higher error rates. This mainly affected Product Team and Commercial Team job postings, which as a result are included in this analysis in a far more limited scope.

Model Classification

To categorize each job posting into one of the predefined AI workforce categories, the AMPLYFI team developed a novel machine learning classification model.[82] In parallel to the annotation process, the AMPLYFI team created multiple iterations of this classifier model, updated with newly annotated data and varied data input formats, parameter optimizations, and model architecture designs. Different parts of the classifier were developed and tested independently and iteratively. With each variation, the model was tested using commonly used metrics: accuracy, precision, recall, F1-score, and F0.5-scores were calculated to determine the highest-performing version.[83]

The AMPLYFI team decided to use a large open-source pretrained transformer model known as the Bidirectional Encoder Representation from Transformers (BERT)-base-Chinese model,[84] due to its effectiveness in text classification and the availability of models pretrained in Chinese.[85] Other models were tested but added no significant improvement, and the team used default hyperparameters for most testing before final tuning at the end of development.

The AMPLYFI team optimized the model for this custom application by varying four hyperparameters. These hyperparameters were: the number of times the model passed through the entire training set (epochs); the number of samples the model reviewed before updating its internal parameters (batch size); the maximum number of words a model was able to process at once, a constraint imposed by the model architecture itself (sequence length); and a mechanism for determining how fast the model learned to solve the problem, because of the tradeoff between fast learning and reaching an optimal solution (learning rate).

Once the tuning of hyperparameters was complete, the AMPLYFI team tried additional ways to improve classifier performance to ensure the best fit. This included merging job categories into different combinations, training on relevant fields in isolation and then in different combinations, altering token input formation and length, training on normalized fields instead of original values, integrating annotator disagreement in the model input, and discarding labels of which annotators were less confident. These experiments either confirmed previous assumptions or provided no increase in performance, providing the AMPLYFI with confidence in the final classifier model.

Field Normalization

Once the dataset was harvested, deduplicated, and each posting classified into AI job category, extracted fields needed to be standardized for analysis. Several fields of interest contained related information in differing formats, requiring a dedicated effort for creating standard values and formats. Fields that were normalized for analysis included desired experience, desired education, salary, posted date, and location. Details on the field normalization process for each variable are provided in Appendix C.

Appendix C: Field Normalization for Analysis

Across the extracted job postings contained in the corpus, field normalization—the process of standardizing fields into consistent values—was required across several fields. These included salary, posted date, desired experience, desired education, and location. Null values for the main normalized fields used in this report are provided in Table C1.

Table C1. Null Count for Fields Used in this Analysis

Field Title	Null Count	Percentage of Job Postings
desired_education_exact_level_norm	1,073,933	15.8
desired_experience_exact_level_norm	431,263	6.3
salary_minimum_value_norm	270,192	4.9
salary_minimum_annual_value_norm	439,632	6.5
salary_maximum_value_norm	262,768	3.9
salary_maximum_annual_value_norm	432,210	6.3
location_province_en_norm	125,860	1.8
location_province_zh_norm	125,860	1.8

Source: CSET and AMPLYFI Chinese job postings dataset.

For each of these fields, the AMPLYFI team attempted to fill null values. Specifically, when the education, experience, or salary fields were null, AMPLYFI attempted to fill them using information from the job description field. AMPLYFI performed regular expression (RegEx) searches on the description—similar to those used to extract values from other fields—to fill in missing information about education, experience, and salary. For example, when the education field was null, AMPLYFI searched the description field for "博士" to identify additional job postings that desired a doctoral

degree. AMPLYFI used a different method to fill null values in the geographic location fields; further details are provided in that subsection.

Extracted information within the corpus was also pulled in more than one language. Therefore language detection and distribution analysis was an important first step. The AMPLYFI team was able to detect which languages were used and at what frequency. They determined that about 97 percent of the job postings were mainly Chinese, with more than half containing some English words or phrases, and more than 2 percent bilingual in Chinese and English. The high proportion of postings in Chinese allowed the AMPLYFI team to focus models solely on Mandarin, removing the need for multilingual processing of job postings.

Desired Education

The desired education for each job posting was provided as either an exact, maximum, or minimum level of educational attainment. About 84 percent of postings were in the exact category. Table C2 gives an overview of Chinese educational levels, which differ from the U.S. system at the secondary level.

Table C2. Chinese Primary and Secondary Education Levels

High-level Chinese education trajectory	
Age	School
3-6	Pre-school education
6-12	Primary education
12-18	Secondary education
>18	Higher education
Chinese secondary education	
Age	School
12-15	Junior middle school
15-18	Senior middle school (if students pass the entrance exam) Specialized / Vocational / Technical middle school (if students don't pass the entrance exam)

Source: AMPLYFI.

After secondary school, the options for postsecondary education are similar to those in the American system. However, there is no direct equivalent of the U.S. community college system. Instead, there are two- to three-year specialized or vocational-technical colleges that are similar to U.S. technical colleges and career and technical education programs at community colleges granting the equivalent of a U.S. associate's degree.

When extracting education requirements from job postings, there was a set number of specific Chinese phrases within the corpus. These phrases and their English equivalents are shown in Table C3. The phrases "junior middle school," "specialized/vocational/technical middle school," and "senior middle school" were all normalized to "middle school" in the final version of the field.

Table C3. Extracted Chinese Characters and Their English Translations

Extracted Education Requirement in Chinese	English Translation of Education Requirement
初中	junior middle school degree
中技	specialized/vocational/technical high school degree
中专	specialized/vocational/technical high school degree
中专/中技	specialized/vocational/technical high school degree
高中	high school degree
大专	associate's degree
本科	bachelor's degree
硕士	master's degree
MBA	master's degree
EMBA	master's degree
MBA/EMBA	master's degree
博士	doctoral degree

Source: AMPLYFI.

Similar to how the desired experience field was handled, using the description field to fill in missing values fairly effective. The process filled in 22 percent of missing values in the desired education field.

Desired Experience

Desired experience can be expressed either in years required or in level. In both cases, it can be an exact value or a range of values. The normalization process for values is straightforward, and Table C4 shows the conversion used between formats.

Table C4. Desired Experience Range Normalization from Years to Levels

Range	Level	Synonyms/Translations
0 years	No experience	在校生/应届生 (current/graduate students), 无需经验 (no experience required), 应届 (last year student), last year student
0-5 years	Junior	一年以下 (less than 1 year), 实习 (internship), internship, entry, associate, intermediate
5-15 years	Mid	Mid level (also numerical values from 5 to 15 years)
>15 years	Senior	lead, executive, director

Source: AMPLYFI.

Using the description field to fill in missing values was more effective for this field than for salary, discussed next. This strategy filled in 17 percent of all missing values.

Salary

The salary field required significant normalization in order to conduct meaningful analysis. In general, the job posting listed a value or a range of values (e.g., minimum to maximum), a currency, and a recurrence. The AMPLYFI team rescaled values to monthly and annual rates for easier comparison across job postings.

Normalization proved a difficult task because salaries were provided by year, month or hour, and sometimes contained errors or lacked pieces of information. Variable weekly schedules further complicated conversion from hourly or daily salaries to monthly and yearly. The weekly Chinese standard is 40 hours, similar to the U.S. work week, but some start-ups follow the 9-9-6 system where people work 72 hours a week.

An investigation into the hourly and daily salaries showed that not only was it not clear which hourly rate the job paid, but some values were also clearly wrong. Some postings had a very high value, for instance 5,000 yuan per hour or 15,000 yuan per day for low qualification jobs. Reading the descriptions indicated that in those cases the value was not correct, but rather a way to show that the company was willing to pay well. Further, just over 170,000 postings, or 1.65 percent of the corpus, contained no payment frequency (e.g., hourly, annually), and slightly more than 3 percent of the postings did not list a salary. Future analysis may include imputed salary data for missing values.

Geographic Location

The normalization of the job location field required several steps. China uses administrative divisions to describe geographic areas at different levels of granularity. From largest to smallest, the divisions relevant to analysis are: provincial, prefectural, county, and township levels.

A few locations do not fit perfectly in the administrative division system: the National Economic and Technological Development Zones, the National High-Tech Industrial Development Zones and the Border Economic Cooperation Zones. Sometimes these zones also have their own administrative division, but this can be on the county level as well as the township level. These zones are very interesting because they likely have a different job vacancy landscape from their geographical environment, since they have a high concentration of technical jobs. Because of this, AMPLYFI created a separate category for these zones.

Job postings did not always specify the same level of administrative division. To rectify this issue, AMPLYFI developed a process to disambiguate locations to all levels above the one given. For example, if the location 九江 (Jiujiang) is given at the county level, we need the prefecture level list to disambiguate this to 九江市 (Jiujiang City) and that list can then tell us that Jiujiang City lies in Jiangxi Province. This allows for analysis at higher levels of administrative zones.

This process provided normalized fields for each job posting for every available administrative zone in both Chinese and English. Table C5 demonstrates with an example.

Table C5. Example of Geographic Normalization by Administrative Zone

Beilun District, Zhejiang, China	
Field Name	Field Content
Continent EN	Asia
Continent ZH	亚洲
Country EN	China
Country ZH	中国
Province EN	Zhejiang Province
Province ZH	浙江省
Prefecture EN	Ningbo City
Prefecture ZH	宁波市
County EN	Beilun District
County ZH	北仑区
Level	county
Code	330206

Source: AMPLYFI.

Appendix D: AI Keyword List

Table D1 provides the keyword list used for the AI-specific analysis of Technical Team 1 and Technical Team 2 job postings. The list is provided in English and Chinese; analysis was done in Chinese, as the origin language of the posting. The list was developed using CSET staff expertise and a literature review of research that similarly has attempted to identify AI-specific data in areas such as grants and patents. Some words were intentionally stemmed (e.g., "recog") for more inclusive analysis.

Table D1. AI Keyword List

English Term	Chinese Term
artificial intelligence	人工智能
\bAI\b	
machine learning	机器学习
supervised learning	监督学习
unsupervised learning	无监督学习
reinforcement learning	强化学习
deep learning	深度学习 or 深层学习
neural network	神经网络
computer vision	计算机视觉
natural language processing	自然语言处理
facial recog	人脸识别 or 面部识别 or 面像识别 or 面容识别 or 面孔识别
biometric	生物识别 or 指纹识别
robotics	机器人技术 or 机器人
augmented reality	增强现实
autonomous	自主性
human-machine	人机 or 人机交互

English Term	Chinese Term
learning algorithm	学习算法
classification algorithm	分类算法
image recog	图像识别
data mining	数据挖掘
text mining	文本挖掘
topic model	主题建模 or 主题模型
sentiment analysis	情感分析 or 评论分析 or 评价分析
opinion mining	意见挖掘 or 观点挖掘
pattern recog	模式识别

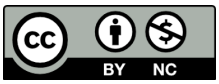
Source: CSET.

Authors

Diana Gehlhaus is a research fellow at the Center for Security and Emerging Technology, where Sara Abdulla and Jacob Feldgoise are data research analysts, and Luke Koslosky and Dahlia Peterson are research analysts. Joanne Boisson is a senior machine learning engineer at AMPLYFI.

Acknowledgments

The authors thank Catherine Aiken, Igor Mikolic-Torreira, Emily Weinstein, Anna Puglisi, Ngor Luong, Owen Daniels, Amy Chao and Christine Cleary DuFour for their valuable feedback and assistance. They are grateful to Howard Shatz and Erik Britton for providing most insightful reviews, to Daniel Chou for his critical data expertise, and to Ben Murphy for his translations and Chinese character verification. They also thank Oliver Hayman at AMPLYFI for his dedicated management of this collaboration, and Luis Espinosa-Anke for his technical expertise as AMPLYFI's head of machine learning. Finally, we thank Dale Brauner for fact-checking and the outstanding editorial support provided by CSET's external affairs team.



Document Identifier: doi: 10.51593/20220003

Endnotes

[1] These teams are broadly defined to include technical occupations that "could be" as opposed to "actively are" working in AI. Technical Team 1 is more directly related to AI, including most computer and mathematical science occupations. Technical Team 2 focuses more on engineering and science occupations.

[2] Remco Zwetsloot, James Dunham, Zachary Arnold, and Tina Huang, "Keeping Top AI Talent in the United States" (Center for Security and Emerging Technology (CSET), December 2019). See also Jack Corrigan, James Dunham, and Remco Zwetsloot, "The Long-Term Stay Rates of International STEM PhD Graduates" (CSET, April 2022).

[3] Graham Allison and Eric Schmidt, " The U.S. Needs a Million Talents Program to Retain Technology Leadership," *Foreign Policy*, July 16, 2022, https://foreignpolicy.com/2022/07/16/immigration-us-technology-companies-work-visas-china-talent-competition-universities/.

[4] Dahlia Peterson, Kayla Goode, and Diana Gehlhaus, "AI Education in China and the United States: A Comparative Assessment" (CSET, September 2021).

[5] For example, the U.S. Bureau of Labor Statistics tracks job postings explicitly for this purpose, and publishes a monthly update. For more, see: https://www.bls.gov/jlt/jltprovq.htm.

[6] 51Job is one of the top 100 internet companies in China and was listed on Nasdaq in 2004. Many well-known firms and Fortune 500 companies are recruiting talent and posting positions on 51Job. Every week, users upload more than 50 million resumes to the platform, including for jobs related to internet and digital industries. For more, see "找工作软件排行榜2022 国内十大找工作app排行榜" (top 10 job search apps in China in 2022), 安粉丝 (Anfensi, January 13, 2022), https://perma.cc/CK9K-Z23P.

[7] Zachary Arnold et al., "Using Machine Learning to Fill Gaps in Chinese AI Market Data" (CSET, February 2021).

[8] Original CSET translation, "AI Innovation Action Plan for Institutions of Higher Education" (教育部关于印发《高等学校人工智能创新行动计划》的通知), Chinese Ministry of Education, April 2, 2018, https://cset.georgetown.edu/research/ai-innovation-action-plan-for-institutions-of-higher-education.

[9] Original CSET translation, "AI Innovation Action Plan."

[10] A forthcoming CSET publication will provide more detail on China's AI Institutes.

[11] Rogier Creemers, Hunter Dorwart, Kevin Neville, Kendra Schaefer, et al., "Translation: 14th Five-Year Plan for National Informatization, Dec. 2021," DigiChina at Stanford University, January 24, 2022,

https://digichina.stanford.edu/work/analyzing-chinas-2021-2025-informatization-plan-a-digichina-forum/.

[12] Creemers, Dorwart, Neville, Schaefer, et al., "Translation."

[13] Zou Shuo, "AI now most favored major at universities," *China Daily*, March 3, 2021, https://archive.ph/jKNYz.

[14] The colleges offer the MOE-approved Artificial Intelligence major, along with other standardized majors such as Data Science and Big Data Technology and Intelligent Science and Technology. A forthcoming CSET publication will provide more detail on China's AI Institutes.

[15] Chinese Ministry of Education, National Development and Reform Commission, and Ministry of Finance, "关于"双一流'建设高校促进学科融合 加快人工智能领域研究生培养的若干意见" ("Notice on the Publication of "Certain Opinions on Promoting Curricula Merging at 'Double First-Class' Institutes of Higher Education and on Accelerating the Cultivation of Graduate Students in the AI Field") January 21, 2020, https://archive.vn/f7iGx. See CSET's translation at https://cset.georgetown.edu/publication/notice-on-the-publication-of-certain-opinions-on-promoting-curricula-merging-at-double-world-class-institutes-of-higher-education-and-on-accelerating-the-cultivation-of-graduate-students-in-the/.

[16] Peterson, Goode, and Gehlhaus, "AI Education in China and the United States: A Comparative Assessment."

[17] "Certain Opinions on Promoting Curricula Merging at 'Double First-Class' Institutes of Higher Education and on Accelerating the Cultivation of Graduate Students in the AI Field."

[18] The central government also requests that companies train graduate students by having them solve industry needs. For PhD students, enterprises are encouraged to open up "scenario-driven," application-oriented courses, as well as their data, case studies, tools, and training platforms. Enterprises can also utilize industry alliances, joint R&D labs, entrepreneurship and skills competitions, and certification training to help graduate students grow in the field. Instructors are asked to incorporate the latest AI research findings into PhD courses.

[19] Peterson, Goode, and Gehlhaus, "AI Education in China and the United States: A Comparative Assessment."

[20] Universities are encouraged to coordinate various resources, such as financial investment and scientific research income, and to obtain angel investment and venture capital financing, to boost major AI projects and applied research within universities while assisting with talent training.

[21] Peterson, Goode, and Gehlhaus, "AI Education in China and the United States: A Comparative Assessment."

[22] National Bureau of Statistics of China, "China Statistical Yearbook, 2021" (Beijing: China Statistics Press), http://www.stats.gov.cn/tjsj/ndsj/2021/indexeh.htm.

[23] See, for example, "China Artificial Intelligence Talent Training Report," an issue group at the Institute of China's Science, Technology and Education Policy, Zhejiang University and Baidu, January 2022.

[24] See "China's top talent trends for 2020: tech innovation to drive job market," Hays, https://www.hays-china.cn/en/press-release/content/-2020-?s=d3d3LmhheXMuY24=; A. J. Cortese, "China's Tech Pivot (Part II): STEM Talent Shortage Stymies Core Innovation?" Macro Polo, May 10, 2022, https://macropolo.org/chinas-tech-pivot-stem-talent-shortage/?rp=e.

[25] China Information Security Evaluation Center (中国信息安全测评中心), "Research Report on the Status of China's Information Security Professionals (2018–2019)" (中国信息安全从业人员现状调研报告 [2018-2019] 年度)," translated Center for Security and Emerging Technology (CSET), https://cset.georgetown.edu/publication/research-report-on-the-status-of-chinas-information-security-professionals-2018-2019/.

[26] That is, one person for every 10 openings for AI engineers and technicians. This talent is defined by the Ministry as "engineering and technical personnel engaged in the analysis, research and development of artificial intelligence–related algorithms, deep learning and other technologies, and design, optimization, operation and maintenance, management and application of artificial intelligence systems." For more, see Ministry of Human Resources and Social Security, "新职业——人工智能工程技术人员就业景气现状分析报告" ("New Occupations — Analysis Report on Artificial Intelligence Engineers' Current Employment Status"), April 30, 2020, https://archive.ph/Dvk4p.

[27] Ministry of Human Resources and Social Security, "New Occupations."

[28] Ministry of Industry and Information Technology Talent Exchange Center (工业和信息化部人才交流中心), Artificial Intelligence Industry Talent Development Report (2019-2020 Edition) [人工智能产业人才发展报告（2019-2020年版）], translated by the Center for Security and Emerging Technology (CSET), https://cset.georgetown.edu/publication/artificial-intelligence-industry-talent-development-report-2019-2020-edition/.

[29] Kai Shen, Xiaoxiao Tong, Ting Wu, and Fangning Zhang, "The next frontier for AI in China could add $600 billion to its economy," McKinsey, June 2022, https://www.mckinsey.com/business-functions/quantumblack/our-insights/the-next-frontier-for-ai-in-china-could-add-600-billion-to-its-economy.

[30] Bloomberg News, "China's Youth Unemployment Spikes as Students Graduate," August 16, 2021, https://www.bloomberg.com/news/articles/2021-08-16/china-s-youth-unemployment-spikes-as-students-graduate; Martin Quin Pollard, "Analysis: Record numbers of Chinese graduates enter worst job market in decades," Reuters, June 23, 2022, https://www.reuters.com/world/china/record-numbers-chinese-graduates-enter-worst-job-market-decades-2022-06-23/.

[31] Yaling Jiang, "Job crisis looms for Chinese graduates as tech, new energy vehicle firms stop hiring amid Covid-19 lockdowns," *South China Morning Post*, May 2022, https://www.scmp.com/tech/big-tech/article/3178963/job-crisis-looms-chinese-graduates-tech-new-energy-vehicle-firms-stop.

[32] Jinshan Hong and Spe Chen, "China Lacks the Right Workers to Boost Xi's Favored Tech Jobs," Bloomberg, September 27, 2022, https://archive.ph/pUrg5#selection-4215.0-4215.61.

[33] Statistics here include only "regular" institutions of higher education, not adult- or web-based ones. Because China's Ministry of Education does not oversee private or provincial institutions, nor the Seven Sons of National Defence institutions, all recipients of bachelor's degrees are not included in this overview. See the official statistics from China's Ministry of Education at http://www.moe.gov.cn/jyb_sjzl/moe_560/2020/quanguo/202108/t20210831_556364.html.

[34] Laura He, "China is encouraging college graduates to work in the countryside," CNN Business, June 13, 2022, https://www.cnn.com/2022/06/13/economy/china-youth-unemployment-countryside-recruitment-intl-hnk/index.html. The struggle with college graduate underemployment is not new. See Ralph Jennings, "China's ant tribe' poses policy challenge for Beijing," Reuters, February 17, 2010, https://www.reuters.com/article/us-china-middleclass/chinas-ant-tribe-poses-policy-challenge-for-beijing-idUSTRE61H01220100218.

[35] In 2022, an estimated record 10.8 million bachelor's degrees were awarded in China, and another 11.9 million youth took the notoriously difficult national college entrance exam. For more, see the official enrollment data, National Bureau of Statistics of China, "China Statistical Yearbook, 2021."

[36] Peterson, Goode, and Gehlhaus, "AI Education in China and the United States: A Comparative Assessment."

[37] Remco Zwetsloot et al., "China Is Fast Outpacing U.S. STEM PhD Growth" (CSET, August 2021).

[38] 猎聘网 (Liepin), "2019年中国AI&大数据人才就业趋势报告" ("2019 China AI & big data talent employment trend report"), https://perma.cc/3Y9W-2WKZ.

[39] Corrigan, Dunham, and Zwetsloot, "The Long-Term Stay Rates of International STEM PhD Graduates."

[40] For example, one study on the effectiveness of China's "Youth 1000 Talents Program" found that while recruiting "top-tier" researchers back to China remains elusive, recruiting "next best" young researchers struggling to secure funding in the United States and European Union has been relatively effective. See Dongbo Shi, Weichen Liu, and Yanbo Wang, "Has China's Young Thousand Talents Program Been Successful in Recruiting and Nurturing Next-generation Chinese Scientists?" (SSRN, July 12, 2022), https://ssrn.com/abstract=4043516. In response to this program, some in the United States are calling for a reciprocal U.S. program. See Allison and Schmidt, "The U.S. Needs a Million Talents Program to Retain Technology Leadership.".

[41] Emily Weinstein, "Mapping China's Sprawling Efforts to Recruit Scientists," DefenseOne, November 30, 2020, https://www.defenseone.com/ideas/2020/11/mapping-chinas-sprawling-efforts-recruit-scientists/170373/.

[42] The BOSS (Zhipin) homepage included a section for overseas returnees, at times with a caption advertising "exclusive spot for high paying jobs for returnees." https://perma.cc/JQ98-349X.

[43] This is not a statement about the relative quality of the talent, an aspect that is not addressed here. With that said, more Chinese universities are gaining standing in international rankings. For more on stay rates for Chinese STEM PhDs in the United States, see Corrigan, Dunham, and Zwetsloot, "The Long-Term Stay Rates of International STEM PhD Graduates."

[44] Diana Gehlhaus and Santiago Mutis, "The U.S. AI Workforce: Understanding the Supply of AI Talent" (CSET, January 2021).

[45] It is a widely used measure in economics. For example, the U.S. Bureau of Labor Statistics tracks job postings explicitly for this purpose, and publishes a monthly update: https://www.bls.gov/jlt/jltprovq.htm.

[46] As noted in an earlier section, China's MIIT Talent Exchange Center also estimates and evaluates AI occupation job vacancy rates for this reason.

[47] See Gehlhaus and Mutis, "The U.S. AI Workforce." Some characteristics of the background demanded, such as educational attainment and experience, are likely to reflect talent supply.

[48] For more on CyberSeek, see https://www.cyberseek.org/. The number of job openings in cybersecurity based on this tool is routinely used for cybersecurity workforce policy. For example, see https://www.whitehouse.gov/oncd/briefing-room/2022/10/03/office-of-the-national-cyber-director-requests-your-insight-and-expertise-on-cyber-workforce-training-and-education/.

[49] ChinaHR (中华英才网), Zhaopin (智联招聘), Zhipin (直聘), and Lagou were also considered, but were not included due to their implemented restrictions and safeguards against web scraping. This was likely in part due to new, stricter privacy rules on these Chinese job boards, imposed by the Beijing Municipal Human Resources and Social Security Bureau in May 2021 after a large-scale data breach ("Regulations on the Administration of Online Recruitment Services"). After this breach, the Bureau summoned executives from Zhaopin and Liepin, and additional reports at the time suggest that 51Jobs was also affected. For more, see http://rsj.beijing.gov.cn/xwsl/mtgz/202103/t20210322_2313339.html; http://rsj.beijing.gov.cn/xwsl/mtgz/202103/t20210322_2313341.html; and

http://rsj.beijing.gov.cn/xwsl/mtgz/202104/t20210416_2360300.html. However, an updated harvest will include BOSS (Zhipin) and ChinaHR using a modified data collection approach.

[50] "Top 10 job search apps in China in 2022."

[51] Because of this study's broad interpretation of what constitutes "AI" job postings, which include all positions with the ability to design, develop, and deploy AI, the number of jobs actively working in AI is overstated. To address this, the study later provides data for positions that explicitly mention AI.

[52] Levels of education in China share many similarities to those in the United States, but there are some differences in secondary (middle and/or high school) education. From age 12 to 15, students attend junior middle school. From 15 to 18, if they pass an entrance exam they attend senior middle school; otherwise they go to an occupational-based middle school. While there is no one-to-one equivalent in China of the U.S. community college system, there are two- to three-year specialized or vocational technical colleges that are similar to U.S. technical colleges and to career and technical education at community colleges. For more information on Chinese post-secondary institutions, see Gerard Postiglione, "Community Colleges in China," *International Higher Education*, March 25, 2015, https://www.researchgate.net/publication/313411891_Community_Colleges_in_China.

[53] When data was missing from the "desired education" field, AMPLYFI attempted to fill those cells with data from the general job description field. Validation checks were conducted, and it was determined that including the additional data did not change by more than 0.4 points any of the percentage levels reported in Figure 1 and Table 2.

[54] Significance testing was also run for pairings between Technical Team 1 and each other group, for both bachelor's and associate's degrees, using two-sample Z-tests. In addition, two other pairings were tested with similar proportions: share of associate's degrees for Product Team versus Technical Team 2, as well as share of bachelor's degrees for Product Team versus Commercial Team. All pairwise differences were statistically significant except for associate's degrees between Technical Team 2 and Product Team. The Bonferroni Correction was applied to each significance test to compensate for the number of tests run.

[55] Gehlhaus and Mutis, "The U.S. AI Workforce."

[56] Appendix C provides more detail on using specified years of experience to define entry-, junior-, mid-, and senior-level positions.

[57] Significance testing also showed that differences in the share of jobs requiring no experience for non-AI postings relative to each category of AI job postings were statistically significant. Endnote 54, above, provides more detail on the approach.

[58] Felicia F. Tian, "Is guanxi unfair? Market reform and the public attitude toward guanxi in urban China," *Journal of Chinese Sociology* (December 22, 2020), https://journalofchinesesociology.springeropen.com/articles/10.1186/s40711-020-00138-3.

[59] Nancy DiTomaso and Yanjie Bian, "The Structure of Labor Markets in the U.S. and China: Social Capital and Guanxi," *Management and Organization Review* (March 28, 2018), https://www.cambridge.org/core/journals/management-and-organization-review/article/structure-of-labor-markets-in-the-us-and-china-social-capital-and-guanxi/335D57AF58C845208A83E60991F5A2BA.

[60] Jiang, "Job crisis looms for Chinese graduates as tech, new energy vehicle firms stop hiring amid Covid-19 lockdowns."

[61] Appendix C provides more information on the normalization process.

[62] Significance testing also showed that differences in median salary were statistically significant between non-AI and Technical Team 1 jobs, as well as between non-AI and Technical Team 2 jobs. This was true for median minimum and median maximum salaries. To test differences in median values, the independent two-sample Wilcoxon rank sum test was applied.

[63] All Chinese yuan to U.S. dollar purchasing power conversions were calculated using the 2021 exchange rate of 4.187, retrieved from the Organization for Economic Co-operation and Development (OECD). For more information, see https://data.oecd.org/conversion/purchasing-power-parities-ppp.htm#:~:text=Purchasing%20power%20parities%20(PPPs)%20are,in%20price%20levels%20between%20countries. Other PPP measures are available, but this one was selected because of its widely accepted and used status.

[64] Although not directly comparable, a reference point may be insightful. Technical Team 1 closely aligns with computer and mathematical occupations as defined by the U.S. Standard Occupation Classification (SOC) system, minus web designers, web developers, and actuaries. In the United States in 2021, median earnings for computer and mathematical occupations were $97,540. See for more https://www.bls.gov/oes/current/oes150000.htm.

[65] For example, the annual salary range for Technical Team 1 tracks closely with official salary statistics for the "information, software, and IT," and for "scientific research and technical services" sectors as published by the Chinese government in its 2021 Annual Statistical Yearbook. For more, see National Bureau of Statistics of China, "China Statistical Yearbook, 2021," Table 4-11.

[66] National Bureau of Statistics of China, "Communiqué of the Seventh National Population Census (No. 3)," May 11, 2021, http://www.stats.gov.cn/english/PressRelease/202105/t20210510_1817188.html.

[67] PR Newswire, "Top 10 Chinese provincial regions with strongest GDP in 2020," April 2021, https://www.prnewswire.com/news-releases/top-10-chinese-provincial-regions-with-strongest-gdp-in-2020-301270482.html. For official numbers, see estimates for 2020 as reported by the National Bureau of Statistics of China at https://data.stats.gov.cn/english/easyquery.htm?cn=E0103.

[68] Using the OECD PPP 2020 conversion rate of 4.179. For China's 2020 per capita GDP, see https://fred.stlouisfed.org/series/NYGDPPCAPKDCHN.

[69] Significance testing showed that these differences were statistically significant between Technical Team 1 job postings with an AI keyword versus without. Endnote 54, above, provides detail on the approach.

[70] A corresponding wide range of fields of study were cited that were very aligned with the nature of the associated AI application.

[71] Publication in highly cited journals is incredibly difficult, and in many cases involves affiliation with leading institutions, academics, or collaborators. In effect, some of these hiring organizations discriminate in very explicit ways.

[72] Ngor Luong, Zachary Arnold, and Ben Murphy, "Understanding Chinese Government Guidance Funds" (CSET, March 2021).

[73] Ryan Fedasiuk, Jennifer Melot, and Ben Murphy, "Harnessed Lightning: How the Chinese Military Is Adopting Artificial Intelligence" (CSET, October 2021).

[74] The URLs of the set of identified duplicates differed only in the last segment of their strings.

[75] Out of an abundance of caution, the AMPLYFI team created a separate deduplicated dataset of job postings, leaving the initial non-deduplicated set intact for data quality checks.

[76] The lack of cross platform duplicates identified may be in part the result of the harvesting approach, in which only a fraction of the daily postings were downloaded to limit risk of anti-scraping measures. The AMPLYFI team did conduct a separate manual and semi-automated review of postings across platforms to validate the non-duplicity of job postings. While this study does not claim that employers posted a given job vacancy only on one website, no duplicates were found in this check.

[77] The larger training dataset is used to train the classification model. The validation set, or development set, is used to iteratively test performance of the model while adjusting model parameters to increase performance. The test set is kept away from the training of the model and is used as the final unbiased benchmark of model performance.

[78] An intentional mix of job postings selected at random and not at random. Some were not selected at random because of the relatively small share of jobs that were likely considered part of the AI workforce, in order to include postings across all categories. Moreover, given the nuance of category definitions, some postings were intentionally selected as edge cases for the model.

[79] The skill "data analytics" can appear in occupations across Technical Teams 1 and 2, and the proper label depends on whether or not the role requires advanced statistical techniques. Engineers also come

in many varieties; some fall under Technical Team 1, others under Technical 2, still others under Commercial Team, while some are not in the AI workforce at all.

[80] For example, Technical Team 2 includes mechanical engineers while omitting civil engineers, and the Commercial Team includes sales engineers while omitting most all other sales-based occupations.

[81] Discussion of how this was determined is provided in the section on classification.

[82] The classifier actually generates a predicted probability distribution across all five categories along with a predicted label, but functionally this amounts to sorting each posting to the category with the highest probability.

[83] Accuracy measures the number of correct predictions made by the model over all predictions made. An F0.5-score is similar to an F1-score. While an F1-score equally weights precision and recall, an F0.5 puts more weight on precision and less on recall. This means that it values how many of the labels within the category are correctly labeled more than do job postings left out of the correct category.

[84] This is using a formal model naming convention that identifies types of language processing. For more, see Qing Xie, Xinyuan Zhang, Ying Ding, and Min Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *Journal of Informetrics* 14, no. 3 (2020), https://doi.org/10.1016/j.joi.2020.101055;

Wei-Lin Liao and Wei-Yun Ma, "Roof-BERT: Divide Understanding Labour and Join in Work," ArXiv, December 2021, https://arxiv.org/pdf/2112.06736.pdf.

[85] A transformer is a type of recurrent neural network architecture that differentially weighs the significance of each part of the input data and is particularly suited for tasks such as natural language processing.