Tracking China's Growing AI Innovative Capacity Using Publications, Patents, and People

Sagar Baviskar, Lee Branstetter, Cameron Drayton, Doug Hanley, Eduard Hovy, Lei Li, Xibao Li, Prasanna Tambe, Jiarui Wu, Liujie Wu, Huiyan Zhang

Abundant evidence exists of China's growing scientific prowess and even intellectual leadership in key fields. Yet standard bibliometric analyses of article texts can struggle to ascertain where and to what extent this exists. Chinese computer scientists working for industrial firms and academic institutions publish large numbers of papers documenting their advances, but determining the true scientific value and innovativeness of Chinese publications is complicated by well-documented biases in both Chinese scientific publications and the citations in these documents that are not unique to computer science but are certainly present in it. Strong incentives to publish lead some Chinese scientists, especially at lower-tier universities, to rush publications, cut corners, and occasionally fabricate data; 60% of science article retractions have a Chinese author (Van Norden, 2025). Chu, Steinwinder, and Azoulay (2025) document a striking "home bias" in Chinese scientists' citations of other Chinese academic publications -- a bias so strong it substantially distorts the true impact of Chinese academic science across fields.

Similar challenges obtain in the case of Chinese patents. An expanding literature documents the strong incentives in China to patent low quality inventions and the distortions this generates for the Chinese innovation system (Boeing et al, 2026; Branstetter et al, 2025; Cao et al, 2024; Chen et al., 2021; Dang and Motohashi, 2015; Li, 2012; Wei et al, 2021; Sun et al, 2021; Hu, 2014). Simply counting and comparing "AI" patents granted in the U.S. and in China is neither appropriate nor sufficient as a means of evaluating Chinese AI prowess relative to the U.S.; attempts to do so generate a number of anomalies (Fang et al, 2026). The emergence of Chinese technological leadership in important fields of engineering is undeniable, but, as with papers, standard NLP approaches to the analyses of patent texts can obscure as much as they reveal.

Despite these challenges, publications and patents reveal important truths about the advance of China's growing AI capabilities that cannot be obtained elsewhere. Properly used, these data not only permit broad international comparisons but also illustrate how and where in the technological product space Chinese institutions and firms are generating major advances. The key is to separate the signal contained in documentation of real advances in academic science (by papers) and industrial technology (by patents) from the noise generated by low-quality patents and papers. This is difficult to accomplish in practice using standard NLP tools because of the role patents and papers play in signaling compliance with and support of government policy priorities. Authors of these documents can manipulate their text to reflect the "novel" fields in which authorities wish to see greater scientific and R&D effort. Yet, the appearance of fashionable buzzwords and the apparent elaboration of novel-sounding concepts in these documents can mask a limited degree of truly substantive innovation. It represents one of the world's best manifestations of Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."

How can we separate the wheat of real AI innovation documented in patents and papers from the chaff of patent and paper texts deliberately constructed to make this separation difficult?

The approach taken in this paper will focus on author and inventor identity, and the position of authors and inventors within the Chinese (and global) networks of AI coauthorship, coinvention, and cowork. As in the United States, the current coincidence of scientific opportunity and economic value in AI generates significant flows of talent between elite universities, leading firms, and disruptive start-ups. With some well-documented exceptions, leading AI innovation in China is generally led by top executives, research directors, and/or leading engineers trained in elite university labs in China or the West. Top Chinese academic and industrial labs regularly present at top international conferences and have their work cited extensively by elite researchers outside China. While a rising tide of derivative, low-quality academic work outside these elite circles is emerging in response to government directives, our work will identify the authors, academic centers, and industrial labs with strong network positions and international impact. Similarly, we will use the approaches developed in Baviskar et al. (2025) to screen 8.5 million Chinese invention patents, identify those that are AI-related, and then screen Chinese AI patents on the basis of assignee and inventor status within the computer science network, the propensity of inventors to patent the invention outside China, citations received from non-Chinese inventions, and the propensity of inventors to pay fees to keep patents in force through their entire legal life.

Having identified high-quality AI publications and patents, our next key objective is to identify experts on fine-grained technology topics and track the evolution of this expertise over their career trajectories. Our approach includes five key steps:

1) Constructing a large-scale expert-document corpus by mining from diverse sources. Experts include both researchers who publish papers over a career that moves back and forth between academia and industry and key engineers in industry whose employers may not allow them to publish. To measure expertise across these diverse backgrounds, we extract data from a variety of sources, including papers from Elsevier's Scopus database, obtained through an existing partnership with Elsevier, Arxiv (English), the CSL dataset (Chinese papers), authors and papers from DBLP, the professor-university database from CSRanking, person profile data from OpenReview, user profiles and activities from Github (only public), user profiles from ORCID, patents from PatentsView and Chinese sources of patent data. We also crawl public websites from top universities and industry labs globally. By automatically parsing professors' homepages, we extract (partial) listings of graduate students who have worked with those professors.

2) Automatically constructing and validating (with humans in the loop) an AI topic taxonomy using topic models and pre-trained neural embeddings. We use existing multilingual named entity recognition and Bayesian topic modelling tools (e.g., a Hierarchical Topic Model) and develop neural embedding-based methods to build technical topics in the subfields of AI described in Baviskar et al. (2025). We develop a method to predict the uncertainty of topic terms and use it to guide the human validation of the topic taxonomy. A graduate student in AI efficiently validates subset(s) of highly uncertain topics.

3) Linking and disambiguating experts and associating expert-organization linkages using information extraction and LLM-based prediction. We develop an LLM-based approach to generate parsers to extract author affiliations. One critical challenge is that persons with the same name may correspond to different entities even if they are in the same organization. Another challenge is associating each person with the correct organization over the right period, when individuals are mobile across organizational boundaries. We start with a highly selective set of top AI experts and develop a semi-automatic pipeline to correct their affiliated organizations.

4) Identifying temporal evolution of fine-grained expertise for experts based on analysis of technical articles. We develop a new index to measure the expertise-level for experts based on their papers, patents, and open-source software contributions.

5) This data base of top Chinese AI experts serves as a map of AI innovative capability in China, identifying fine-grained technological areas where a critical mass of top Chinese talent exists. It also links these individuals to their employing organizations, their career histories, and their papers, patents, and publicly identified open-source software contributions.