Research Report

JON SCHMID, PRATEEK PURI, VITOR MELO, ANTHONY HAKIM

AI Development in the United States and China

Evidence from a New Dataset of AI Developer Firms

For more information on this publication, visit **www.rand.org/t/RRA5043-1**.

Published by the RAND Corporation, Santa Monica, Calif.

RR-A5043-1

About This Report

The emerging U.S. and Chinese commercial artificial intelligence (AI) developer sectors remain poorly specified. The aim of our study was to systematically characterize the commercial AI ecosystems in the United States and China and thus broaden the empirical foundation for business intelligence and policy analysis. To this end, we constructed a novel dataset of AI developers headquartered in the United States and China that meet a set of innovation criteria. We then analyzed this dataset, focusing on three categories of information: basic descriptive information about the population of U.S. and Chinese AI developers, the commercial orientation of firms in each country, and the technical approaches taken by firms in each country. This report fills an important gap in the literature on explaining the modern AI development ecosystem, a field that has largely relied on anecdotal evidence or official policy statements rather than systematic cross-national comparison of the firms shaping advanced AI. It should be of interest to policymakers and researchers studying innovation systems and technology competition between the United States and China.

Center on AI, Security, and Technology

RAND Global and Emerging Risks is a division of RAND that delivers rigorous and objective public policy research on the most consequential challenges to civilization and global security. This work was undertaken by the division's Center on AI, Security, and Technology, which aims to examine the opportunities and risks of rapid technological change, focusing on artificial intelligence, security, and biotechnology. For more information, contact cast@rand.org.

Funding

This research was independently initiated and conducted within the Center on AI, Security, and Technology using income from operations and gifts and grants from philanthropic supporters. A complete list of donors and funders is available at www.rand.org/CAST. RAND clients, donors, and grantors have no influence over research findings or recommendations.

Acknowledgments

The authors thank two peer reviewers—Igor Mikolic-Torreira and Shawn Chao—whose careful and insightful comments on an earlier draft of this document led to substantive improvements. Comments by Jason Etchegaray and Kyle Siler-Evans also improved the document.

Summary

We constructed a new dataset of 1,181 artificial intelligence (AI) developer firms headquartered in the United States and China. Using open-source intelligence, we gathered basic descriptive information for each firm along with detailed information on commercial posture and technical approach to AI development. Analysis of this dataset yields a rich description of the AI development ecosystem in each country, which we present in the body of this report. Synthesis of these findings yields three key insights.

First, the two AI developer firm populations studied in our sample are structurally more similar than commonly assumed. On most dimensions—e.g., model architecture, commercial positioning, foundation-model development, and learning paradigm—U.S. and Chinese AI developer firm ecosystems closely resemble each other. Both are transformer-led, both have roughly 20 percent foundation-model developers, and the vast majority of firms in both countries build applied AI products rather than operating model-as-a-service or inference-as-a-service delivery models.

Second, the largest divergence is physical embodiment versus software. Sixty-one percent of U.S. firms are software-only, compared with just 26 percent of Chinese firms. Chinese firms are heavily concentrated in embodied form factors (such as humanoid robots, ground robots, and autonomous vehicles) and in industry verticals (such as manufacturing, transportation, and energy). U.S. firms concentrate in knowledge-intensive, software-delivered verticals, such as health care, scientific research, and cybersecurity. The two ecosystems are, by these measures, building AI for different environments and domains of application.

This finding speaks to the ongoing debate over whether the United States is insufficiently diversified in its approach to AI. In the AI developer firm ecosystem, the strongest version of the concern—that the United States pursues a single architectural approach while China pursues many—is not supported by our data. However, should the path to artificial general intelligence ultimately require codevelopment with robotics, world models, and physical-environment learning, the comparatively broad embodied AI portfolio that we observe from Chinese firms represents a meaningful hedge that the U.S. ecosystem lacks.

Finally, we analyzed usage data from OpenRouter, a model routing provider, to assess adoption patterns of Chinese and U.S. models among developers. In this population, Chinese open-weight models rose from near zero to parity with U.S. models in token volume between late 2024 and April 2026, likely driven by Chinese open-weight model releases that offer frontier-like benchmark performance at roughly one-quarter to one-fifth the price of leading U.S. proprietary models. Because OpenRouter structurally overrepresents open-weight models and its user base skews toward technically sophisticated, cost-sensitive developers, this finding is best interpreted as a directional signal about open-weight developer adoption in this specific population rather than a measure of broader global market share.

Although many of our findings have been noted elsewhere, these claims have been supported by anecdote. In this work, we provide empirical grounding, rooted in open-source intelligence, for discussions of evolving U.S.-China AI competition and ecosystem development.

Three limitations bear noting when interpreting these findings. First, the dataset captures a snapshot of each ecosystem as of January–February 2026; given the pace of change in the AI industry, some firm-level classifications might no longer reflect activities. Second, all data are drawn from publicly available sources, which have uneven coverage across the two countries: Chinese firms have higher rates of missing information, meaning the counts for named categories for these firms should be treated as lower bounds. Third, the dataset encompasses only commercial AI developer firms; universities, government research centers, and other non-firm actors are excluded from scope.

Contents

Available at www.rand.org/t/RRA5043-1

Annex

AI Development in the United States and China: Annex

Figures and Tables

Figures

Tables

Chapter 1

Introduction

The emerging U.S. and Chinese commercial artificial intelligence (AI) developer sectors remain poorly characterized. Significant churn in the sector driven by entry, exit, and reorientation has resulted in the persistence of major gaps in the understanding of industry composition in both countries. In this data-poor environment, much of what is claimed about the commercial AI ecosystems in the United States and China is derived from the activities of a few high-profile developers and policy priorities articulated in national planning documents.

For example, in the United States, leading AI developers—such as OpenAI, Google, Anthropic, Meta, and xAI—have collectively invested tens of billions of dollars in the scaling paradigm, in which performance improvements are achieved by training increasingly large language models (LLMs) on vast datasets using massive computational resources. This concentration of effort around transformer-based architectures has led some observers to argue that U.S. AI development is insufficiently diversified across alternative model architectures and approaches.[1]

Similarly, China's policy initiatives since 2025 have emphasized diversification in AI research and application domains. In March 2025, "embodied AI" was formally designated a national development priority, and in 2026, China released the *Humanoid Robot and Embodied Intelligence Standard System*,[2] establishing a nationwide framework for humanoid robot design and embodied intelligence development. Similarly, China's inclusion of neuromorphic computing in the *Made in China 2025* strategy has raised questions about its potential to gain an advantage in non-transformer-based architectures.[3] These initiatives, combined with evidence of Chinese investment in robotics, have prompted concerns that the United States might be lagging behind China in embodied AI and the integration of AI into industrial and robotic systems.[4]

The veracity of these claims is consequential. If true, this imbalance could leave the United States poorly positioned to be the first to achieve artificial general intelligence (AGI) should the scaling paradigm encounter significant limits to performance improvement. However, given the scale and dynamism of the AI industry, focusing solely on a handful of major firms risks providing a distorted view of the broader commercial AI developer landscape in each country. Moreover, official plans might not accurately reflect the actual distribution of technical capabilities or commercial activity for the United States or China. Assessing these claims requires moving beyond policy documents,

---

[1] Marcellino, "Hedging Our Bets"; Marcellino et al., *Charting Multiple Courses to Artificial General Intelligence*; Hannas et al., *Chinese Critiques of Large Language Models*.

[2] Ministry of Industry and Information Technology, People's Republic of China, *Humanoid Robot and Embodied Intelligence Standard System*.

[3] State Council of the People's Republic of China, "Made in China 2025" ["中国制造2025"].

[4] Chang, Arcesati, and Junusova, *Embodied AI*; Hannas et al., *China's Embodied AI*.

anecdotal insider testimonies, and trends in high-profile firms to systematically characterize each country's AI developer ecosystem. Our study undertook that task: We constructed a broad empirical foundation against which claims about concentration, diversification, comparative advantage, and other topics can be evaluated.

The remainder of this report is organized as follows. Chapter 2 describes the process by which we built the dataset—the identification of candidate organizations, the application of inclusion criteria, and the population of firm-level variables. Chapter 3 presents basic descriptive statistics of the U.S. and Chinese AI developer industries. Chapter 4 examines the commercial orientation of firms in each country, covering commercial positioning, model provenance, application domain, and open-weight status. Chapter 5 analyzes the technical approaches taken by firms in each country across five dimensions: model architecture, learning paradigm, data modality, physical form factor, and core functionality. We conclude in Chapter 6 with a synthesis of findings and a discussion of limitations and directions for future work.

Four appendixes provide additional technical details on methodology. Appendix A describes the full set of categories to which firms were assigned. Appendix B discusses the AI-driven data collection and classification process used to categorize firms, Appendix C presents the key prompts employed in that process, and Appendix D provides the full LLM prompt used to implement the Phase 2 AI developer inclusion criterion.

An annex to this report that lists the firm names used during the data collection process is available at www.rand.org/t/RRA5043-1.

Chapter 2

Building the Dataset

As we will describe in more detail later in this chapter, we defined our target population in this study as the set of firms in each country that are directly involved in the development of AI technology (i.e., *AI developers*). Specifically, we aimed to include organizations that develop, train, or release AI models or AI-enabled systems in which the core AI capabilities are internally developed. We excluded (1) firms that primarily provide downstream deployment or integration services for models developed by others and (2) organizations whose principal activity is supplying AI infrastructure or components (e.g., chips, cloud computing, or data) without developing AI models themselves. Our focus on AI developers is driven both by tractability and by the assumption that they represent a primary locus of private-sector AI innovation, although we fully acknowledge that governments, universities, infrastructure providers, and downstream integrators also shape national AI ecosystems.

Throughout this report, we use the phrase *AI developer ecosystem* as shorthand to refer to the ecosystem of commercial AI developer firms. This term refers exclusively to the population of profit-seeking companies meeting the AI developer criteria described here and in Appendix D; it does not encompass universities, government agencies, research laboratories, or other non-firm actors, even for cases in which such actors play important roles in national AI development.

Our data collection process proceeded in three phases. First, we constructed a broad set of candidate organizations that exhibit evidence of AI development activity. Second, we filtered this list according to our inclusion criteria, retaining only those organizations that are firms, are headquartered in the United States or China, and qualify as AI developers. Third, for each firm in the resulting target population, we populated a set of variables using LLM search agents, complemented by several rounds of human validation. We describe each of these phases in greater detail in the following sections.

Phase 1. Building a Set of Candidate Organizations

Consistent with our definition of the target population—firms that directly develop, train, or release AI models or AI-enabled systems—we identified candidate AI-developing firms using complementary indicators of AI development activity. In the absence of a single authoritative registry of such firms, we drew from several data sources that capture distinct and observable manifestations of direct AI development.

Specifically, we constructed our list of candidate organizations using four sources: organizations employing individuals in AI development roles, organizations publishing AI research, organizations holding AI patents, and organizations ranking on AI leaderboards.

Firms Employing Individuals in AI Development Roles

Employing personnel in AI development roles provides strong evidence that a firm is engaged in the development of AI technology. Although some employees in AI-related roles might support out-of-scope activities—such as applying preexisting models to internal business processes—we filtered out such false positives during Phase 2 of our process.

The data source we leveraged for this category was Enrich Layer, which provides longitudinal career trajectory information on more than 790 million professionals. We defined a set of AI development–relevant occupational roles (e.g., machine learning [ML] engineer), queried Enrich Layer for individuals holding these roles, and extracted the firms employing them in the United States and China. We ran two queries: one in January 2026 and one in February 2026. Using this approach, we captured firms actively investing in AI-specific human capital, a necessary condition for sustained in-house AI development.

Firms Publishing AI Research

Although not all firms that are engaged in AI development publish academic research, the publication of original research in AI is a clear indicator of direct involvement in advancing AI capabilities.[5] Firms that publish AI research are, by definition, generating new technical knowledge rather than merely adopting existing models or tools.

Our data source for this category was SciVal, a repository of scientific publication metadata. We queried SciVal for corporate entities that published articles classified under the "Artificial Intelligence" research category during the 2024–2025 period. Any firm with at least one AI publication was included in our candidate dataset.

Firms Holding AI Patents

In addition to publishing research, firms might encode their progress in building AI technology through patents. Patent assignees represent organizations that have invested in the development of novel, legally protectable, AI-related inventions. As a result, patenting activity provides a complementary and more application-oriented signal of AI development than either employment or publication data alone. Including patent assignees in our population helped capture firms that conduct proprietary or product-focused AI research and development (R&D) that might not be visible through academic publishing, thereby reducing bias toward research-intensive firms and improving coverage of the broader commercial AI ecosystem.

The data source we used for this category was the Derwent Innovation Index patent data. Our focus is the contemporary AI commercial ecosystem, so we limited results to the 2024–2025 period. To build a set of AI patent assignees, we leveraged a methodology developed in 2025 by the Organisation for Economic Co-operation and Development (OECD) to identify AI patents.[6] Specifically, we used a combination of patent classification codes and AI-related keywords to produce

[5] Schmid, *An Open-Source Method for Assessing National Scientific and Technological Standing*.

[6] OECD, *Identifying Emerging AI Technologies Using Patent Data*.

a set of AI patents. From this set, we generated a list of corporate assignees from the United States and China.[7]

Firms Ranking on AI Leaderboards

Submitting models to AI benchmarks or leaderboards provides direct evidence that a firm has developed, trained, and released an AI model for external evaluation. Participation in such benchmarks requires internal model development capabilities and is therefore closely aligned with our target-population definition.

We drew on two leaderboards: The first was a performance leaderboard aggregating benchmark results across providers,[8] and the second was a ranking of Chinese open-model developers.[9]

Phase 2. Applying Inclusion Criteria

After removing duplicate organizations,[10] Phase 1 yielded a set of 26,629 candidate entities. In Phase 2, we refined this list to ensure that the dataset focuses on the target population of corporate AI developer firms headquartered in either the United States or China. To achieve this, we applied three inclusion criteria:

- The organization is a company.
- The organization is headquartered in the United States or China.
- The organization qualifies as an AI developer.

We applied each criterion using a web search–enabled LLM-assisted labeling process, which was subsequently refined and validated through manual review.

The first criterion (i.e., the organization is a company) ensured adherence to the study's focus on the commercial AI sector. A positive classification required that the focal organization be a profit-seeking company. Universities, government research laboratories, and individual researchers were therefore excluded from the sample.

The second criterion excludes organizations not headquartered in the United States or China. Country of origin was determined by the geographic location of the firm's global headquarters.

---

[7] Because the patent data we used did not allow searching on Cooperative Patent Classification codes, as in the OECD methodology, we relied instead on the Derwent classification system. The final search comprised all Derwent Manual codes containing the phrase *artificial intelligence* or any of the 274 AI-related keywords identified in the OECD methodology. Finally, we limited results to firms with more than ten AI patents during the period; this list was what we included in our candidate dataset.

[8] Artificial Analysis, "LLM API Providers Leaderboard—Comparison of over 500 AI Model Endpoints."

[9] Lambert and Brand, "China's Top 19 Open Model Labs."

[10] To remove duplicate firms, we constructed an affiliation graph from approximately 30,000 company profiles using declared parent–subsidiary relationships to cluster entities and selected a canonical representative for each group. Remaining duplicates were resolved through a series of deterministic matching rules (shared internal identifiers, website domains, and constrained fuzzy name matching), and external records from publications, patents, and model registries were matched against the canonical set to ensure that subsidiaries resolved to their parent firms.

The final criterion ensured that the resulting sample contained only organizations meeting our definition of AI developers. In this report, an *AI developer* is an organization actively engaged in developing new AI technology—creating, training, or releasing its own models or AI-enabled systems—rather than primarily applying or enabling models developed by others. A firm qualifies as an AI developer if it meets at least one of two criteria:

- It develops or trains AI models or systems that advance the state of the art in its domain, including foundation models, embodied AI or robotics control systems, neuromorphic or cognition-inspired architectures, novel training paradigms, or other technically novel AI contributions documented through model releases, peer-reviewed publications, patents, or detailed technical documentation.
- It demonstrates sustained institutional commitment to AI R&D through dedicated AI research divisions, recurring model releases or publications, or ongoing investment in new AI development directions. Examples of firms that qualify are a company that trains its own LLM from scratch and releases it via application programming interface (API) (e.g., DeepSeek), a robotics firm that develops and publishes its own locomotion control models (e.g., Unitree Robotics), and a medical AI company that trains proprietary diagnostic models on its own clinical dataset (e.g., Tempus AI).

Our definition excludes firms that, however sophisticated their AI use, do not themselves develop new models or paradigms. Banks deploying AI for fraud detection, cloud providers offering APIs built on others' models, systems integrators applying existing AI tools to customer workflows, and companies whose AI activity is limited to fine-tuning, prompt engineering, or retrieval-augmented generation using third-party models are all excluded. In all ambiguous cases, the default classification was "no": Absent direct, concrete evidence of in-house AI development (such as a named model release, a technical publication describing in-house training, or a documented AI research division), the firm was excluded. Speculative or inferential language in the classifier's reasoning (e.g., "likely," "appears to," "may") automatically triggered a "no" classification regardless of other evidence. The full classification prompt used to implement this criterion is provided in Appendix D.

Phase 3. Populating the Dataset

Once the target population was finalized, we populated a standardized set of variables for each firm using a system of LLM-based search agents. Each agent was scaffolded with a web search tool—primarily Gemini 2.5 Pro with online search[11]—and tasked with conducting background research on the firm and extracting the relevant information. Variables collected were basic descriptive characteristics (such as year of founding, publicly traded status, and military AI relationship), commercial orientation details (such as commercial positioning mode and model provenance), and technical variables (such as model architecture, learning paradigm, data modality, and physical form factor). The full variable set and taxonomy definitions are provided in Appendix A.

[11] Although results in this report were obtained leveraging Gemini 2.5 Pro, future work could investigate the cost, reliability, and latency of other LLM annotators and the rate at which they converge on annotation decisions.

We refined our prompt design by sampling outputs from approximately 100 randomly selected firm-variable pairs and inspecting the results manually for fidelity. During this inspection, we also examined the reasoning traces underlying each classification, which helped us pinpoint the sources of LLM disagreement and adjust prompts accordingly.

For certain firm variables, we supplemented our web search with additional data sources. When classifying the more-technical attributes of a firm—such as its dominant AI architectures or the learning paradigm underlying its products and services—we granted the search agent access to Perplexity and OpenAlex (a scientific publication repository), enabling it to more readily analyze the firm's open-source technical publications.

After all variables had been extracted for the complete firm set, we performed a final human evaluation on a holdout random sample of 25 firms, separate from any firm-variable pairs used during prompt refinement. For each audited firm-variable classification, a human rater categorized their agreement with the agent's classification as Strongly Agree, Agree, Disagree, or Strongly Disagree. We define *human-agent agreement* as the share of audited classifications receiving either Agree or Strongly Agree.

This evaluation serves two related but distinct purposes. First, it provides a variable-level estimate of human-agent agreement. For example, an observed agreement rate of 90 percent for $n = 25$ implies an approximate 95-percent margin of error of 11.8 percentage points under the standard binomial approximation for the human-agent agreement measurement.

Second, the audit helps bound the possible measurement error in the binary fractions estimated from agent data annotations that we report in later chapters. For a binary classification variable, the difference between the agent-reported fraction and the corresponding human-reference fraction is bounded above by the human-agent disagreement rate, assuming that the audit is representative and the human rating is treated as the reference standard. Thus, an observed human-agent agreement rate of 90 percent implies a disagreement rate of 10 percent and a conservative measurement-error bound of approximately plus or minus 10 percentage points for the corresponding AI-coded fraction. This bound is conservative because it assumes that all classification errors move the estimated fraction in the same direction; if false positives and false negatives partially offset one another, the resulting bias in the reported fraction would be smaller.

In the audit, nearly all variables retained for analysis showed high (more than 90 percent) observed agreement between human raters and the classification agent over the evaluation dataset. Variables with low human-agent agreement were excluded from the analysis reported in this study. For example, we initially considered tracking the country of education for founders and chief executive officers of firms in our dataset. However, both the AI search agents and internal team members had difficulty obtaining consistent classifications for this variable because of conflicting early-career narratives, ambiguous corporate hierarchies, and rapid movement of leaders across firms. As a result, we excluded this variable, along with others that faced similar reliability challenges, from the analysis.

Scope and Data Limitations

A Firm-Level View of a Broader AI Ecosystem

By excluding universities, government agencies, government research laboratories, individual researchers, and other non-firm entities from our study, we produced a sample that, although analytically useful, did not capture the full AI ecosystem in either country. The excluded sectors play important roles in AI development, and those roles might differ substantially between the United States and China. The findings in this report should therefore not be interpreted as a complete census of all advanced AI activity in either country.

We nevertheless believe that this firm-centric subset is a valuable object of analysis for three reasons. First, commercial firms are central actors in the development, deployment, and diffusion of contemporary AI systems. They translate research into products, platforms, infrastructure, and embodied systems, and they often determine which technical approaches reach users, customers, and downstream developers. Second, focusing on firms standardizes the analysis and enables systematic cross-national comparison using such observable indicators as employment, patents, publications, model releases, commercial offerings, and public technical documentation. Third, many policy-relevant claims about U.S. and Chinese AI competition concern the industrial structure of each country's AI sector, focusing on such questions as whether firms are concentrated in foundation models, whether they are building embodied AI systems, whether they are publishing open-weight models, and how innovation activity is distributed across technical approaches. A firm-level dataset is well suited to evaluating these claims.

At the same time, using commercial AI firms to draw broader conclusions about national AI ecosystems requires caution. Such suppositions assume that the commercial AI sector captures a meaningful portion of each country's advanced AI activity; that the effects of national AI strategies, including Chinese policies emphasizing embodied AI and technological diversification, are at least partly visible in commercial firms; and that AI developer firms are a useful lens through which to observe broader ecosystem orientation. These assumptions likely hold more strongly for questions about commercial deployment, product orientation, and industrial diffusion than for questions about early-stage basic research, long-horizon experimental architectures, or government-directed research programs. For example, some of the most speculative work on neuromorphic computing, brain-inspired architectures, or other non-transformer approaches might occur primarily in universities or government-funded research centers rather than in commercial firms. To the extent that this is true, our firm-level approach could understate forms of technical diversification occurring outside the commercial sector.

Coverage Limits in Public Data Sources

The process used to construct both the candidate firm list and the firm-variable dataset contains uneven levels of coverage across the United States and China. For example, when identifying firms, we use multiple data sources to reduce dependence on any single indicator; however, each source has limitations. Employment-based discovery relies on Enrich Layer's longitudinal career trajectory data, which draws on professional profile information that might have uneven coverage across countries and

sectors. Professional profile penetration in China is substantially lower than in the United States, meaning this step likely underidentifies Chinese individuals in AI-development roles and the firms employing them. This effect might be especially pronounced for individuals who work at firms with ties to the People's Liberation Army, Ministry of State Security, Ministry of Public Security, or other Chinese security services and who are unlikely to maintain public professional profiles. This undercount is partially offset by our patent-based extraction step: The Chinese government provides strong incentives to patent, and the number of Chinese firms surfaced through patent data exceeded the number of U.S. firms in our extraction, which likely counteracts some of the employment-data undercount of Chinese firms.[12] Publication-based discovery through SciVal is similarly valuable but might miss firms whose research appears in venues, languages, or databases less visible to the indexing systems we use. This effect is likely partially offset by strong prestige and career incentives for Chinese AI researchers to publish in top English-language venues.[13] A parallel language bias in OpenAlex is discussed in Appendix B; the issue applies to SciVal as well.

Similarly, when populating firm-level variables (e.g., model architecture, learning paradigm) we rely on academic publications, white papers, and online search agents. Coverage of Chinese firms in these data sources is sparser than coverage of U.S. firms. This issue affects how cross-national comparisons should be read. When one country has a substantially higher unknown rate, differences in named categories might reflect both real ecosystem differences and differences in public disclosure. For example, if a larger share of Chinese firms lacks public documentation about model architecture or learning paradigm, then lower observed shares for named technical categories might partly reflect data availability rather than true absence. Later chapters therefore discuss unknown rates directly when interpreting technical comparisons.

The estimates in this report should be interpreted with attention to both missingness and the size of observed differences. Large cross-national gaps, such as differences in software-only versus embodied form factors, are more likely to be robust to plausible coverage and annotation error. Smaller differences, especially those of only a few percentage points, should be treated as descriptive and tentative. Finally, all findings in this report reflect publicly available data collected in January and February 2026. Given the pace of change in the AI developer industry—firms enter, exit, pivot, and reorient rapidly—some classifications might no longer reflect a firm's activities, and the cross-national comparisons presented here should be understood as a snapshot of each ecosystem at that point in time rather than a durable characterization.

[12] Schmid and Wang, "Beyond National Innovation Systems"; Long and Wang, "China's Patent Promotion Policies and Its Quality Implications."
[13] Quan, Chen, and Shu, "Publish or Impoverish"; Luo and Hyland, "International Publishing as a Networked Activity."

Chapter 3

Descriptive Statistics of U.S. and Chinese AI Developer Industries

This chapter presents basic descriptive information about the AI developer firms in our dataset. We begin by presenting summary statistics for five key firm-level characteristics—headquarters country, year founded, publicly traded status, military AI relationship, and diversified firm status—broken out by country. We then examine the geographic distribution of firms in each country and the temporal pattern of firm founding. Table 3.1 presents summary statistics for the five firm-level variables of interest broken out by headquarters country. Categorical variables are reported as counts and column percentages; year of founding is reported as the median with interquartile range (IQR).

Table 3.1. Descriptive Summary of U.S.- and China-Based AI Developers

Variable	United States (*n* = 743)	China (*n* = 438)	Total (*n* = 1,181)
Headquarters country	743 (62.9%)	438 (37.1%)	1,181 (100.0%)
Year founded, median [IQR]	2016 [2010–2019], *n* = 734	2014 [2003–2017], *n* = 431	2015 [2007–2019], *n* = 1,165
Publicly traded			
Yes	101 (13.6%)	148 (33.8%)	249 (21.1%)
No	641 (86.3%)	289 (66.0%)	930 (78.7%)
Unknown	1 (0.1%)	1 (0.2%)	2 (0.2%)
Military AI relationship			
Yes	223 (30.0%)	71 (16.2%)	294 (24.9%)
No	518 (69.7%)	366 (83.6%)	884 (74.9%)
Unknown	2 (0.3%)	1 (0.2%)	3 (0.3%)
Diversified firm			
Yes	214 (28.8%)	248 (56.6%)	462 (39.1%)
No	529 (71.2%)	188 (42.9%)	717 (60.7%)
Unknown	0 (0.0%)	2 (0.5%)	2 (0.2%)

NOTE: Categorical variables show count (% of country column). Year founded shows median [25th–75th percentile] and *n* with non-missing values. Numbers might not sum to 100 because of rounding.

We assign each firm to a national ecosystem using its headquartered country, defined as the country in which the firm reports its primary corporate headquarters. Of the 1,181 firms in the consolidated dataset, we identify 743 as headquartered in the United States and 438 in China.

For each firm, we record the year founded, defined as the calendar year of incorporation as reported in public filings or the firm's own communications. The median founding year is 2016 for U.S. firms (IQR 2010–2019) and 2014 for Chinese firms (IQR 2003–2017).

We further classify firms by their public listing status. The publicly traded indicator is labeled "yes" if the firm is listed on any stock exchange, distinguishing publicly listed firms from privately held firms. In our sample, 13.6 percent of U.S. firms and 33.8 percent of Chinese firms are publicly traded; the remainder are privately held or of unknown status.

The military AI relationship indicator is labeled "yes" if publicly available sources directly document a commercial or R&D relationship between the firm and its domestic military—the U.S. Department of War or the People's Liberation Army—specifically related to AI. Procurement contracts, contract awards, or documented defense R&D partnerships constitute qualifying evidence.[14] We find evidence of such a relationship for 30.0 percent of U.S. firms and 16.2 percent of Chinese firms, though the Chinese figure is likely an undercount given the limited public disclosure of Chinese defense-industrial AI activity already noted.

Finally, the diversified firm indicator measures whether a firm belongs to a larger parent organization that produces a broad variety of non-AI products or services rather than focusing exclusively on AI development. Examples include Meta and Google in the United States and Alibaba and Tencent in China. By this measure, 28.8 percent of U.S. firms and 56.6 percent of Chinese firms in the dataset are diversified.

Geographic Distribution

U.S. firms are grouped heavily on the coasts in major technology clusters. The five states with the largest number of firms are California (350), Massachusetts (75), New York (70), Texas (36), and Pennsylvania (25). Chinese firms are concentrated in the eastern and southern provinces; the five province-level regions with the largest counts are Beijing (159), Guangdong (90), Shanghai (58), Zhejiang (29), and Jiangsu (26). Figures 3.1 and 3.2 display these distributions on choropleth maps shaded by the number of firms per state or province.

[14] Executive Order 14347, signed September 5, 2025, authorized the use of Department of War as a secondary name for the Department of Defense. This publication refers to the secretary and department by their current statutory names under Public Law 81-216, National Security Act Amendments of 1949.

Figure 3.1. U.S.-Based AI Developers, by State

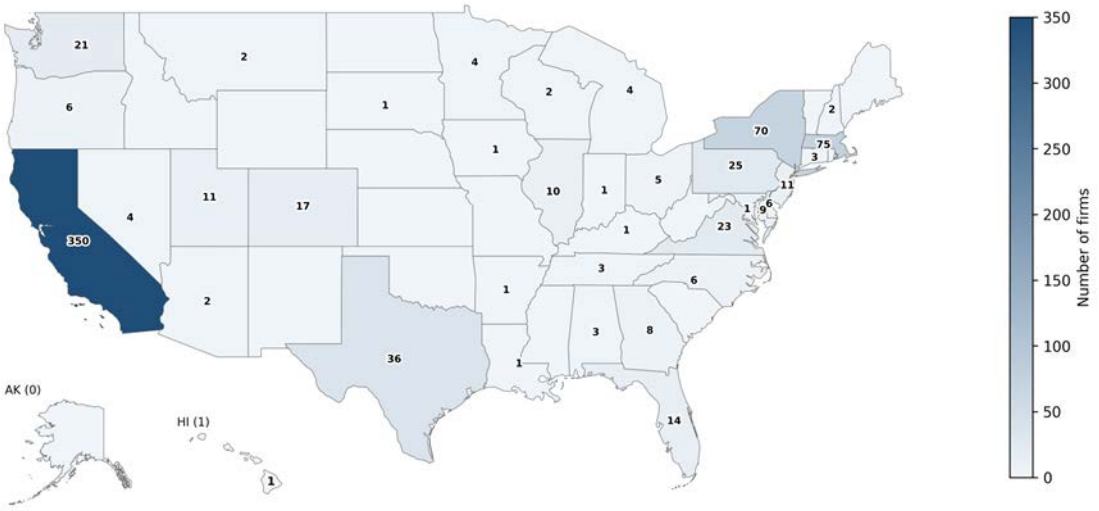
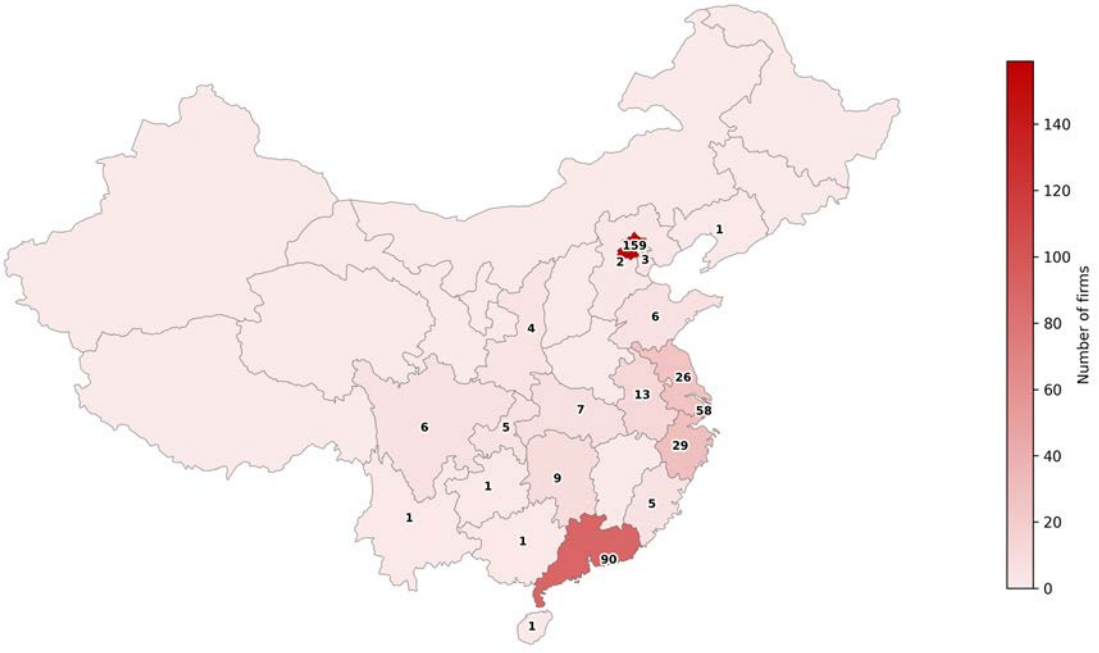


Figure 3.2. China-Based AI Developers, by Province-Level Region

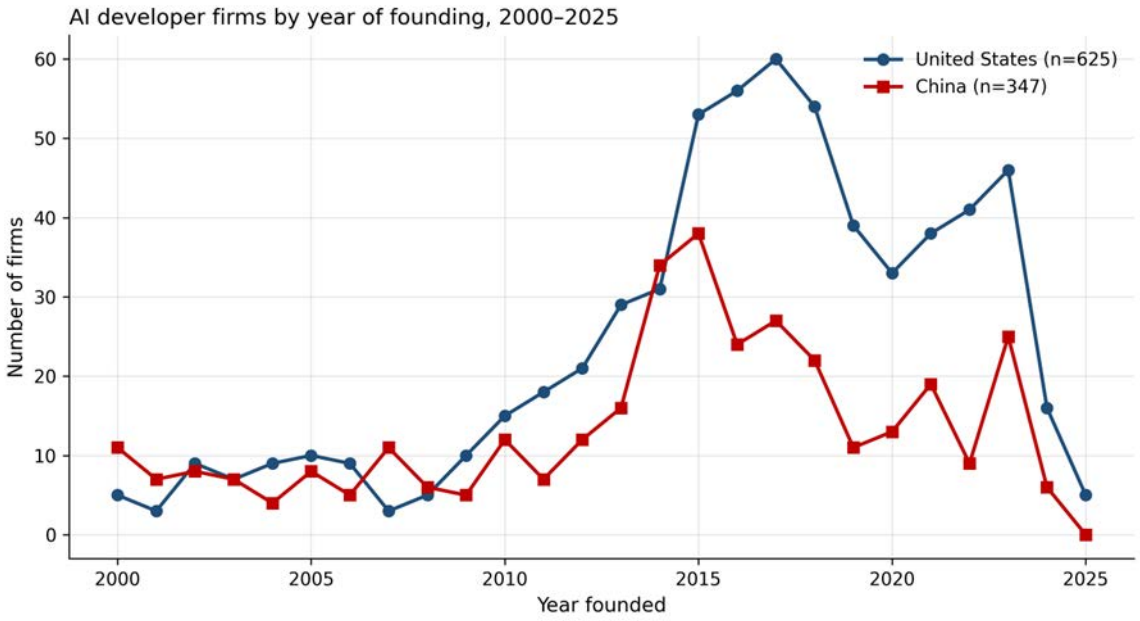


NOTE: Provinces could not be determined for a small number of firms, which are therefore not shown on this province-level map.

Year of Founding

The median year of founding is 2016 for U.S. firms and 2014 for Chinese firms, with the largest cohorts in both countries established between 2014 and 2018. Figure 3.3 displays the count of firms founded in each year from 2000 to 2025; 188 firms founded before 2000 are omitted from the figure but retained in all other analyses. This clustering is consistent with the wave of commercial AI investment and firm formation that followed breakthroughs in deep learning in the early to mid-2010s.

Figure 3.3. AI Developer Firms, by Year of Founding (2000–2025)



Chapter 4

Commercial Orientation of U.S. and Chinese AI Developer Industries

In this chapter, we examine the commercial orientation of U.S. and Chinese AI developer firms. We present our results for four firm-level dimensions: commercial positioning, model provenance, application domain, and open-weight status. We then use an additional data source, OpenRouter, to test whether the firm-level picture is borne out in actual developer adoption.

Four Dimensions of Commercial Orientation

Figure 4.1 summarizes the four dimensions used in this chapter and the categories that fall under each. The categories in a dimension are nonexclusive. For example, a firm can be tagged as a foundation developer and as a model adapter if it legitimately engages in both activities. Firm-level percentages in a dimension are therefore allowed to sum to more than 100. Across dimensions, the categorizations are independent of one another.

Figure 4.1. Commercial Orientation Dimensions and Categories

Commercial Positioning

Our commercial positioning dimension is adapted from the layered service-model taxonomy established for cloud computing and its subsequent application to AI-specific business models.[15] It distinguishes among four primary commercial positioning modes:

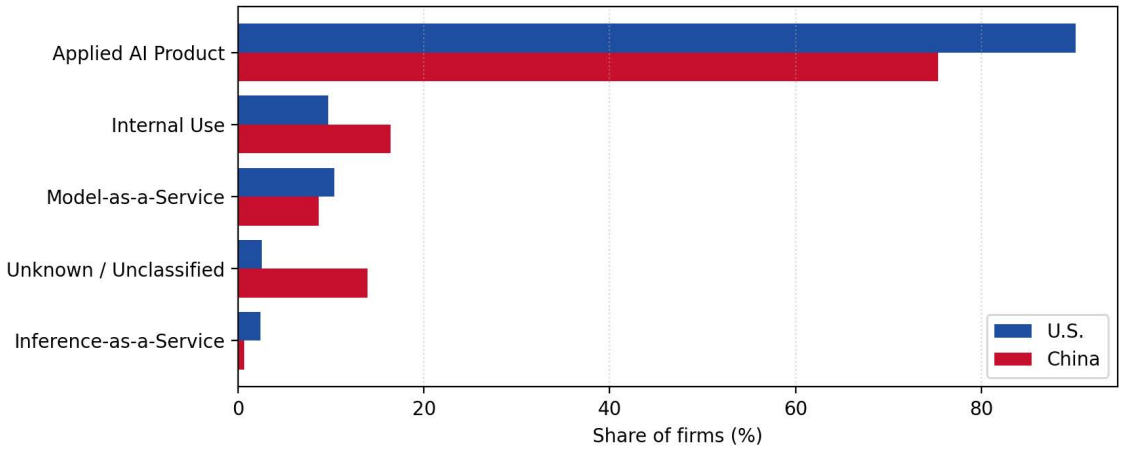
- **model-as-a-service**: selling access to trained models directly, typically through an API
- **inference-as-a-service**: selling optimized inference for third-party models
- **applied AI product**: selling an end-user or vertical product whose value to the customer is delivered through an AI-enabled application

[15] National Institute of Standards and Technology, *The NIST Definition of Cloud Computing*; Corea, *Artificial Intelligence and Exponential Technologies*.

- **internal use**: developing AI capability primarily for use inside the parent organization without commercializing the model itself.

The dominant commercial positioning mode in both countries is the development of applied AI products. Roughly 90 percent of U.S. firms and 75 percent of Chinese firms fall into this category. The model-as-a-service category—populated by such frontier API providers as OpenAI, Anthropic, Alibaba, and Zhipu—accounts for roughly 10 percent of firms in each country. The popular framing of the commercial AI race as a contest among a set of frontier API providers therefore describes a real but quantitatively small (in terms of number of firms) slice of the AI developer ecosystem in both countries. The majority of firms are building applications rather than selling model access. Figure 4.2 presents the full distribution across all four commercial positioning categories.

Figure 4.2. Distribution of Firms, by Commercial Positioning (United States and China)



The largest asymmetry is in the internal use category. About 16 percent of Chinese firms occupy this category, compared with 10 percent of the U.S. sample. In China, many of the firms in this category are large state-owned enterprises, utilities, regulated banks, and telecommunications firms. The pattern is consistent with the observation that a larger share of Chinese firms has a principal application domain of manufacturing, transportation, or energy.

Model Provenance

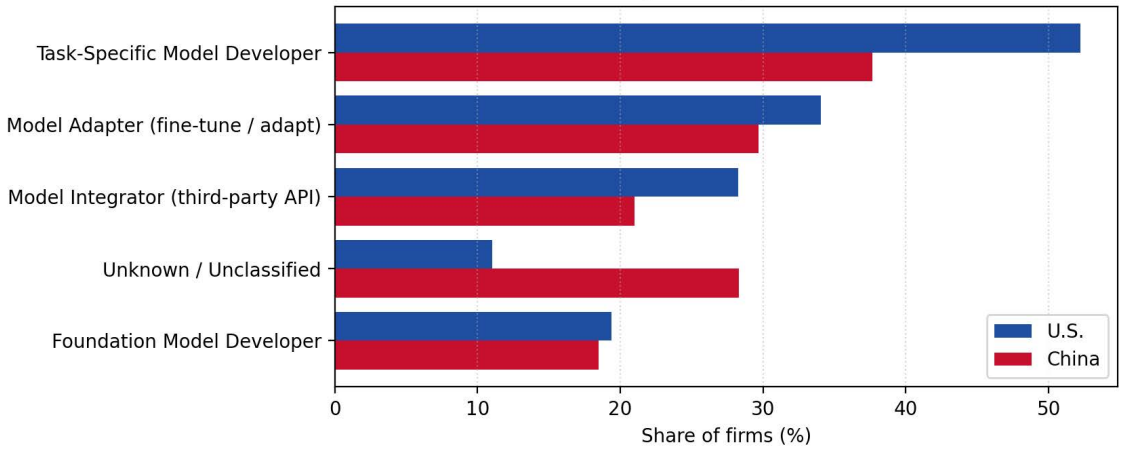
The commercial positioning dimension asks how a firm sells AI, whereas the model provenance dimension asks what a firm does with models. We define four categories of model provenance:

- **foundation model developer**: firms that pretrain general-purpose models at scale from scratch
- **task-specific model developer:** firms that train domain-specific or task-specific models from scratch, typically on proprietary data
- **model adapter**: firms that fine-tune or otherwise adapt third-party base models with custom weights

- **model integrator:** firms that build products on top of third-party model APIs without modifying weights.

Although foundation model development absorbs the largest share of media attention, only 19 percent of U.S. firms and around 19 percent of Chinese firms in our dataset are tagged as foundation developers—although the high rate of Chinese firm data missingness (28 percent) might be masking subtler differences between the two countries across this dimension. According to the available data, it is more common in both countries for a firm to train narrow, task-specific models; to adapt existing foundation models with proprietary data; or to integrate third-party model APIs into vertical products. Figure 4.3 shows the distribution across all four model-provenance categories.

Figure 4.3. Distribution of Firms, by Model Provenance



Task-specific model development is more prevalent in the U.S. sample (52 percent) than in the Chinese sample (38 percent). Some illustrative examples of the common profiles we observed in our sample for the United States are Abridge (see case study in Appendix B), which trains custom speech-recognition models from scratch for medical conversations; Recursion Pharmaceuticals, which trains task-specific cell-imaging and drug-response models for compound screening; and Tempus AI, which trains oncology-specific models on its proprietary clinical dataset. In China, on the other hand, Unitree Robotics trains task-specific locomotion and manipulation controllers for its humanoid and quadruped robots, iFlytek trains custom Mandarin speech models for industrial voice interfaces, and BYD trains autonomous-driving perception and planning models for its vehicle line.

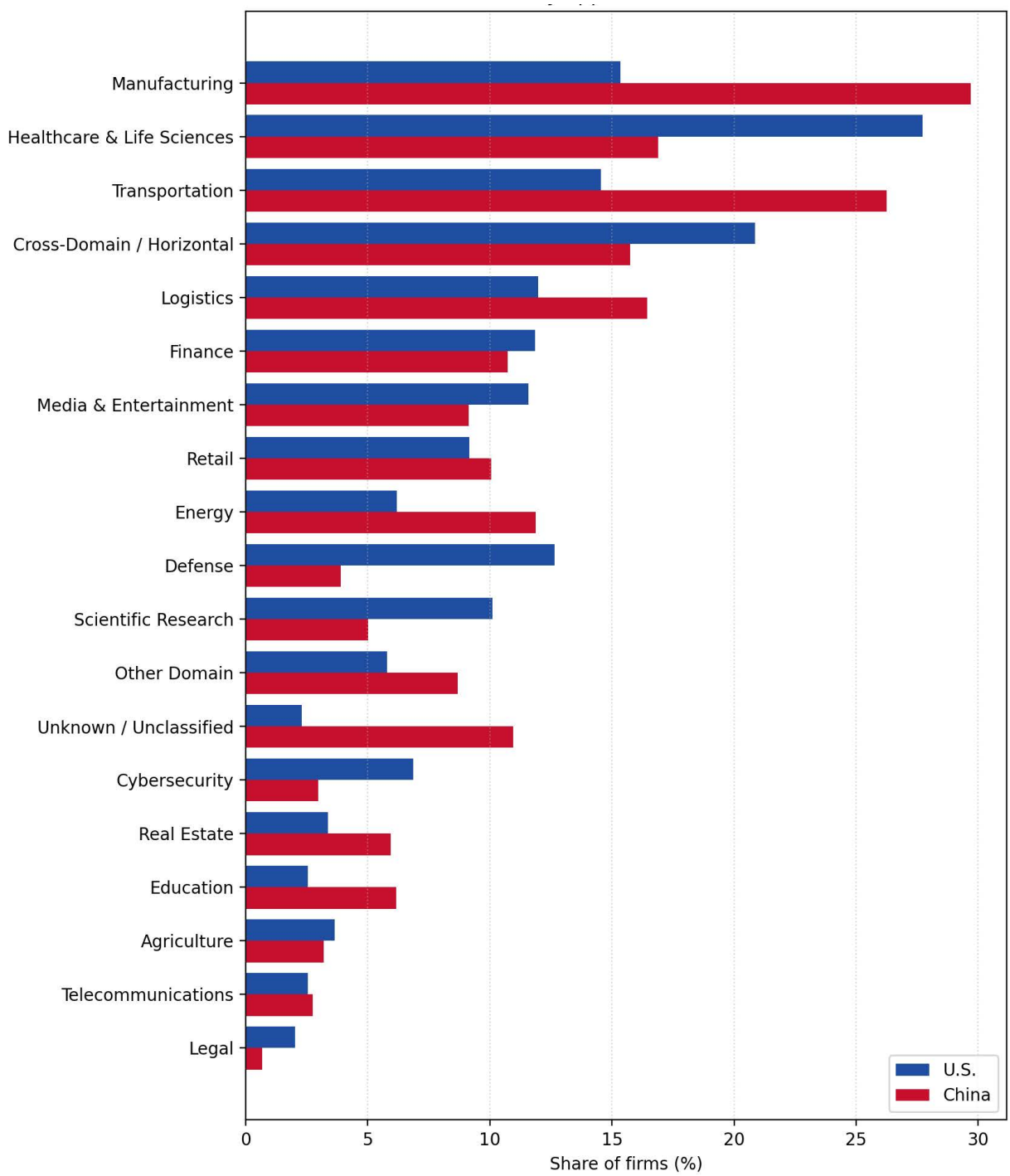
Again, it is worth noting that interpreting this gap requires caution. As discussed in Appendix B, the "unknown" share of Chinese firms is consistently higher than the U.S. share across positively assigned dimensions; Chinese counts should therefore be treated as lower bounds. Given the elevated Chinese unknown rate on this dimension (28 percent), we cannot be confident that the observed gap in task-specific model development reflects a real underlying difference rather than an artifact of higher Chinese missingness.

Application Domain

The application domain dimension records the verticals that each firm targets. A firm can serve more than one domain, and our classifier assigned a primary domain alongside any clearly evidenced secondary domains.

Application domain is where the clearest structural difference between the two ecosystems appears. Chinese firms are concentrated in physical-economy verticals: manufacturing (30 percent of Chinese firms versus 15 percent of U.S. firms), transportation (26 percent versus 15 percent), and energy (12 percent versus 6 percent). In contrast, U.S. firms are concentrated in knowledge-intensive verticals: health care and life sciences (28 percent versus 17 percent), scientific research (10 percent versus 5 percent), and cybersecurity (7 percent versus 3 percent). Figure 4.4 plots the share of U.S. and Chinese firms tagged with each domain.

Figure 4.4. Distribution of Firms, by Application Domain

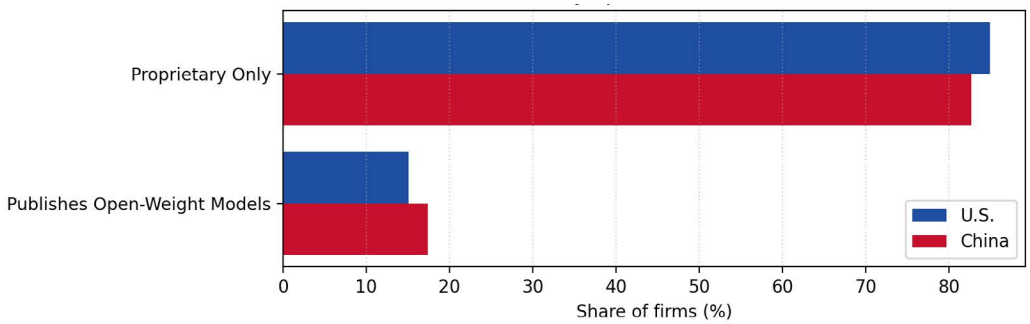


The most pronounced gap is in defense; 13 percent of U.S. firms are tagged versus 4 percent of Chinese firms. We are cautious about reading this gap as a real industrial difference; Chinese defense-industrial AI work is often conducted in state-affiliated enterprises whose public disclosure is limited.

Open-Weight Status

The fourth commercial dimension we consider in this chapter is whether a firm releases any open-weight models. Open-weight release is a strategic choice that crosses the commercial positioning dimension. For example, a model-as-a-service firm, such as DeepSeek, can release its weights publicly while still selling a managed API. Similarly, an applied-AI-product firm, such as Anthropic, can build its commercial offering entirely around proprietary, closed-weight models. Figure 4.5 compares the share of firms in each country that publish at least one open-weight model.

Figure 4.5. Share of Firms Publishing Open-Weight Models



Assessing the U.S. and Chinese ecosystems by counting firms suggests a similar concentration: Approximately 15 percent of U.S. firms publish at least one open-weight model, versus 17 percent of their Chinese counterparts. The remainder of firms in each country are proprietary only. Some keep their model usage internal (applied product and internal use firms); others release access through closed APIs.

However, the argument we develop in the next section is that mere publication of open-weight models tells only part of the story. Model usage also matters. Whichever set of models developers actually build on is the set that shapes the products, conventions, and de facto standards around which the global AI ecosystem will consolidate.

OpenRouter Data: A Window into Developer Adoption

OpenRouter is a developer-facing gateway that routes API requests across roughly 300 frontier and open-weight models from dozens of model providers and inference vendors. It publishes daily token-volume statistics by model and by developer. We use OpenRouter as a complement to the firm-level taxonomy, not as a substitute for it. The data capture developer-mediated traffic—requests that third-party developers route through OpenRouter as part of their own software—and therefore exclude consumer chatbot usage and direct use of provider APIs (e.g., requests sent straight to OpenAI, Anthropic, or Google rather than through OpenRouter). The relevant property for our purposes is that OpenRouter is the principal way to access many open-weight Chinese models from inside the United States, which makes the data well suited to measure how open releases are used in this specific developer population.

Three caveats apply to the OpenRouter data and should be borne in mind throughout this section. First, OpenRouter represents an unknown but likely small fraction of total global developer API traffic; the majority of developer calls to models such as GPT-4o, Claude, and Gemini are made directly to provider APIs and are not captured here.

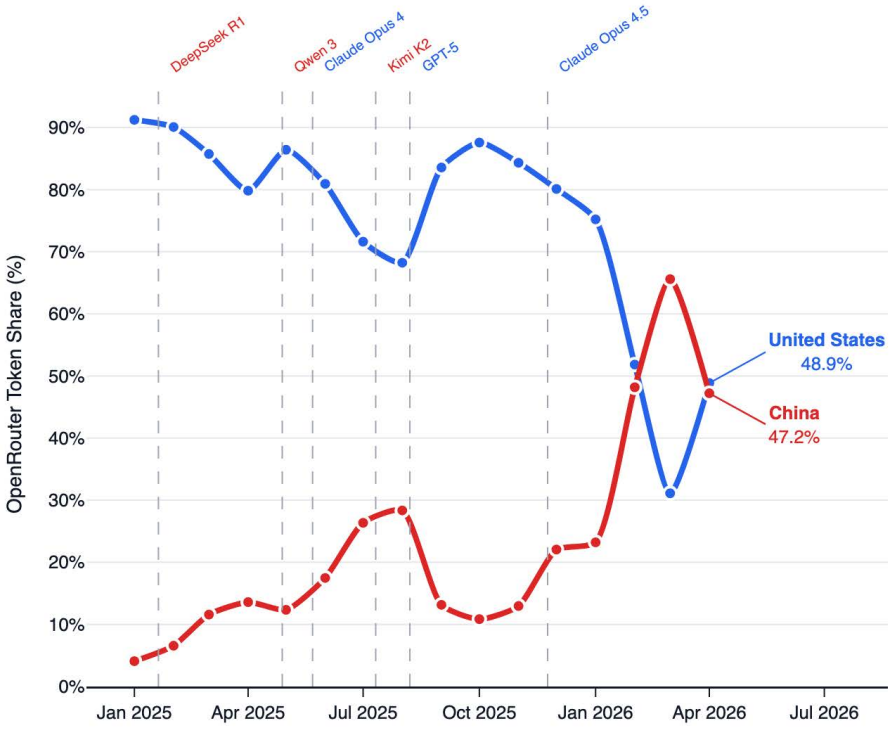
Second, because OpenRouter structurally overrepresents open-weight models relative to the broader market—and because Chinese frontier models are predominantly open-weight, whereas leading U.S. frontier models are predominantly proprietary—the platform's user base is not a neutral sample of global developer preferences and might be most relevant for describing relative trends rather than absolute ones. This interpretation is consistent with a 2026 RAND analysis of global LLM adoption patterns using broader web traffic data, which found Chinese models rising from roughly 3 percent to 13 percent of global market share over a similar period, a meaningful gain but well short of the parity we observe on OpenRouter (discussed next).[16]

Third, usage patterns on OpenRouter might exclude enterprise or consumer AI engagement, which often happen through internal and/or customized APIs that are not captured on the OpenRouter platform. With these caveats in mind, we interpret the OpenRouter data as a directional signal about the competitive trajectory of Chinese open-weight models among cost-sensitive developers rather than as a definitive measure of global market share.

Examining the data reveals that usage of Chinese models has caught up to that of the U.S. models. In late 2024, U.S. models accounted for roughly 85–90 percent of the top-seven token volume on OpenRouter, with Chinese models near zero. By April 2026, the two countries were at parity on the platform. The Chinese share rose to roughly 25 percent through the first half of 2025 on the strength of DeepSeek-V3 and the Qwen series, retracted as new U.S. frontier releases pulled developers back, and then re-accelerated sharply through the first quarter of 2026. Figure 4.6 shows the trajectory and marks several pivotal model releases.

[16] Wang and Siler-Evans, *U.S.-China Competition for Artificial Intelligence Markets*.

Figure 4.6. U.S. and Chinese Share of OpenRouter Token Volume

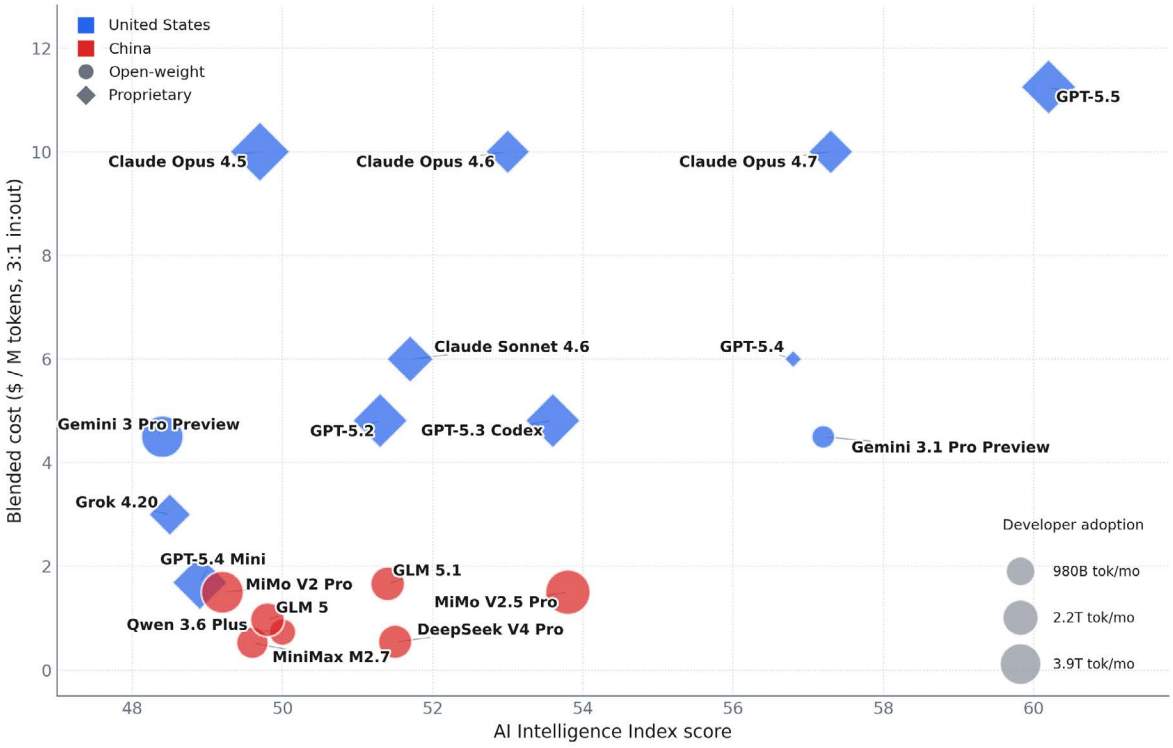


SOURCE: OpenRouter daily rankings (March 2026 onward) supplemented with Wayback Machine snapshots of OpenRouter rankings for the earlier period.
NOTE: Dashed lines mark notable model releases that shifted developer adoption.

At the endpoint of the chart, on April 30, 2026, the nine providers accounting for the most OpenRouter traffic are led by Anthropic, Google, and OpenAI from the United States, followed by Tencent, Moonshot AI, DeepSeek, Alibaba (Qwen), MiniMax, and Zhipu AI from China—so, six of the nine are Chinese. Among the Chinese providers, every flagship model is released open-weight. Among the three U.S. providers, only Google's Gemma model line is open-weight (and only at smaller sizes; Gemini frontier-tier models remain proprietary).

Why is the Chinese open-weight cohort displacing closed-API providers in this slice of the market? Part of the answer likely depends on the relationship between price and performance. Figure 4.7 plots the leading U.S. and Chinese frontier models against each other on two axes: an external intelligence-benchmark score and the blended cost per million tokens. The Chinese open-weight models cluster in a narrow band around $1–2 per million tokens at intelligence scores of roughly 48–54. The leading U.S. proprietary models span roughly $1–12 per million tokens at intelligence scores of roughly 48–60. At the top of the intelligence axis, the U.S. proprietary models are still ahead; for the bulk of developer-facing workloads, the Chinese open-weight cohort is delivering comparable measured intelligence at roughly one-quarter to one-fifth of the price.

Figure 4.7. Frontier Models: Blended Cost per Million Tokens Versus External Intelligence-Benchmark Score



SOURCE: Authors' calculations from OpenRouter pricing data and published model evaluations.
NOTE: Chinese open-weight models (red) cluster in a tight low-cost band at intelligence scores comparable with the bulk of the U.S. proprietary cohort (blue). Marker size scales with developer adoption.

Chapter 5

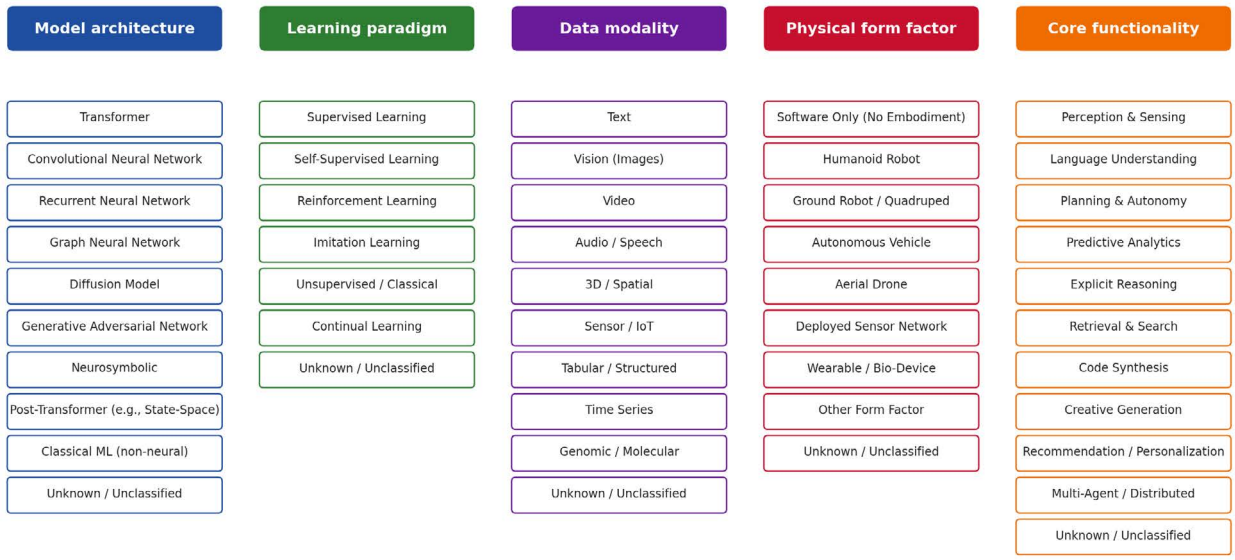
Technical Focus of U.S. and Chinese AI Developer Industries

In this chapter, we examine five technical dimensions of the same firm population: model architecture, learning paradigm, data modality, physical form factor, and core functionality. Finer-grain details on these dimensions are presented in Appendix A, the AI-driven process executed to classify firms along them is discussed in Appendix B, and core prompts are presented in Appendix C.

Five Technical Dimensions

Figure 5.1 summarizes the five technical dimensions used in this chapter and the categories under each. As discussed in Chapter 4, categories in a dimension are non-exclusive, and categorizations across dimensions are independent.

Figure 5.1. Technical Profile Dimensions and Categories



NOTE: IoT = internet of things.

Two data coverage notes, common to much of the analysis in the report, apply to the technical dimensions. First, technical disclosure is uneven across firms. Many firms describe their products but not their underlying model architecture or training methodology. The "unknown" rate for model architecture is 56 percent in the U.S. sample and 68 percent in the Chinese sample; for learning paradigm, the corresponding rates are 35 percent and 60 percent. Second, even when sufficiently granular information is available to classify firms across these variables, it is typically drawn from technical blog posts, white papers, and research publications. Therefore, our data are more likely to capture firms that publicly document their technical activities, creating a potential visibility bias against firms that disclose less.

Model Architecture

The model architecture dimension records the neural-network family (or families) underpinning a firm's AI work in the following categories:

- transformer
- convolutional neural network (CNN)
- recurrent neural network, graph neural network
- diffusion model, generative adversarial network
- neurosymbolic
- post-transformer (state-space and related)
- classical (non-neural) ML
- unknown

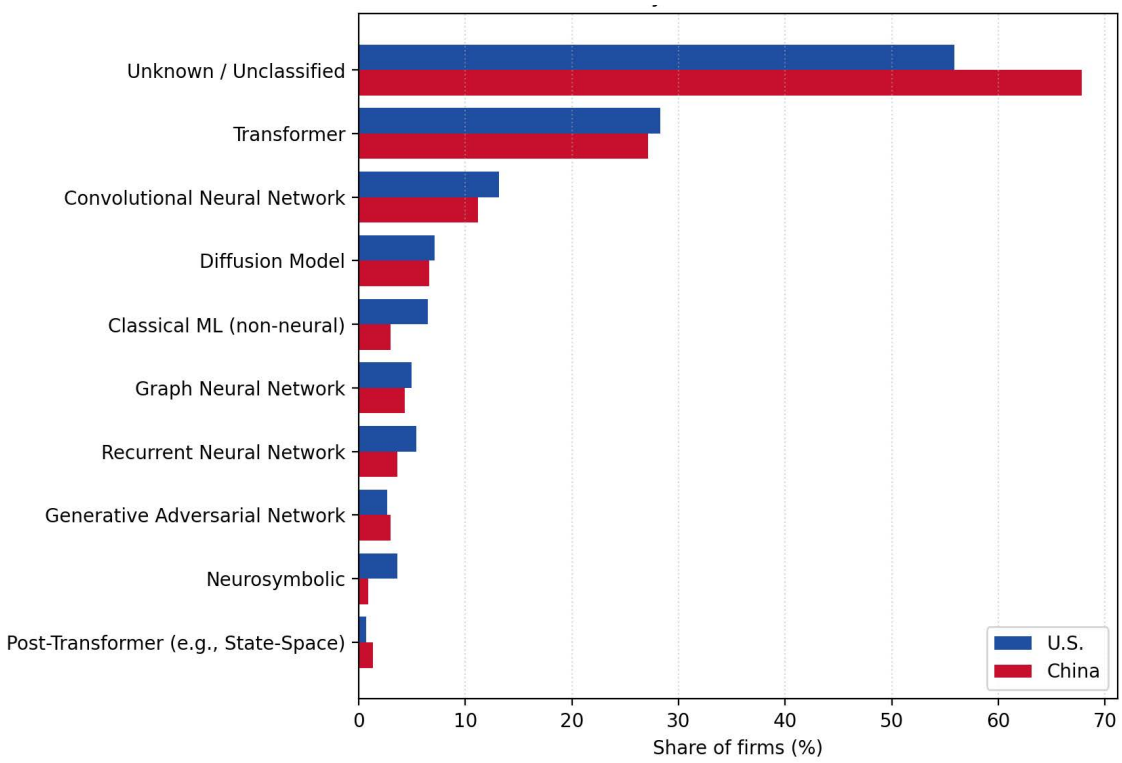
As already discussed, the category of unknown is the largest single share in both countries because many firms do not disclose architecture-level detail publicly.

The structure of our taxonomy was motivated by merging AI/ML technical taxonomies established by such conferences as NeurIPS and ICML, the arXiv preprint server, and the Stanford AI Index.[17] We combined, deduplicated, and blended these external taxonomies using insight from internal subject-matter experts; further guidance was provided by exploring the technical details available across our dataset.

Among the firms for which architecture could be confidently classified, transformer is the most prevalent category in both countries, tagged on 28 percent of U.S. firms and 27 percent of Chinese firms. Convolutional networks are second (13 percent and 11 percent, respectively), consistent with the continued prevalence of CNNs in vision and perception pipelines. Diffusion models appear at roughly 7 percent of U.S. firms and 7 percent of Chinese firms, principally in image and video generation. Overall, the architecture shares are very similar across the two countries. Figure 5.2 presents the share of U.S. and Chinese firms tagged with each architecture category.

[17] Stanford Institute for Human-Centered Artificial Intelligence, *The AI Index 2026 Annual Report*.

Figure 5.2. Distribution of Firms, by Model Architecture



Learning Paradigm

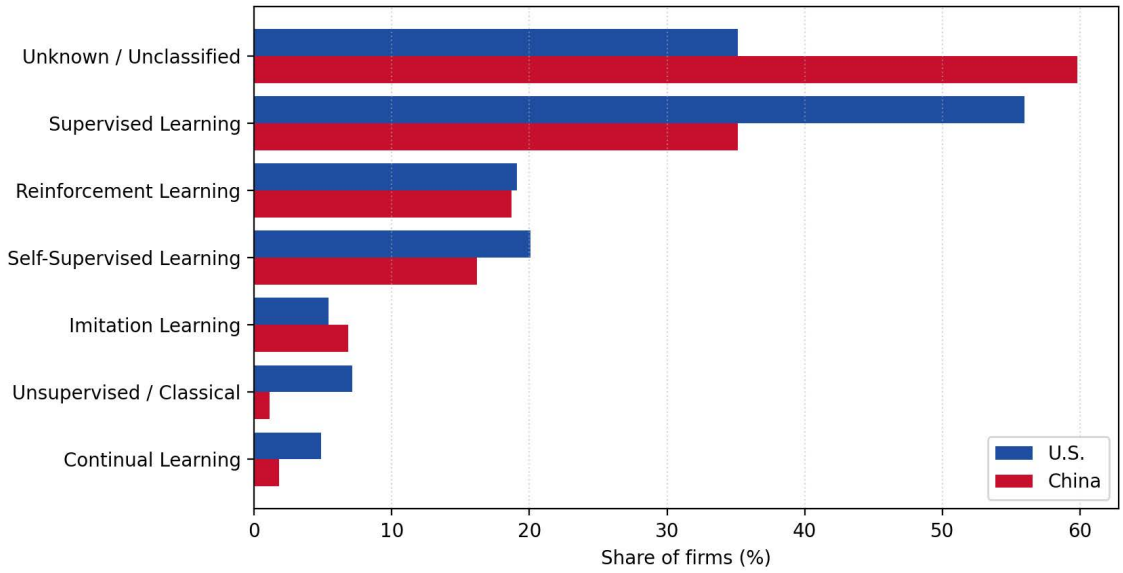
The learning paradigm dimension records how a firm trains its models. Following the taxonomy discovery process mentioned in the previous section, seven categories are distinguished:

- supervised learning
- self-supervised learning (e.g., next-token prediction)
- reinforcement learning (RL) (e.g., environment RL for robotics and preference-based RL for language models)
- imitation learning
- unsupervised or classical methods
- continual learning
- unknown

Figure 5.3 shows the distribution of firms across all learning-paradigm categories for each country.

Supervised learning is the most common paradigm in the U.S. sample (56 percent) and is substantially less common in the Chinese sample (35 percent). Self-supervised learning, the paradigm underlying LLM pretraining, is tagged on 20 percent of U.S. firms and 16 percent of Chinese firms, consistent with the similar foundation-developer cohort sizes described in Chapter 4. However, it is important to recognize for both of these claims that the elevated Chinese unknown rate (60 percent versus 35 percent) should be borne in mind.

Figure 5.3. Distribution of Firms, by Learning Paradigm



RL is an interesting category for comparison because its application is different in the two ecosystems. RL is tagged on 19 percent of U.S. firms and 19 percent of Chinese firms, but the firms apply this learning paradigm in distinct ways. After a qualitative read of firm-level information, on the U.S. side, the concept is most often associated with preference-based RL for language models (RL from human feedback [RLHF], direct preference optimization, group relative policy optimization-style reasoning training at OpenAI, Anthropic, and similar labs). On the Chinese side, the concept is more often associated with environment RL for robotics, autonomous-vehicle control, and industrial automation—as reflected in the joint distribution with the application domain taxonomy.

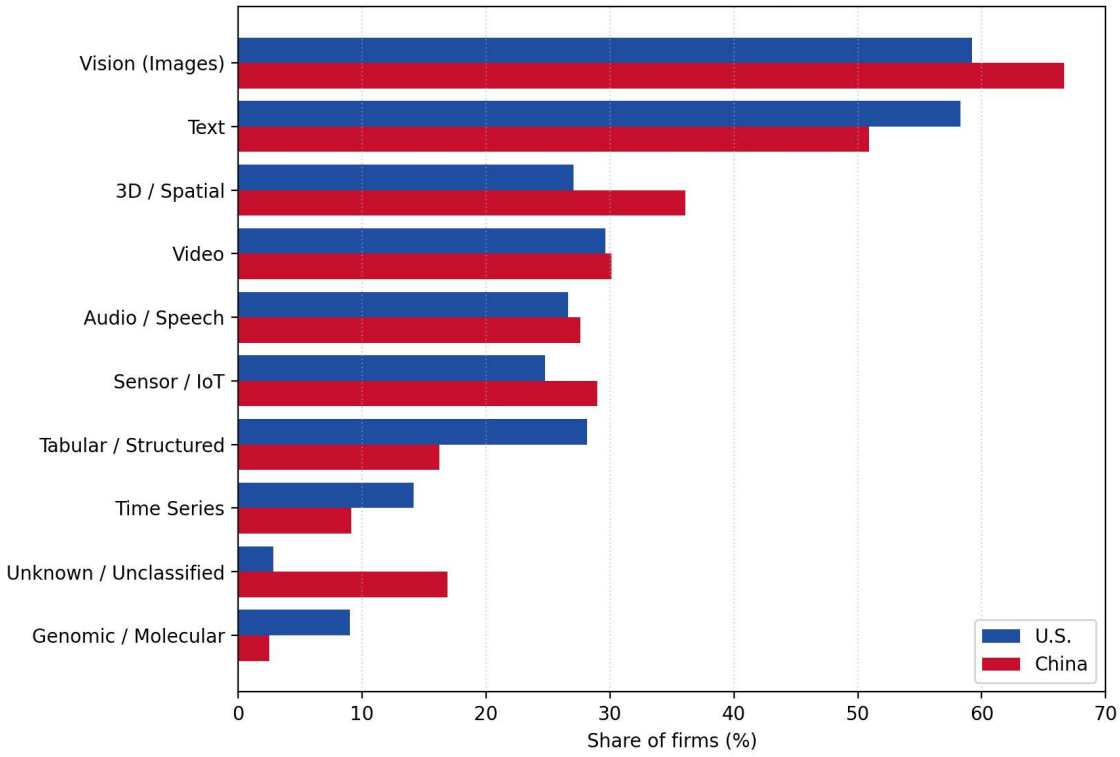
Data Modality

The data-modality dimension records the kinds of data a firm's models consume. Categories are the following:

- text
- vision (images)
- video
- audio or speech
- three-dimensional (3D) or spatial data
- sensor
- IoT signals
- tabular structured data
- time series
- genomic and molecular data
- unknown.

Vision and text are the two most prevalent modalities in both ecosystems, and they have similar shares: Vision is tagged on 60 percent of U.S. firms and 67 percent of Chinese firms, and text is tagged on 58 percent and 51 percent, respectively. Video and audio are similar across the two countries as well. Figure 5.4 presents the share of U.S. and Chinese firms tagged with each data modality. Although the rate of Chinese firm data missingness across this variable is nontrivial at 17 percent, the degree of missingness is smaller than that of the other variables in this chapter, suggesting comparative assessments are more robust here than for other variables.

Figure 5.4. Distribution of Firms, by Data Modality



The data-modality dimension provides additional evidence of China's strength in embodied AI. Chinese firms have large prevalence percentages in the 3D or spatial and the sensor or IoT modalities most directly associated with embodied systems. For example, 3D or spatial data is tagged on 36 percent of Chinese firms versus 27 percent of U.S. firms, and sensor or IoT is tagged on 29 percent versus 25 percent. On the other hand, for the tabular or structured, time-series, and genomic or molecular data modalities (which are more closely related to knowledge-work), U.S. firms exhibit slightly larger prevalence percentages than Chinese firms, although most differences are in the noise floor set by data missingness.

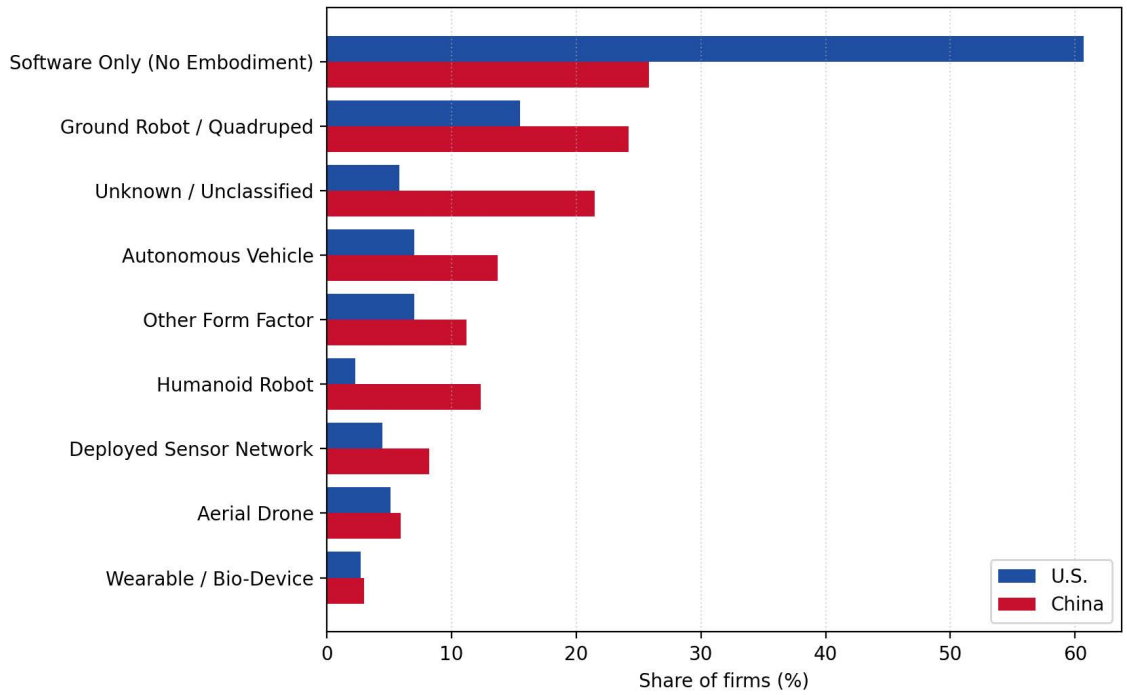
Physical Form Factor

The physical form factor dimension records what physical instantiation, if any, a firm's AI is delivered in for the following categories:

- software-only (no embodiment)
- humanoid robot
- ground robot or quadruped
- autonomous vehicle
- aerial drone
- deployed sensor network
- wearable or bio-device
- other form factor
- unknown.

Figure 5.5 shows the distribution of firms across these form-factor categories for each country.

Figure 5.5. Distribution of Firms, by Physical Form Factor



Physical form factor is the dimension on which the two ecosystems diverge most sharply. Sixty-one percent of U.S. firms in our dataset are software-only versus only 26 percent of Chinese firms. The Chinese unknown share on this dimension is 22 percent, and because positively assigned counts of Chinese firms should be treated as lower bounds (Appendix B), the true software-only gap is likely real in direction but could be considerably smaller in magnitude than the observed figures suggest.

The measured distribution of embodiment modes reflects that humanoid-robot firms are more than five times as prevalent in the Chinese sample (12 percent) than in the U.S. sample (2 percent). Similarly, ground robots and quadrupeds are tagged on 24 percent of Chinese firms versus 16 percent of U.S. firms, and autonomous vehicle firms are tagged at 14 percent versus 7 percent.

This difference is consistent with the application-domain asymmetry discussed in Chapter 4. Chinese overrepresentation in manufacturing, transportation, and energy verticals is mediated through embodied form factors—humanoid and quadruped robots for manufacturing inspection and material handling, autonomous vehicles for transportation, and sensor-network deployments for grid and energy monitoring. The U.S. concentration in health care, scientific research, cybersecurity, and developer tooling, by contrast, is overwhelmingly software-only.

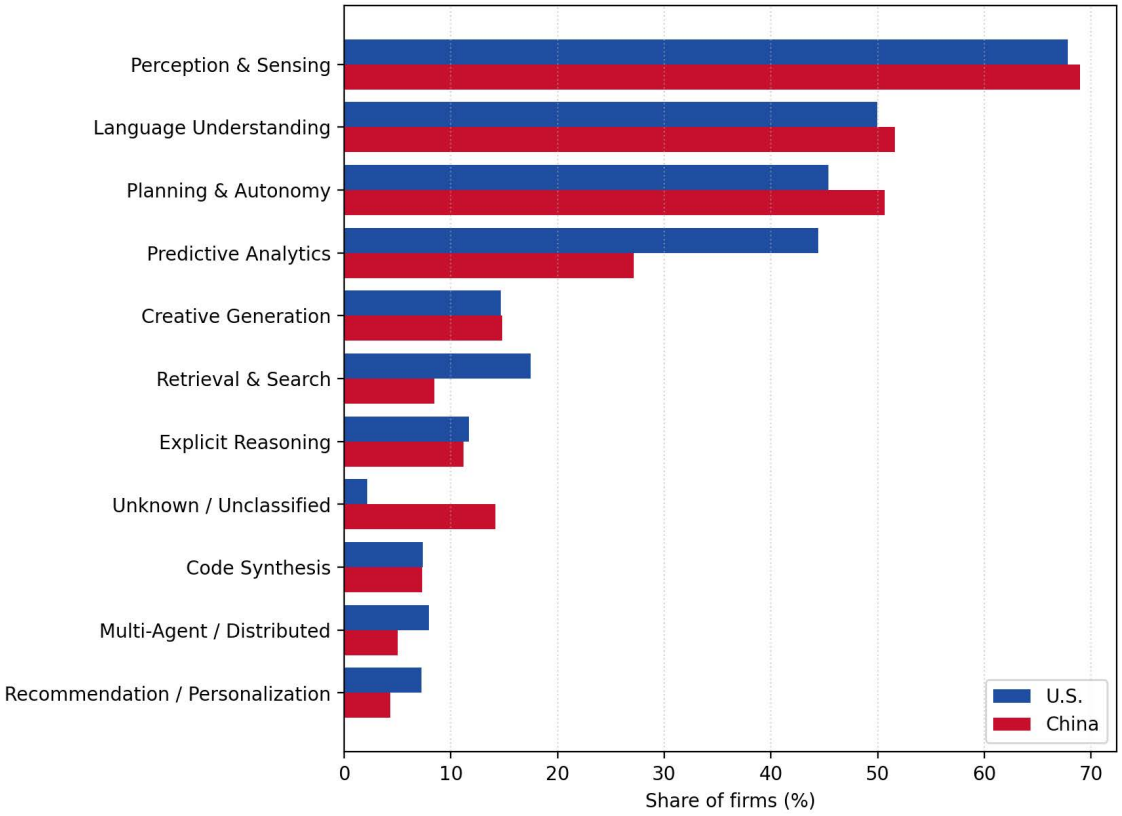
Core Functionality

The core functionality dimension records what an AI system is being used to do. The categories are not mutually exclusive—a product can perform both perception and planning, or both language understanding and code synthesis—and the classifier assigns all categories that are clearly evidenced. The category list is as follows:

- perception and sensing
- language understanding
- planning and autonomy
- predictive analytics
- explicit reasoning
- retrieval and search
- code synthesis
- creative generation
- recommendation/personalization
- multi-agent/distributed
- unknown/unclassified.

Figure 5.6 presents the share of U.S. and Chinese firms tagged with each core functionality category.

Figure 5.6. Distribution of Firms, by Core Functionality



Most categories show only modest U.S.–China differences. Perception and sensing, language understanding, and planning and autonomy are the three most prevalent functionalities in both ecosystems, each tagged on roughly one-half or more of firms. The largest apparent U.S.–China gaps appear in predictive analytics (44 percent of U.S. firms versus 27 percent of Chinese firms) and retrieval and search (17 percent versus 8 percent), both of which lean toward knowledge-work use cases consistent with the U.S. vertical mix. However, because positively assigned counts of Chinese firms should be treated as lower bounds (Appendix B), the true magnitude of both gaps might be considerably smaller than shown. By contrast, the China-leading gap on planning and autonomy appears more robust and is consistent with China's national emphasis on robotics and embodied AI.

Chapter 6

Conclusion

This report provides a systematic empirical foundation for analyzing the AI developer firm ecosystems in the United States and China. By constructing and analyzing a dataset of more than 1,100 AI developer firms across both countries, we aimed to ground claims about ecosystem structure, concentration, and comparative advantage in observable firm-level data rather than inference from national plans, expert intuition, or the activities of a handful of prominent firms.

One of the central motivating concerns in the debate over U.S.-China AI competition is whether the United States is insufficiently diversified in its approach to AI development relative to China's approach. China's national AI strategies have emphasized diversification through such actions as a formal designation of embodied AI as a national development priority and promotion of neuromorphic computing and other post-transformer architectures via government-funded universities and research centers. Our data shed light on how this agenda is—and is not—reflected in the AI developer firm ecosystem. In our dataset, we find no substantial evidence that Chinese AI developer firms are pursuing neuromorphic or post-transformer architectures to any greater degree than their U.S. counterparts. Rather, in our study sample, both ecosystems are transformer-led and have nearly identical distribution of firms across other architectures we explored.

Chinese AI developer firms diverge most clearly from their U.S. counterparts in the physical instantiation of AI. Only 26 percent of Chinese firms are software-only, compared with 61 percent of U.S. firms. Chinese firms are far more likely to deliver AI through humanoid robots, ground robots and quadrupeds, and autonomous vehicles, and they are overrepresented in physical-economy verticals, such as manufacturing, transportation, and energy. Chinese firms' RL activity is more often based in a physical environment and centers on application to robotics locomotion, autonomous-vehicle control, and industrial automation—a contrast to the preference-based RL that is dominant in the United States.

Taken together, our findings offer a more empirically grounded assessment of the U.S.-China AI competition than has previously been available. The strongest version of the concentration concern—that the United States pursues a single architectural approach while China pursues many—is not well supported at the level of model architecture by our dataset. However, we do observe that the U.S. AI developer firm ecosystem is more oriented toward software-delivered, language-model-centric AI, whereas the Chinese AI developer firm ecosystem has made structurally larger bets on embodied AI, physical-world applications, and the integration of AI into manufacturing, transportation, and industrial systems.

If the path to transformative AI capabilities ultimately requires codevelopment with robotics, world models, and physical-environment learning, China's heavier investment in physical form factors in the AI developer firm ecosystem represents a meaningful hedge that the U.S. ecosystem lacks.

Without systematic cross-national firm-level data, this distinction would be difficult to draw or support with empirical rigor.

As AI development accelerates and geopolitical competition intensifies, empirical grounding becomes more critical. The findings presented here suggest that U.S.-China AI competition is less a race between monolithic national strategies than a complex interplay of diverse, dynamic ecosystems pursuing overlapping yet distinct pathways. Understanding these pathways requires systematic firm-level analysis. This report provides that foundation, but the data will require continual updating as both ecosystems evolve.

Appendix A

Technical Taxonomy Description

We tag each AI firm in our sample against a structured nine-dimension taxonomy of AI development activity. In Table A.1, we present each taxonomy dimension, the definition of the dimension itself, and the subcategories in it. These are the definitions that were ingested by the LLM agents involved in our classification loop. We consolidate or truncate our definitions in certain cases, for brevity; these entries end in ellipses (. . .).

Table A.1. Taxonomy of AI Development Technical Dimensions

Dimension and Definition	Possible Categories
Model Provenance What does this company DO with AI models? Captures the company's primary model-production activity—whether they train foundation models from scratch, train narrow task-specific models from scratch, adapt existing pretrained models, or integrate third-party models without weight modification.	• **Foundation model developer:** Company that trains AI models from scratch at significant compute scale, designed for REUSE across multiple downstream tasks, customers, or applications. • **Task-specific model developer:** Company that trains narrow neural networks from scratch for a SINGLE well-defined application, without starting from a large pretrained foundation model. • **Model adapter (fine-tunes, distills, quantizes existing models):** Company that takes EXISTING pretrained models (their own or third-party) and modifies the weights through fine-tuning, instruction tuning, distillation, LoRA, quantization, alignment training, continued pretraining, or model merging. • **Model integrator (uses third-party models without weight modification):** Company that builds sophisticated AI products on top of third-party foundation models WITHOUT modifying model weights. • **Unknown/not disclosed:** Model provenance cannot be determined from available evidence. Use as fallback when it is unclear whether the company trains, adapts, or integrates models.

Dimension and Definition	Possible Categories
Model Architecture The fundamental computational structure of the AI model, independent of training method, task, or application. Answers: "What kind of neural architecture (or non-neural algorithm) is this company actually building or modifying?"	• **Transformer architecture:** Transformer architectures across any modality—text (GPT, LLaMA, Claude, BERT, T5, Qwen, DeepSeek, Mistral), vision (ViT, DINOv2, Swin, MAE, SigLIP), or multimodal (GPT-4V, Gemini, Qwen-VL, LLaVA, InternVL). • **Convolutional neural networks:** Architectures built primarily on convolutional operations for spatial or grid-structured data. Includes ResNet, EfficientNet, YOLO, U-Net, MobileNet, and similar CNN families. • **Post-transformer architectures:** Neural sequence architectures that are explicit alternatives to the standard transformer. • **Recurrent neural networks:** Sequential architectures with explicit recurrent connections (LSTM, GRU). Largely superseded by transformers for modern language tasks but still used in time-series, signal processing, and legacy industrial systems. • **Graph neural networks:** Architectures designed to operate on graph-structured data through message passing, graph convolutions, graph attention, or spectral methods. Common in drug discovery, molecular modeling, knowledge graphs, and network analysis. • **Diffusion and flow-matching models:** Generative architectures that learn to reverse a noise-adding process or learn continuous flow trajectories. Includes DDPM, score-based generative models, stable diffusion, flow matching, and rectified flow. • **Generative adversarial networks:** Architectures using adversarial training between generator and discriminator networks. Includes StyleGAN, CycleGAN, and similar GAN families. • **Neurosymbolic architectures:** Systems that explicitly integrate neural network components with formal symbolic reasoning — theorem provers, constraint solvers, satisfiability solvers, probabilistic programming, or formal logic engines. • **Classical machine learning:** Non-deep-learning approaches including decision trees, random forests, gradient boosting (XGBoost, LightGBM, CatBoost), SVMs, linear models, and ensemble methods. • **Unknown/not disclosed:** Architecture cannot be determined from available evidence. This is the DEFAULT classification when the evidence does NOT explicitly name an architecture, model family, or technique. USE THIS LIBERALLY.

Dimension and Definition	Possible Categories
Commercial Positioning How the AI is offered to market—what the company actually sells and how it captures value. Modeled on classical cloud-service layering (IaaS → PaaS → SaaS) adapted to AI, with a Model-as-a- Service layer for companies that train and sell their own models, a distinct Inference-as-a-Service layer for companies that serve models efficiently (often open-weight models they did not train), an Applied AI Product layer for finished AI-powered solutions (which now includes AI infrastructure / dev tooling sold to ML engineers), and an Internal Use category for companies that build AI for their own operations without commercializing it.	• **Model-as-a-Service:** The company trains its own AI models and sells access to them — via API, downloadable weights, or managed inference endpoints. • **Inference-as-a-Service:** The company provides optimized inference infrastructure for running AI models — typically open-weight models trained by others, though some providers also host their own fine-tunes. • **Applied AI Product:** A finished product or service whose core value proposition IS the AI's output, sold or offered to external customers. • **Internal Use / Operational AI:** The company develops AI primarily for its own internal operations and does NOT commercialize the AI as a standalone external offering. • **Unknown/not disclosed:** Commercial positioning cannot be determined from available evidence. Use this as the fallback when the company's go-to-market model is unclear.
Learning and Optimization Paradigm How the system learns from data and makes decisions over time. This captures the optimization or control logic.	• **Supervised learning:** Learning a mapping from inputs to outputs using human- or automatically-labeled examples. Includes both classification (discrete labels) and regression (continuous targets), and covers fine-tuning on labeled downstream data. • **Self-supervised learning:** Learning representations from unlabeled data by constructing a pretext task from the data itself (e.g., next-token prediction, masked reconstruction, or contrastive invariance). • **Unsupervised learning (classical):** Classical unlabeled-data methods that discover structure without constructing a learned pretext task: clustering (k-means, GMM, hierarchical), dimensionality reduction (PCA, t-SNE, UMAP), density estimation, anomaly/outlier detection, topic. . . • **Reinforcement learning (environment RL and RLHF):** Training models through reward signals — covering BOTH environment RL (agents interacting with simulated or real environments, learning from environment rewards) AND preference-based RL (LLM alignment via human or AI preferences, used to sh. . . • **Imitation learning:** Learning a policy by supervised fitting to expert demonstrations of the desired behavior. • **Continual or lifelong learning:** Incrementally updating a deployed model's weights on new data or tasks over time while explicitly mitigating catastrophic forgetting of prior capabilities. • **Unknown/not disclosed:** Learning paradigm is not specified.

Dimension and Definition	Possible Categories
Core Functionality Capability The broad type of intelligence or capability the system provides, independent of application domain, user interface, or delivery mechanism. Answers the question: "What kind of AI problem does this company solve?"	• **Perception and sensing:** Extracting structured information from raw sensor data—images, video, audio, lidar, radar, thermal, depth, IMU, or other physical signals. • **Language understanding and generation:** Processing, interpreting, and generating human language in text form. Covers the full NLP spectrum: text classification, named entity recognition, translation, summarization, question answering, conversational language generation, and large. . . • **Code and software synthesis:** Generating, completing, understanding, or analyzing source code as a primary task. Includes code completion assistants, code generation models, automated refactoring, code review AI, and specialized code LLMs. • **Creative content generation:** Generating novel media content—images, video, audio, music, 3D assets, animations, or multimodal creative output. • **Explicit reasoning and inference:** Explicit multi-step reasoning as a primary capability. Includes: (a) specialized reasoning models trained for chain-of-thought, extended thinking, or problem-solving (e.g., OpenAI o1/o3, DeepSeek R1, Claude extended thinking, Gemini thinki. . . • **Predictive analytics and statistical prediction:** Statistical prediction from learned models where the primary output is a numerical prediction, probability, or class label. • **Recommendation, ranking, and personalization:** Ranking, matching, and personalizing content, products, or actions for individual users or contexts. • **Retrieval and search:** Finding and retrieving relevant information from a corpus in response to a query or context. • **Planning, control, and autonomy:** Selecting and executing sequences of actions in dynamic environments to achieve goals. • **Multi-agent and distributed intelligence:** Coordinated intelligence involving multiple AI agents that communicate, cooperate, or compete to accomplish a goal. • **Unknown/not disclosed:** Functional capability cannot be determined from available evidence. Use this as the fallback when the company's core AI capability is unclear.

Dimension and Definition	Possible Categories
Physical Form Factor If the company deploys or integrates AI with physical hardware, what is the form factor? This dimension captures the physical embodiment type for companies that build AI-powered hardware systems. Software-only companies are tagged 'not_applicable'.	• **Humanoid robot:** Bipedal or human-form robots designed to operate in human environments. The defining physical signature is a human-like body plan (legs, arms, torso) enabling locomotion and manipulation in spaces designed for humans. • **Aerial drone/UAV:** Unmanned aerial vehicles and drone systems with AI-powered autonomy, perception, or decision-making. • **Autonomous vehicle:** Self-driving cars, trucks, buses, robotaxis, and other road vehicles with AI-powered autonomous driving. • **Ground robot (non-humanoid, non-vehicle):** All non-humanoid, non-road-vehicle ground robots with AI. Covers wheeled/tracked delivery robots, warehouse mobile robots, industrial robotic arms and manipulators, patrol and security robots, cleaning robots, cobots, quadrupeds, agricultur. . . • **Deployed sensor network:** Fixed sensing infrastructure that the company deploys in the physical world—camera networks, lidar arrays, radar installations, environmental sensor grids, acoustic monitoring systems. • **Wearable or biodevice:** AI-powered wearable devices, biosensors, or brain-computer interface hardware. Covers smartwatch AI, fitness/health wearables with on-device ML, EMG/EEG wristbands, BCI headsets, and medical wearable sensors with AI processing. • **Other physical form factor:** Physical AI hardware that does not fit the categories above. Covers underwater ROVs/AUVs, space robotics, agricultural machinery with AI, mining equipment with AI, and other specialized physical platforms. • **Form factor unknown:** The company operates physical AI hardware but the specific form factor cannot be determined from the evidence. • **Not applicable (software only):** ONLY use when there is POSITIVE evidence the company is software-only (SaaS, API, web app, mobile app, cloud platform).

Dimension and Definition	Possible Categories
Application Domain The real-world sector, industry, or context in which the AI system is primarily deployed or intended for use. Answers: "What domain does this company's customers come from, and what real-world problem does the AI solve?"	• **Agriculture and food systems:** Farming, crop management, livestock monitoring, precision agriculture, food production and supply chain, agricultural robotics, soil and yield analysis. • **Banking, finance, and insurance:** Financial services—banking, payments, lending, trading, asset management, risk and credit modeling, fraud detection, insurance underwriting, regulatory compliance for financial institutions, fintech products. • **Education and learning:** AI for teaching, learning, tutoring, curriculum, assessment, academic content generation, K-12 and higher education tools, corporate training and learning. • **Energy and utilities:** Power generation, transmission, and distribution; grid operations and forecasting; renewable energy management; oil and gas exploration and production; utility operations; smart grid and demand-response systems; battery and storage optimiza. . . • **Healthcare and life sciences:** Clinical care AI, medical imaging, diagnostics, electronic health records, hospital operations, telemedicine, drug discovery and pharmaceutical R&D, clinical trials, genomics and precision medicine, mental health, biotechnology. • **Legal services:** Legal research, contract analysis and drafting, e-discovery, legal compliance, litigation support, legal practice management AI. Companies whose primary customers are law firms, in-house legal teams, or compliance functions. • **Logistics and supply chain:** Supply chain optimization, warehousing automation, last-mile delivery, freight and shipping, inventory management, demand planning for supply chains, customs and trade, fleet management. • **Manufacturing and industrial:** Industrial production, factory automation, quality inspection and defect detection, predictive maintenance for industrial equipment, process optimization for chemical/ materials/heavy industry, robotic manufacturing, industrial IoT. • **Media, entertainment, and creative industries:** Creative content generation (text, image, video, music, 3D), gaming and game development AI, streaming and recommendation, advertising and marketing creative, journalism AI, post-production and visual effects, social media content. • **Military and defense:** Military and defense applications—autonomous defense systems, surveillance and ISR (intelligence, surveillance, reconnaissance), defense logistics, command and control, electronic warfare, missile defense, defense intelligence analysis, m. . . • **Real estate and construction:** Real estate transactions, property management, construction AI, building information modeling (BIM), architecture and design AI, smart buildings, property valuation. • **Retail and e-commerce:** Retail operations, e-commerce platforms, recommendation and personalization for shoppers, store-level inventory management, omnichannel retail, dynamic pricing, virtual try-on, store automation. • **Science and research:** Scientific research and discovery—AI for scientific experimentation, materials discovery, climate and earth science, physics, chemistry, biology research (basic science as opposed to applied medicine), astronomy and space science, scienti. . . • **Security and cybersecurity:** Cyber threat detection and response, security operations (SOC) automation, malware analysis, network security, identity and access management AI, fraud detection for security purposes, vulnerability assessment, security analytics.

Dimension and Definition	Possible Categories
	• **Telecommunications and networking:** Telecom network operations, 5G and network optimization, network management, customer service for telcos, billing and OSS/BSS systems, RAN intelligence, fiber/network planning. • **Transportation and mobility:** Autonomous vehicles, ride-hailing and mobility platforms, public transit AI, aviation and aerospace AI, maritime and shipping route AI, traffic management, fleet routing and dispatch, automotive ADAS. • **Cross-domain or general-purpose:** General-purpose AI INFRASTRUCTURE or foundation technology designed for use across many industries. • **Other identifiable domain:** Use ONLY when the evidence clearly identifies the company's application domain but it does not fit any of the listed categories above (e.g., religion, sports, hospitality, art, philanthropy, niche consumer hardware verticals). • **Unknown/insufficient evidence:** Use ONLY when the evidence is insufficient to determine the company's application domain. The classifier could not identify which industry or use case the AI primarily serves from the available evidence.
Data Modality What types of data does the company's AI product primarily process? This dimension captures the data modalities the company's CORE PRODUCT operates on—not what architecture it uses internally, and not the modality of training data alone. Answers: "When this company's AI does its job, what kind of data is it consuming?"	• **Vision (still images):** Still image data—photos, medical imaging (X-rays, MRI, pathology slides, ultrasound), satellite and aerial imagery, document images, individual video frames analyzed as images. • **Text (natural language and code):** Natural language text (documents, articles, chat messages, social media posts, emails), source code (as text tokens), and other symbol sequences processed as language. • **Audio (speech and sound):** Audio signals—speech, music, environmental sound, industrial acoustic monitoring. • **Video (temporal image sequences):** Video data—temporal sequences of image frames with motion. Covers video understanding (action recognition, scene understanding over time, surveillance analytics), video generation (text-to-video, image-to-video, video-to-video), and real-. . . • **3D and spatial data:** 3D point clouds (from LiDAR, depth sensors, photogrammetry), 3D meshes, volumetric data (voxel grids, medical CT/MRI volumes treated as 3D), depth maps, geospatial data, and SLAM/reconstruction data for robotics and AV perception. • **Tabular and structured data:** Structured relational data—databases, spreadsheets, financial records, transaction logs, clinical records, enterprise resource data, customer relationship data. Data organized as rows and columns with typed fields. • **Time series (temporal numeric streams):** Structured numeric streams indexed by time where the TEMPORAL PATTERN is the primary signal. • **Genomic and molecular data:** DNA and RNA sequences, protein structures and sequences, molecular graphs, chemical compound representations (SMILES, molecular fingerprints), drug compound data, and related bio/chem structured data. • **Raw sensor and IoT data:** Raw or minimally-processed multi-channel signals from physical sensors—IMU (accelerometer, gyroscope, magnetometer), radar, sonar, industrial sensor networks (pressure, temperature, vibration, flow), EMG and biosensor arrays, environmenta. . . • **Unknown / insufficient evidence:** Use when the evidence is insufficient to determine which data modality the company's AI primarily processes. The classifier could not identify a specific modality from the available evidence.

Dimension and Definition	Possible Categories
Open-Source Status Does the company publish its own AI models as open source? This is binary by design—exactly ONE of the two categories should be tagged per company.	• **Publishes open-source models:** The company has released one or more of its OWN AI models for download—model weights, training code, or runnable inference code, on HuggingFace, GitHub, or the company's own site. • **Proprietary only (does not publish open-source models):** The company does NOT release its own AI models for download. All AI models are proprietary, accessed via API, hosted product, or used internally.

NOTE: Because this material was ingested by the LLM agents, we have reprinted it here verbatim, although we consolidate or truncate certain passages that end in ellipses (. . .). We also intentionally did not define abbreviations because the specific terms are not essential for understanding the general concept being presented.

Each dimension carries an "unknown" fallback so that low-evidence firms are flagged rather than guessed. In addition to the information already presented, our taxonomy for every category provides anchor examples, strong and moderate signal phrases, and explicit exclusion criteria (the latter blocking such common false-positive patterns as inferring transformer architecture from the existence of a chatbot or inferring foundation-model status from marketing language).

Appendix B

Technical Taxonomy Assignment

What We Collected for Each Firm

When assigning the technical properties to each firm, we assembled two streams of evidence for each firm in the dataset and used them as inputs to an LLM-assisted classifier. The streams were as follows:

- **web evidence collected through Perplexity's web-search interface.** Each firm received roughly 30 targeted web queries covering its products, model releases, training methodology, open-weight posture, and commercial offerings. The retrieved sources span vendor product pages, press releases, technical blog posts, GitHub repositories, Hugging Face model cards, news coverage, and similar public material. A separate LLM pass then classified each retrieved source by type (peer-reviewed paper, preprint, technical blog, product documentation, marketing copy, and so on) so that more-authoritative sources would carry more weight in the classifier's downstream aggregation.
- **academic papers retrieved from OpenAlex.** We pulled the firm's authored or affiliated papers in AI venues over the 2021–2025 period and applied a relevance filter to remove misattributed work. Most firms with active publication programs contributed around five to ten papers after filtering.

For each node in our taxonomy, the LLM classifier assigned which categories were most relevant to each firm along with a *relevance score* based on what information was available to support the classification. For example, information based on peer-reviewed articles or conference papers is considered high signal—and thus boosts relevance scores—whereas information based on marketing copy is treated as fairly low signal and is thus weighted low in relevance. Relevance scores were fed alongside each block of retrieved information to an LLM classifier to make a final assessment for each taxonomy category; the relevance scores helped the LLM adjudicate decisions when encountering conflicting or inconsistent information.

Two caveats follow from the evidence-collection process and apply throughout. The first is that the underlying evidence is dominated by English-language sources; firms with small English-language web footprints (disproportionately Chinese firms) surface less evidence per query. The second is that Chinese commercial AI activity is, on average, more opaque than U.S. activity by virtue of institutional differences. A meaningful share of Chinese AI development occurs inside state-owned enterprises, regulated utilities, and industrial groups whose public disclosure of AI work is limited. Throughout the study, the result is that the "unknown" share of Chinese firms is consistently higher than the U.S. share across positively assigned dimensions. Readers should therefore treat counts of Chinese firms for any positively assigned category as lower bounds.

Two Worked Examples: DeepSeek and Abridge

To illustrate the classification process across technical variables, we present two case study examples for the firms DeepSeek and Abridge, displaying a subset of assigned taxonomy nodes along with the LLM-supplied reasoning traces. The two firms were chosen because they sit at almost opposite corners of the technical dimensions: DeepSeek is a Chinese frontier-model developer with a model-as-a-service commercial model and an open-weight release strategy; Abridge is a U.S. vertical-AI product firm whose offering is a finished workflow for clinical documentation. Reading the two profiles together shows how the dimensions interact in a single firm and how the classifier treats the supporting evidence.

DeepSeek (China)

DeepSeek pretrains general-purpose LLMs from scratch and sells access to those models under paid usage through its API while also releasing model weights and inference code publicly through Hugging Face and GitHub under permissive licenses. The company's headline product line is the text-only DeepSeek V3, V3.2, or V4 series of MoE (mixture of experts) architectures, together with the DeepSeek-R1 reasoning model trained primarily through RL. Beyond the headline text models, DeepSeek has also released a smaller multimodal line (DeepSeek-VL and DeepSeek-VL2 for vision-language understanding, and the Janus, Janus-Pro, or JanusFlow models for unified image understanding and generation; this line is the basis for the secondary perception-and-sensing tag in Table B.1. The headline V or R series is text-only; the perception capability sits as a secondary research line, which is reflected in the lower relevance score that the classifier assigned.

Table B.1. DeepSeek—Assigned Categories and LLM Reasonings for Each Category

Dimension	Assigned Category	LLM Reasoning
Commercial positioning	• model_as_a_service • applied_ai_product	The primary commercial mode is selling API access to the DeepSeek model family under paid usage; a consumer chat product accounts for the secondary Applied AI Product tag.
Model provenance	• foundation_developer • model_adapter	The firm pretrains its own base models at scale (foundation developer); it also releases distilled checkpoints fine-tuned on top of Qwen 2.5 and Llama 3 bases (model adapter).
Application domain	• cross_domain	Models are general-purpose. The firm sells API access used across verticals rather than targeting one industry; hence a single Cross-Domain tag.
Core functionality	• language_understanding • reasoning_decision • code_synthesis planning_autonomy • creative_generation • perception_sensing	Text-only LLM workloads dominate (language understanding, reasoning, code synthesis, planning or agentic). The lower-relevance perception tag reflects DeepSeek-VL and Janus, a smaller multimodal line distinct from the V/R flagship.
Open-weight status	• publishes_open_source	All released models are open-weight under permissive licenses (MIT or similar) on GitHub and Hugging Face.

Abridge (United States)

Abridge sells a finished AI workflow rather than a model. Its product converts physician–patient conversations into structured clinical notes that integrate with such electronic health record systems as Epic. The model stack behind the product combines a custom speech-recognition system tuned for medical conversations with fine-tuned foundation models for summarization and structured-note generation; no components are sold or released as stand-alone models. The firm publishes no model weights and no code.

Table B.2. Abridge—Assigned Categories and LLM Reasonings for Each Category

Dimension	Assigned Category	LLM Reasoning
Commercial positioning	• applied_ai_product	Customers pay for the finished documentation workflow, integrated into existing electronic health record systems. The AI is internal to the product, not an item being sold.
Model provenance	• model_adapter • task_specific_developer	The firm adapts general-purpose foundation models with proprietary medical fine-tuning data (model adapter); it also trains custom speech-recognition models from scratch for medical conversations (task-specific developer).
Application domain	• healthcare	This is single-vertical; it is clinical documentation for health care providers. There is no claim of general-purpose applicability.
Core functionality	• language_understanding • perception_sensing • retrieval_search	The functionality is language understanding (note generation), perception (medical speech recognition with diarization), and retrieval against medical literature.
Open-weight status	• proprietary_only	There are no public model releases, no public code repository, no public datasets. It is a fully proprietary commercial model.

Appendix C

Technical Taxonomy Key Prompts

Here, we provide the core LLM workhorse prompt that was leveraged to assign taxonomy properties to each firm along with the prompt we used to extract information from OpenAlex publications. We truncate portions of the prompt for brevity but provide the most-central components.

Company-Level Classification Prompt (Gemini 2.5 Pro)

This prompt was leveraged to classify a company's posture given information sourced from a Perplexity Sonar search. This information was later merged with artifact-level classifications (discussed in the next section) to produce a final composite profile.

Company-Level Prompt

```
You are an expert AI industry analyst. Classify the company "{COMPANY}" across the
following taxonomy dimensions based on the evidence provided.

## TAXONOMY DIMENSIONS
{taxonomy summary — dimension name, description, multi-coding guidance,
 and for each category: ID, name, and definition (truncated to 200 chars)}

## EVIDENCE ABOUT {COMPANY}
### Synthesized Search Results:
{Perplexity Sonar synthesized answers from 4–5 queries}

### Additional Source Content:
{top 20 citation snippets, each up to 400 chars}

## INSTRUCTIONS
For each dimension, provide:

   1. A list of categories that apply, each with its own relevance score (0.0–1.0)
      indicating how central that category is to the company's PRIMARY activities. Rank
      them from most to least relevant.
   2. An overall dimension-level confidence score (0.0–1.0) reflecting how confident you
      are in the evidence quality for this dimension.
   3. Brief reasoning citing specific evidence.


### Relevance scoring guidance:
     — 1.0: Core to the company's identity and primary business
     — 0.7–0.9: Significant, well-evidenced activity
     — 0.4–0.6: Secondary or emerging activity with some evidence
     — 0.1–0.3: Minor, tangential, or weakly evidenced
     — Only include categories with relevance >= 0.3

### Evidence freshness:
     — Strongly prefer recent evidence (2024–2026) over older sources
     — If older evidence contradicts recent, go with recent
     — If only evidence older than 2023 is available, note "[outdated evidence]", cap
       dimension confidence at 0.4, and prefer the *_unknown primary for that
       dimension

### Classification rules (ENFORCE STRICTLY):

   1. PROHIBITED INFERENTIAL PHRASES — When a primary tag is NOT an *_unknown category,
      reasoning must cite direct evidence. Do NOT use "strongly implies", "almost
      certainly", "typically refers to", "would likely use", "is characteristic of", "is a
      strong indicator of", "in the modern AI context", or "suggests that" to justify a
      non-unknown primary. If the evidence only supports inferential language, the correct
      primary is the *_unknown category for that dimension.
   2. DIRECT-EVIDENCE-ONLY FOR ARCHITECTURE — For model_architecture, a non-unknown tag
      requires the evidence to NAME the architecture directly (e.g., "transformer",
      "Mamba", "ResNet"). "Uses generative AI", "has a chatbot", "large model", and "does
      NLP" are NOT evidence for transformer_architecture. Default to architecture_unknown
      when the architecture is not explicitly named.
   3. SOE LARGE-MODEL RULE (model_provenance) — If the only evidence for
      foundation_developer is "large model" / "大模型" / "工业大模型" / "新能源大模型" / "行
```

```
       业大模型” plus a domain noun, default to model_adapter UNLESS the evidence also
       contains pretraining-scale evidence (token count, parameter count, GPU hours, from-
       scratch dataset, pretraining corpus size).
    4. RL / IMITATION TRAINING RULE (learning_paradigm) —reinforcement_learning and
       imitation_learning require evidence that the COMPANY’S OWN training pipeline uses
       this paradigm (language like “we train”, “training pipeline”, “reward function”,
       “policy gradient”, “PPO”, “SAC”, “demonstrations dataset”). A research paper
       mentioning RL in an experiment, or a product that “would use” RL, is insufficient.
       Otherwise, use learning_unknown.
    5. FORM_FACTOR_UNKNOWN vs NOT_APPLICABLE — Use not_applicable ONLY for companies with
       POSITIVE evidence of being software-only (SaaS, API, web app, cloud). If the
       evidence suggests physical hardware but does not name the specific platform, use
       form_factor_unknown. Never use not_applicable as a fallback for missing information.
    6. UNKNOWN / CONFIRMED MUTUAL EXCLUSION — *_unknown and confirmed categories MUST NOT
       co-occur as significant tags within the same dimension. If you have confirmed
       evidence, don’t list *_unknown; if evidence is ambiguous, use only *_unknown.

## OUTPUT FORMAT
Return a JSON object with this structure:
{
 “dimensions”: {
  “model_provenance”: {
   “categories”: [
    {“id”: “category_id1”, “relevance”: 0.95},
    {“id”: “category_id2”, “relevance”: 0.6}
   ],
   “confidence”: 0.85,
   “reasoning”: “Brief explanation with evidence”
  },
  “model_architecture”: {...},
  “commercial_positioning”: {...},
  “learning_paradigm”: {...},
  “core_functionality”: {...},
  “physical_form_factor”: {...},
  “application_domain”: {...},
  “data_modality”: {...},
  “open_source_status”: {...}
 }
}

Respond with ONLY the JSON object, no additional text.
```

Per-Paper Artifact-Tagging Prompt (Gemini 2.0 Flash, Batches of 25)

These tags were issued once per technical artifact uncovered for each firm. All technical artifacts for a given firm were classified following the process listed with the results aggregated to form a final consolidated firm-level profile.

Artifact-Tagging Prompt

```
You are classifying academic papers by the AI company "{COMPANY}" into taxonomy categories.

For EACH paper below, identify the relevant categories IN THE PAPER ITSELF. Only tag
categories that the paper provides direct evidence for.

## VALID CATEGORY IDs

model_architecture (pick all that apply per paper):
 {architecture_unknown, classical_ml, convolutional_nn, diffusion_models, gan, graph_nn,
  neurosymbolic, post_transformer_architectures, recurrent_nn, transformer_architecture}

learning_paradigm (pick all that apply per paper):
 {continual_learning, imitation_learning, reinforcement_learning, self_supervised,
  supervised_learning, unsupervised, ...}

data_modality (pick all that apply — what data does the paper work with?):
 {vision, text, audio, video, 3d_spatial, tabular_structured, genomic_molecular,
  time_series, ...}

model_provenance (pick ONE primary per paper — how was the model created?):
 {foundation_developer, task_specific_developer, model_adapter, model_integrator,
  provenance_unknown}

## PAPERS
[0] "{TITLE}" ({YEAR}, {VENUE}, {CITATIONS} citations)
  Abstract: {first 300 chars of reconstructed abstract}
[1] ...

## INSTRUCTIONS
    - Use ONLY the exact category IDs listed above
    - Only tag what the paper actually describes — don't infer from the company name
    - model_architecture: tag the specific architecture mentioned. Use
      architecture_unknown if the abstract doesn't name one
    - learning_paradigm: tag the training method described. Use learning_unknown if
      not disclosed
    - data_modality: tag the data types the paper works with
    - model_provenance: "Trained from scratch" with reuse intent → foundation_developer.
      "Fine-tuned X model" → model_adapter. Narrow task-specific model trained from
      scratch → task_specific_developer. Uses third-party model without modification →
      model_integrator
    - Also classify source_type as one of: peer_reviewed, preprint, technical_report

Return ONLY a JSON array with exactly {N} entries:
[
 {
  "paper_index": 0,
  "model_architecture": ["transformer_architecture"],
  "learning_paradigm": ["self_supervised", "reinforcement_learning"],
  "data_modality": ["text"],
  "model_provenance": ["foundation_developer"],
  "source_type": "peer_reviewed"
 },
 ...
]
```

Appendix D

Phase 2 AI Developer Inclusion Criteria Prompt

Here, we provide the core LLM prompt used to implement the Phase 2 inclusion criterion determining whether a candidate organization qualifies as an AI developer. We present the prompt in full because the precise wording of inclusion and exclusion rules, the evidence hierarchy, and the mandatory classification rules are all substantively important to understanding how the dataset was constructed.

This prompt was applied to each of the 26,629 candidate organizations identified in Phase 1. For each organization, the classifier was provided with web-sourced evidence retrieved via a search-enabled LLM agent. Classifications were subsequently refined and validated through manual review as described in Chapter 2. Readers will note that the prompt uses the phrase "frontier AI research and development aimed at advancing the state of the art," which might appear to set a higher bar than the plain-English definition in Chapter 2. The two are intended to capture the same population. The prompt's "frontier" language is an operational device designed to exclude firms that merely apply, integrate, or deploy models developed by others—not to restrict inclusion to only the most cutting-edge or largest-scale laboratories. A robotics firm training its own locomotion controllers, a medical imaging firm training proprietary diagnostic models, or a startup releasing an open-weight language model all satisfy the prompt's criteria on equal footing with hyperscale foundation model developers. The determining factor in both the body definition and the prompt is whether a firm itself creates new AI models or systems, not the scale or prestige at which it does so.

Inclusion Criteria Prompt

“““““⚠️ DEFINITION CLARIFICATION — CRITICAL ⚠️
DISCARD any previous knowledge or assumptions about the term “AI developer”. You must operate EXCLUSIVELY with the definition provided in this prompt, as all actions and decisions must be guided solely by the rules, examples, and objectives established herein.

DEFINITION OF “AI DEVELOPER” (For this system):
An “AI Developer” is an organization that is actively engaged in frontier AI research and development aimed at advancing the state of the art, rather than primarily applying or enabling existing models. This specifically includes organizations that develop their own AI models from scratch, have dedicated AI research teams, and release proprietary AI technology that advances the frontier of AI. Organizations that only use, apply, integrate, invest in, or provide infrastructure for AI are NOT considered AI developers according to this system.

IS THE ORGANIZATION AN AI DEVELOPER? Determine if the organization directly develops AI technology.

DEFAULT RULE: If there is no direct evidence of frontier AI development, code as “NO”.

CLASSIFY AS “YES” ONLY IF AT LEAST ONE OF THE FOLLOWING CRITERIA IS MET:

- Frontier AI Model or System Development: The organization develops or trains AI models or systems that advance the frontier of AI, such as:
  - Foundation or frontier-scale models (language, vision, multimodal, world models, etc.)
  - Embodied AI or AI-enabled robotics control systems
  - Neuromorphic, cognition-inspired, or hybrid symbolic-neural systems
  - Novel architectures, training regimes, learning paradigms, or hardware-software co-design
  - Evidence may include publicly documented proprietary models, technical blogs, model cards, benchmarks, datasets, open-source releases, or peer-reviewed papers/preprints.
- Sustained Institutional Commitment to Frontier AI R&D: The organization demonstrates a long-term, institutional commitment to developing new AI paradigms or models, such as:
  - Dedicated AI research divisions or labs
  - Recurring technical publications, releases, or model iterations
  - Ongoing investment in frontier AI directions (e.g., foundation models, embodied intelligence, neuromorphic computing, etc.)

EXCEPTIONS TO MANDATORY RULES (Applied when an organization clearly satisfies YES criteria despite meeting exclusion criteria):

- Companies that both invest in other AI companies AND develop frontier AI technology in-house are classified as AI developers if the in-house development is well-documented.
- Companies that both integrate AI and develop frontier AI technology in-house are classified as AI developers if the in-house development is well-documented.
- Infrastructure providers that also develop frontier AI models (like Google with TensorFlow and Gemini models) should be classified as AI developers.

CLASSIFY AS “NO” WITHOUT EXCEPTION IF:

- Organization only invests in/funds other AI companies (INVESTMENT RULE)—EXCEPT when substantial in-house AI development is also documented
- Organization only applies/uses AI in business operations (USAGE RULE)
- Organization only integrates existing AI into products (INTEGRATION RULE)—EXCEPT when substantial in-house AI development is also documented
- Organization is a university without explicit AI laboratories, AI development research or AI development focused studies (UNIVERSITY RULE)
- Organization provides AI infrastructure/tools without developing frontier models (INFRASTRUCTURE RULE)—EXCEPT when substantial in-house AI development is also documented
- Organization shows uncertainty about AI development (UNCERTAINTY RULE)
- Reasoning contains speculative language (SPECULATION RULE)
- Organization’s AI activity is limited to developer tools, platforms, programming environments, simulation software, or infrastructure for others’ AI development
- Organization engages in consulting, systems integration, or customer-specific AI deployments
- Organization focuses on inference, fine-tuning, prompt engineering, RAG pipelines, or application development using third-party models
- Advanced application of existing models—even at large scale or in safety-critical domains—does not qualify unless the organization also contributes novel models, methods, or paradigms to frontier AI research

MANDATORY RULES:
INVESTMENT: Companies investing in AI companies but that are not internally involved in frontier AI development are NOT AI developers.
USAGE: Companies using AI in operations are NOT AI developers. Example: Banks using fraud detection = NO.
INTEGRATION: Companies integrating existing AI are NOT AI developers. Example: Storage companies using AI = NO.
UNIVERSITY: Universities are NO unless specific evidence of AI divisions, AI development studies, or AI laboratories exist.
INFRASTRUCTURE: Companies providing AI tools/frameworks without developing core frontier models are NOT AI developers. Example: NVIDIA (hardware) = NO, OpenAI = YES. EXCEPTION: Companies that both provide infrastructure AND develop frontier AI models in-house are classified as AI developers.
UNCERTAINTY: If reasoning includes phrases like “need further research”, “uncertain”, “not clear”, “might develop”, “possibly”, “likely”, or “extent of AI development is unclear”, then classify as NO.
SPECULATION: If reasoning contains speculative terms like “could”, “would”, “may”, “might”, “should”, “appears to”, “seems to”, “suggests”, “likely”, “potentially”, “possibly”, “reportedly”, “allegedly”, “believed to”, “thought to”, then classify as NO.

CRITICAL CLARIFICATION ON UNCERTAINTY:

- If the AI’s reasoning contains phrases like “need further research to confirm”, “uncertain about the extent”, “not clear if directly developing”, “might be involved”, “appears to be”, “likely involved”, it MUST BE CLASSIFIED AS “NO”.
- Example: “Siemens is involved in AI across various sectors... Siemens China is likely involved in AI development as a branch of Siemens” = NO (due to uncertainty)
- Example: “Need further research to confirm the extent of AI development” = NO (due to uncertainty)

```
CRITICAL CLARIFICATION ON SPECULATION:
     — If the AI's reasoning contains speculative language, it indicates insufficient
       concrete evidence
     — Only concrete, factual evidence of frontier AI development should result in
       "YES"
     — Examples of speculative language: "it seems", "it appears", "likely",
       "possibly", "might", "could", "potential", "reportedly"

EVIDENCE HIERARCHY (HIGHEST TO LOWEST):
     — Official AI model releases by the organization = HIGHEST
     — Technical documentation of proprietary AI models = HIGH
     — Named AI research teams/divisions within the organization = HIGH
     — Patents filed by the organization for AI models/algorithms = HIGH
     — Peer-reviewed publications by the organization's researchers = MEDIUM
     — Product announcements featuring proprietary AI = MEDIUM
     — Model cards, benchmarks, and datasets released by the organization = MEDIUM
     — Marketing claims about AI = LOW
     — Speculative reports = ZERO

POST-PROCESSING RULES (APPLIED AFTER INITIAL CLASSIFICATION):
     — Cases of "Unknown" should be coded as "NO" if there is no direct evidence of
       frontier AI development
     — Organizations with university indicators in name should be coded as NO unless
       specific evidence of AI divisions, AI development studies or AI laboratories
       exist.
     — Organizations with investment indicators should be coded as NO regardless if
       not internally involved in frontier AI development—EXCEPT when substantial in-
       house AI development is also documented
     — Organizations with clear usage/application indicators should be coded as NO if
       not internally involved in frontier AI development
     — Organizations with integration indicators should be coded as NO if not
       internally involved in frontier AI development—EXCEPT when substantial in-
       house AI development is also documented
     — Organizations with speculative language in reasoning should be coded as NO
       regardless of other factors
     — In cases where an organization satisfies both YES criteria and exclusion
       criteria, classify as AI_DEVELOPER = YES if the satisfaction of YES criteria
       is clear and unambiguous

OUTPUT FORMAT:
AI_Developer: YES or NO
Confidence: High, Medium, or Low
Reason: One sentence explaining classification based on direct evidence found
```

Abbreviations

3D	three-dimensional
AGI	artificial general intelligence
AI	artificial intelligence
API	application programming interface
CNN	convolutional neural network
IoT	internet of things
IQR	interquartile range
LLM	large language model
ML	machine learning
OECD	Organisation for Economic Co-operation and Development
R&D	research and development
RL	reinforcement learning
RLHF	reinforcement learning from human feedback

References

Artificial Analysis, "LLM API Providers Leaderboard—Comparison of over 500 AI Model Endpoints," webpage, undated. As of May 28, 2026: https://artificialanalysis.ai/leaderboards/providers

Chang, Wendy, Rebecca Arcesati, and Altynay Junusova, *Embodied AI: China's Ambitious Path to Transform Its Robotics Industry*, Mercator Institute for China Studies (MERICS), 2026.

Corea, Francesco, *Artificial Intelligence and Exponential Technologies: Business Models Evolution and New Investment Opportunities*, Springer, 2017.

Hannas, William C., Huey-Meei Chang, Maximilian Riesenhuber, and Daniel H. Chou, *Chinese Critiques of Large Language Models: Finding the Path to General Artificial Intelligence*, Center for Security and Emerging Technology, 2025.

Hannas, William C., Huey-Meei Chang, Valentin Weber, and Daniel H. Chou, *China's Embodied AI: A Path to AGI*, Center for Security and Emerging Technology, 2025.

Lambert, Nathan, and Florian Brand, "China's Top 19 Open Model Labs," *Interconnects*, Substack, August 17, 2025.

Long, Cheryl Xiaoning, and Jun Wang, "China's Patent Promotion Policies and Its Quality Implications," *Science and Public Policy*, Vol. 46, No. 1, February 2019.

Luo, N. A., and Ken Hyland, "International Publishing as a Networked Activity: Collegial Support for Chinese Scientists," *Applied Linguistics*, Vol. 42, No. 1, February 2021.

Marcellino, William, "Hedging Our Bets: Is the United States Concentrated on a Single AI Strategy?" *Geopolitics of AGI*, Substack, November 20, 2025.

Marcellino, William, Lav Varshney, Anton Shenk, Nicolas M. Robles, and Benjamin Boudreaux, *Charting Multiple Courses to Artificial General Intelligence*, RAND Corporation, PE-A3691-1, April 2025. As of May 28, 2026: https://www.rand.org/pubs/perspectives/PEA3691-1.html

Ministry of Industry and Information Technology, People's Republic of China, *Humanoid Robot and Embodied Intelligence Standard System*, February 28, 2026 (as reported by TechNode). As of July 30, 2026: https://technode.com/2026/02/28/china-releases-first-national-standard-framework-for-humanoid-robots-and-embodied-ai/

National Institute of Standards and Technology, *The NIST Definition of Cloud Computing*, NIST SP 800-145, September 2011.

OECD—*See* Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development, *Identifying Emerging AI Technologies Using Patent Data: A Semi-Automated Approach*, 2025.

Quan, Wei, Bikun Chen, and Fei Shu, "Publish or Impoverish: An Investigation of the Monetary Reward System of Science in China (1999–2016)," arXiv, arXiv:1707.01162, July 4, 2017.

Schmid, Jon, *An Open-Source Method for Assessing National Scientific and Technological Standing: With Applications to Artificial Intelligence and Machine Learning*, RAND Corporation, RR-A1482-3, 2021. As of July 23, 2026:
https://www.rand.org/pubs/research_reports/RRA1482-3.html

Schmid, Jon, and Fei-Ling Wang, "Beyond National Innovation Systems: Incentives and China's Innovation Performance," *Journal of Contemporary China*, Vol. 26, No. 104, March 2017.

Stanford Institute for Human-Centered Artificial Intelligence, *The AI Index 2026 Annual Report*, Stanford University, 2026.

State Council of the People's Republic of China, "Made in China 2025" ["中国制造2025"], State Council Document No. 28, May 19, 2015. English translation and analysis available via the Mercator Institute for China Studies (MERICS). As of July 30, 2026:
https://merics.org/en/report/made-china-2025

Wang, Austin Horng-En, and Kyle Siler-Evans, *U.S.-China Competition for Artificial Intelligence Markets: Analyzing Global Use Patterns of Large Language Models*, RAND Corporation, RR-A4355-1, 2026. As of July 23, 2026:
https://www.rand.org/pubs/research_reports/RRA4355-1.html

About the Authors

Jon Schmid is a senior political scientist at RAND. He specializes in the measurement and assessment of technological innovation and scientific progress, with a particular focus on emerging and military technologies. Schmid holds a Ph.D. in international affairs, science, and technology.

Prateek Puri is an information scientist at RAND. His research focuses on the threats and opportunities posed by AI systems, particularly LLMs and their impact on misinformation spread and prevention. Puri holds a Ph.D. in experimental atomic, molecular, and optical physics.

Vitor Melo is an associate economist at RAND, where he studies labor and health care markets with a focus on the economic impacts of AI. He holds a Ph.D. in economics.

Anthony Hakim is a research software engineer at RAND. He is experienced in developing and automating ML and data pipelines for research and production teams. He holds an M.S. in computational analysis and public policy.