February 2021

Using Machine Learning to Fill Gaps in Chinese AI Market Data

Supervised Learning Finds AI-Related Activity That Leading Datasets Miss

CSET Data Brief

CSET
CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Zachary Arnold
Joanne Boisson
Lorenzo Bongiovanni
Daniel Chou
Carrie Peelman
Ilya Rahkovsky

Executive Summary

Policymakers, market analysts, and academic researchers often use commercial databases to identify artificial intelligence-related companies and investments. High-quality commercial datasets have many advantages, but by design or by accident, they may overlook some AI-related companies. This proof-of-concept brief describes a new means of identifying these “missing” companies. We used machine learning (ML) models developed by Amplyfi Ltd. and Chinese-language web data to identify Chinese companies active in AI, then manually confirmed whether two leading commercial datasets, Crunchbase and PEData/Zero2IPO, included these companies and associated them with AI.

We found that most of the companies identified by Amplyfi’s models were *not* labeled or described as AI-related in these databases. Although our findings are preliminary, the sheer volume of the “hidden” companies suggests that no matter one’s definition of AI activity, using structured data alone—even from the best providers—will yield an incomplete picture of the Chinese AI landscape. ML-based approaches can complement these structured datasets, providing clearer insight into commercial AI activity in China.

Introduction

The private sector is currently the dominant force in AI, and policymakers, technologists, entrepreneurs, and academics are all eager for accurate measurements of AI-related business activity. A typical approach involves using business-oriented databases to identify AI-related companies. The leading databases generally assign descriptive tags to each company.[1] With these, measuring AI-related investment, for example, is as simple as adding up transaction values for all of the companies with the AI tag. Other strategies rely on other information from the datasets. In an earlier paper, for example, we searched two leading databases for companies whose textual descriptions included AI-related keywords, then analyzed trends in the transactions associated with those companies.[2]

These sorts of analyses are useful, but incomplete. There are many reasons why companies might not be tagged, described, or otherwise distinguished in commercial business datasets as relevant to AI:

- In some cases, AI-active companies may be missing from the datasets altogether. AI is a rapidly growing industry; companies may come and go faster than databases can be updated, especially early-stage companies. Many leading data providers deliberately focus on particular types of businesses, such as startups, companies in a particular region, or publicly traded companies, rather than trying to be comprehensive.
- Companies with many different products and activities may be less likely to receive a specific "AI tag," or to have AI referenced in their general business descriptions.
- AI is a new technology. Even if the companies now use or develop it, descriptions in commercial databases may rely on older information that does not mention AI.
- The algorithms and human annotators that assign AI tags to companies in commercial datasets may simply make mistakes.

For these reasons, analyses that focus on "AI companies," as indicated in leading structured databases, will miss some AI-related activity in the private sector.

To help fill these gaps and provide a richer picture of the AI sector as a whole, researchers can choose one of two strategies. First, they can use other types of structured data. For example, databases of scholarly publications can indicate which companies are especially active in AI research and which topics they are working on. Adding more types of structured data adds new perspectives, but each new perspective is still relatively narrow.

Second, they can use semi-structured or unstructured data, such as free text extracted from news reports or internet searches. This unfiltered approach should, in theory, identify more AI-related activity than parsing structured (i.e., pre-filtered) databases—along with considerable quantities of irrelevant information. Historically, humans have had to review the output and separate signal from noise, often making this approach expensive and impractical. But natural language processing (NLP) tools are increasingly able to interpret semi-structured and unstructured data with little human intervention. In this way, NLP tools can "unlock" unstructured data, which could enable more comprehensive measurements of AI-related business activity.

This paper describes our initial steps toward this goal. Together, CSET and Amplyfi collected unstructured, Chinese-language text from web sources related to AI and investment, then developed NLP tools to identify AI-related Chinese companies in the text. In the sections that follow, we describe our approach, then test its results against leading commercial databases.

Methodology

To find AI-related Chinese companies without relying on structured business databases, we collected a large corpus of Chinese-language documents related to technology and technology investment. We used Amplyfi's NLP models to extract the names of possible AI-related Chinese organizations from these documents. Human reviewers identified companies within the extracted data, then confirmed whether and how they were annotated in two leading business-related databases: the English-language Crunchbase dataset and the Chinese-language Zero2IPO/PEData dataset.[3]

Our analysis uses two different ML models. The first, the "Org Model," extracts the names of organizations from Chinese-language text. After evaluating alternatives, Amplyfi based the Org Model on the open source BERT framework[4] and trained it on a combination of four publicly available Chinese-language NLP training datasets. (See the Appendix for technical information on the model, alternative models considered, and the training data.)

Once trained, the Org Model was applied to a corpus of 235 thousand Chinese-language documents (the "Web Corpus"), the vast majority of which were published from 2017-2020. One hundred ten thousand came from *36Kr*, a well-known outlet for Chinese-language technology and financial news. The other 125 thousand documents came from web searches for various Chinese keywords related to AI and investment.[5] The model identified more than 700 thousand apparent organization names within this corpus. Amplyfi applied a set of post-processing rules, described in the Appendix, to clean and standardize this list of organizations and remove non-Chinese organizations.

The second model, the "AI Model," distinguishes sentences that describe an organization being involved in AI ("positive sentences") from sentences that describe some other relation (or no relation) between an organization and AI ("negative sentences"). The AI Model is deliberately simple: when parsing sentences, it does not distinguish between different ways in which organizations may be related to AI (e.g., as researchers, product developers, vendors, investors) or different degrees of AI specialization (e.g., pure-play AI startups versus diversified companies with some AI-related products).

The AI Model uses a Lattice Long Short-Term Memory (LSTM) architecture.[6] To train the model, human annotators gathered and labeled several thousand positive and negative sentences from web searches related to Chinese companies. The trained model is selective, favoring precision over recall; that is, it is designed to extract entities with a higher likelihood of being AI-related, rather than capturing all possible AI-related organizations. (See the Appendix for further detail.)

After training the AI Model on these hand-picked sentences, we ran it on every sentence in the Web Corpus that included both an organization name (as identified by the Org Model) and an AI-related keyword.[7] From these, the AI Model identified positive sentences containing about 30 thousand apparent organizational names. In other words, the Org Model and AI Model together detected about 30 thousand entities that appeared to be organizations active in AI.

From this larger group, we selected a subsample of 3,156 entities that seemed likely to be associated with meaningful AI activity.[8] Chinese-speaking annotators reviewed each of these entities, confirming whether they were in fact companies; as noted above, some of the "organizations" extracted by the AI Model were not actually organizations, or were organizations but not companies. The annotators also noted (1) whether the companies in the subsample were labeled with an AI term in Crunchbase or PEData and (2) whether the companies' descriptions in those databases included AI terms.[9] Annotators worked according to a set of task-specific instructions and examples provided by CSET.

Findings

According to our review, 888 of the 3,156 potential organizations in the AI Model subsample were actual companies. The other records were either variant names, subsidiaries, or products of companies already counted; organizations but not companies (as defined above); or were not organizations at all.[10]

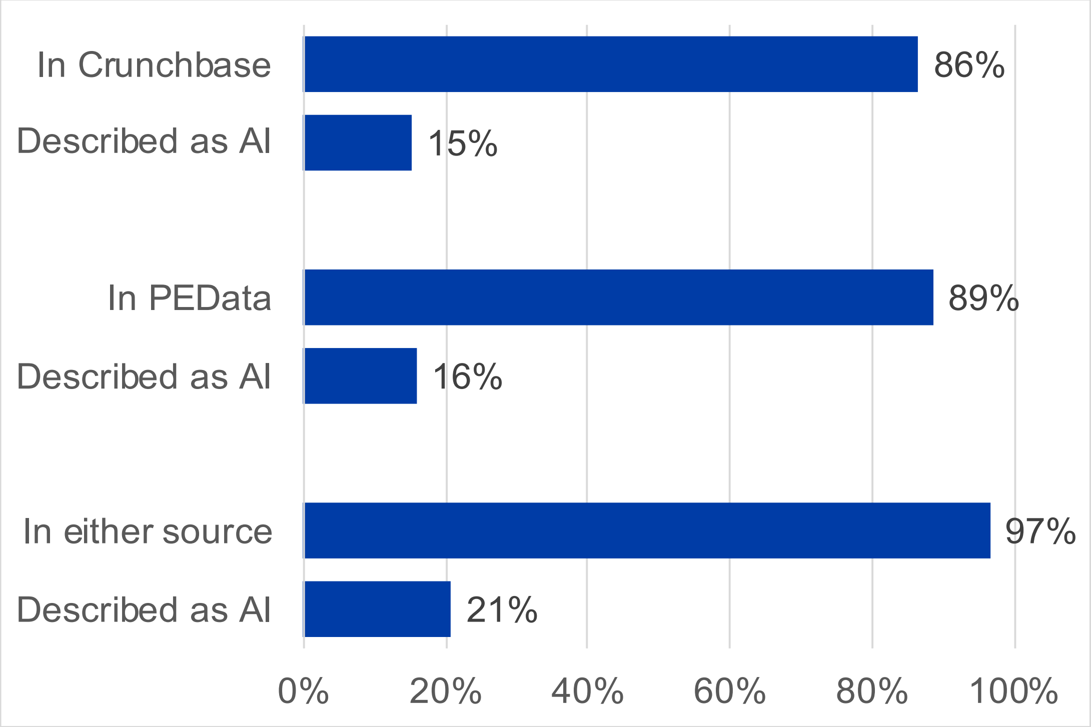
857 of the 888 companies (97 percent) were included in Crunchbase, PEData, or both, but only 184 (21 percent) were labeled or described with an AI term in those datasets. Taken separately, Crunchbase and PEData performed similarly, but with differences in the companies identified by each.

Table 1. Summary of the annotated subsample

Unique companies	888	100%
Included in Crunchbase	*768*	*86%*
Included in PEData	*786*	*89%*
Included in either source	857	97%
Labeled or described with an AI term in Crunchbase	*134*	*15%*
Labeled or described with an AI term in PEData	*140*	*16%*
Labeled or described with an AI term in either source	184	21%

Source: CSET annotation of 3,156 entities identified by the AI Model.

Figure 1. Most of the subsampled companies are present in commercial datasets, but lack AI-related labels or descriptions in those datasets



Source: CSET annotation of a subsample consisting of 888 unique companies identified by the AI Model. Read as: "86 percent of the subsampled companies were present in Crunchbase, but only 15 percent of the subsampled companies were present *and* labeled or defined with an AI term in Crunchbase."

In total, 704 companies in the subsample were neither labeled nor described with an AI term in either Crunchbase or PEData. We randomly selected 40 of these companies for manual review. Twenty-eight (70 percent) appeared to be Chinese companies with at least some meaningful AI activity; Table 2 provides a few examples. (The others were either Chinese companies with an unclear relation to AI or non-Chinese companies.)

This sampling exercise suggests our machine-learning approach identified at least hundreds, and potentially thousands, of AI-related companies that are neither labeled nor described with an AI term in either Crunchbase or PEData.[11] Research strategies relying on labels and business descriptions would overlook these companies.

Table 2. Examples of randomly selected companies from the AI Model subsample that are not labeled or described with an AI term in Crunchbase or PEData

Company	Examples of related sentences from the Web Corpus
BAIC BJEV [北汽新能源]	BAIC BJEV has integrated advanced technologies such as artificial intelligence and deep learning, and independently developed a vehicle artificial intelligence system with self-learning and self-growth capabilities—the Darwin system. [北汽新能源融合人工智能、深度学习等先进技术，自主开发了具有自学习、自成长能力的整车人工智能系统——达尔文系统。]
Sogou [搜狗]	On February 19 this year, Xinhua News Agency and Sogou released a newly upgraded standing AI synthetic news anchor, and launched the world's first AI synthetic female news anchor. This is an important achievement of the in-depth integration of artificial intelligence and news gathering and editing, and opens up new horizons in media integration. [今年2月19日，新华社联合搜狗公司发布全新升级的站立式AI合成主播，并推出全球首个AI合成女主播，是人工智能与新闻采编深度融合的重要成果，为媒体融合向纵深发展开辟了新空间。]
Dada-JD Daojia [达达-京东到家]	Dada Group (formerly "Dada-JD Daojia") went online with its "contactless delivery service." By relying on artificial intelligence technology and a crowdsourcing model, it matches up transportation capacity with frequent fluctuations in real-time delivery orders. . . [达达集团(原"达达-京东到家") 上线"无接触配送服务"， 依托人工智能技术，通过众包模式，针对即时配送中订单的频繁波动来合理匹配运力. . .]
Yunmu AI [云目未来]	As a technology company that relies on deep learning and computer vision technology and uses AI to "understand" video content, Yunmu AI has been established on the basis of a deep learning algorithm model trained on billions of images. Yunmu's core business is AI video technology. We enable intelligent processing of video and other media content for government and business. [作为一家依托深度学习与计算机视觉技术，用AI"理解"视频内容的科技公司，云目未来成立以来以亿级图像训练的深度学习算法模型为基础，以AI视频技术为核心，推动政企在视频等媒体内容领域智能处理。]
Sichuan Changhong Electric Co. [四川长虹]	"Changhong 65Q6N is equipped with Changhong's latest AI4.0 artificial intelligence, with far-field voice control and screen sleep and speaker functions. According to reports, the recognition rate of this voice control can reach 98%, which can understand user needs faster and more accurately, and can recognize some dialects. . ." [长虹65Q6N搭载长虹最新AI4.0人工智能，具备远场语音操控、息屏音箱功能，其语音操控据介绍识别率可达98%， 能更快更精准的读懂用户的需求，还能识别部分方言. . .]

Discussion and next steps

China’s AI industry is broad, deep, and rapidly growing.[12] Our findings suggest that a great deal of Chinese private-sector AI activity is hard to detect in leading commercial databases. This is unsurprising. Business-oriented databases like Crunchbase and PEData are not designed to capture all traces of AI-related activity.[13]

It also bears repeating that the AI Model takes a broad, binary view of which sentences describe a company being involved in AI. As discussed above, sentences are either positive (describing AI involvement) or negative (not describing AI involvement), and the organizations described in the sentences are deemed either “involved in AI” or not. The model does not distinguish between degrees of AI involvement, so the organizations it identifies are not all AI heavyweights. Giant companies like Alibaba and Tencent, tiny AI-focused startups, and longstanding enterprises with a few AI-related products are all grouped as “involved in AI” in our analysis. Depending on one’s analytic goals, some of these organizations might be considered less important than others.

Still, more than three quarters of the model-identified, AI-involved companies we examined are not labeled or described as AI-related in structured datasets. The sheer volume of these “hidden” companies suggests that no matter one’s definition of AI activity, using structured data alone—even from the best providers—will yield an incomplete understanding of China’s AI industry. Researchers, policymakers, and journalists should keep this in mind when using sources that rely on these structured databases.

Although this proof-of-concept analysis is focused on China, we expect that structured datasets are also highly incomplete with respect to the AI sector in the United States and other countries. In future work, we may modify our methodology and NLP models to interpret unstructured data in English, which will allow us to confirm this hypothesis. Other potential next steps include:

- *Improving the models.* As discussed above, most of the “organizations” flagged by the Org Model in the subsample reviewed by human annotators were not really organizations—and by design, the AI Model also probably overlooks many other AI-active organizations. Neither issue affects the specific conclusions drawn in this paper, but they make it difficult to extend our analysis, and the high noise-to-signal ratio in the AI Model results requires time-consuming human review. We may further refine our models and their training data, hopefully improving precision and recall.

- *Extracting other types of information*. We plan to develop models that can extract additional concepts from sentences, such as transactions and their attributes (amount, currency, series, etc.)
- *Unpacking "AI activity."* We may develop models to draw distinctions finer than "AI-involved" or not—for example, distinguishing between different types or degrees of AI activity.
- *Expanding the corpus.* We may augment the 235 thousand document Web Corpus with other raw data, which could help us find new AI-active companies and identify which are most significant.

Authors

Zachary Arnold is a research fellow with CSET. Daniel Chou and Ilya Rahkovsky are CSET data scientists. Joanne Boisson and Carrie Peelman are machine learning engineers with Amplyfi, where Lorenzo Bongiovanni is lead machine learning scientist.

Acknowledgments

Thanks to Theodoros Evgeniou, Igor Mikolic-Torreira, Dewey Murdick, and John Rigsby for useful feedback; Katerina Giannoukou and Kacper Wieczorek for contributions to the models; Alex Engel for help with data harvesting; Cong Niu for data annotation; Ben Murphy for translation support; and Ben Murphy, Xueyan Liu, Emily Xue, Lin Gan, Yao Zhang, and Zungui Lu for help annotating the AI Model subsample.

Document Identifier: doi: 10.51593/20200064

Appendix: Technical information

Org Model: training data

The Org Model training dataset comprises 99 MB of Chinese-language text compiled from four sources:

- The *People's Daily* annotated dataset (71 MB), available from Beijing University.[14] This dataset includes a tag for organizations (机构团体) appearing within the text.
- A publicly available 1 MB sample of the OntoNotes 5.0 Chinese-language dataset.[15] Text in this dataset is annotated with "structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)."[16]
- The open source Boson named entity recognition (NER) dataset (3 MB), which includes annotations for organizations, products, locations and people.[17]
- The open source Microsoft Research Asia (MSRA) simplified Chinese corpus (13 MB), with annotations for locations, organizations, and people.[18]

Org Model: model selection and testing

The Org Model uses a Chinese-language variant of the BERT NLP model to extract organizations from Chinese text.[19] Amplyfi also evaluated a Lattice LSTM-based model and an open source model using conditional random fields as recurrent neural networks (CRF-RNN).[20] Before selecting BERT, Amplyfi tested each of these three models against the MSRA corpus. The evaluation followed the CoNLL task evaluation scheme for named entity recognition. An entity was considered correctly detected if the entity boundaries assigned by the model were exactly the same as the annotated data.[21] Generally, we favored models with higher F1 scores in our testing, rather than applying a threshold requirement for either precision or recall individually.

In the evaluation, the CRF-RNN model performed worse than the other two candidates. The Lattice LSTM model performed relatively well but took several days to train, making it impractical for our analysis. Detailed evaluation results are below.

MSRA training 13MB (training) and MSRA test 14KB (test)

BERT (3 epoch)
Processed 2270 tokens with 78 phrases; found: 75; correct: 70.

	Accuracy	*Precision*	*Recall*	*F1*	*n*[*]
All	99.52%	93.33%	89.74%	91.50	
Location		100.00%	100.00%	100.00	45
Organization		88.89%	100.00%	94.12	9
Person		80.95%	68.00%	73.91	21

Lattice LSTM (5 epoch)
Processed 2270 tokens with 78 phrases; found: 72; correct: 71.

	Accuracy	*Precision*	*Recall*	*F1*	*n*
All	99.56%	98.61%	91.03%	94.67	
Location		100.00%	100.00%	100.00	45
Organization		100.00%	100.00%	100.00	8
Person		94.74%	72.00%	81.82	19

RNN + Conditional Random Fields
Processed 2294 tokens with 78 phrases; found: 70; correct: 65.

	Accuracy	*Precision*	*Recall*	*F1*	*n*
All	99.08%	92.86%	83.33%	87.84	
Location		93.33%	93.33%	93.33	45
Organization		88.89%	100.00%	94.12	9
Person		93.75%	60.00%	73.17	16

[*] The number of locations, organizations, or persons identified by the model, as applicable.

MSRA training 13MB (training) and OntoNotes sample 1MB (test)

BERT (3 epoch)
Processed 138616 tokens with 15759 phrases; found: 15832; correct: 13795.

	Accuracy	*Precision*	*Recall*	*F1*	*n*
All	94.29%	87.13%	87.54%	87.34	
Location		69.68%	89.45%	78.33	2457
Organization		61.41%	70.66%	65.71	1796
Person		94.83%	89.38%	92.03	11579

Lattice LSTM (5 epoch)
Processed 138616 tokens with 15759 phrases; found: 13298; correct: 11198.

	Accuracy	*Precision*	*Recall*	*F1*	*n*
All	88.65%	84.21%	71.06%	77.08	
Location		65.03%	87.62%	74.65	2579
Organization		62.22%	59.19%	60.67	1485
Person		93.10%	69.99%	79.91	9234

Amplyfi performed two extra checks on the BERT model:

- Evaluation against the *People's Daily* dataset (discussed below), using the CoNLL evaluation metric[22] and the character-based evaluation metric included in the Chinese-language BERT module.[23] Every character labeled with the correct entity type (location, organization, person) was considered correct. The model achieved 96.8 percent precision and 97.0 percent recall (f = 96.9%).
- Comparison to the commercial TextRazor NLP AI, using three Chinese-language documents related to AI investment. The sentences contained 51 occurrences of entities of interest. TextRazor successfully extracted 7/51 entities; BERT extracted 50/51.

Org Model: post-processing

The Org Model identified more than 700 thousand apparent organization names. Amplyfi applied a set of post-processing rules to winnow out duplicates, false positives, and companies not likely to be Chinese, including:

- Remove records containing the names of major corporations known to be non-Chinese, such as Facebook, Google, and Samsung.
- Remove records beginning with a country name other than China, Hong Kong, or Taiwan.
- Remove records if they appear in the Web Corpus in close proximity to names of countries other than China, Hong Kong, or Taiwan.
- Convert all records to simplified, lower-case Chinese characters.
- Remove extremely short and extremely long records, records including punctuation, records that typically appear in the Web Corpus, substrings of other extracted entities records, and records that are numbers, dates, or ticker symbols.
- Group all records that contain the name of the same parent organization into a single record (e.g., "Company X Interactive Inc.," "Company X Digital Inc.," and "Company X Software Inc." would all be grouped under "Company X").
- Group together variant records, including records that have the same radical Chinese characters but different suffix characters (e.g., 腾讯研究中心 (Tencent Research Center), 腾讯研究部 (Tencent Research Department), 腾讯研究院 (Tencent Research Institute) and records that include different numbers or non-Chinese characters but are otherwise similar (e.g., 腾讯 wegame and 腾讯 wegeek).
- Remove records that are substrings (truncated versions) of other records—that is, when they most often occur as parts of other records also identified by the model as organizations.
- Using n-gram modeling to identify records that include the same company name with varying prefixes and suffixes, and grouping these records together.[24]

AI Model: training data

The AI Model training dataset consists of positive sentences, defined as sentences that describe an organization being involved in AI, and negative sentences, defined as sentences that define some other relation (or no relation) between an organization and AI.

We gathered positive sentences by searching the internet for sentences that mentioned names of companies marked as AI-related in the ITJuzi startup database,[25] as well as any one of 123 AI-related keywords from a list compiled by Amplyfi.[26] This is a strict standard, favoring precision over recall; many companies may be involved in AI but are not marked as such in ITJuzi, or they may not happen to co-occur in sentences with AI keywords within the Web Corpus. Human annotators manually labeled positive and negative sentences within the search results, reviewing about two thousand sentences in total. Most were positive. Following a distant supervision approach, we then extended the positive set by adding a sample of several hundred sentences from the initial internet search, without manually annotating them.[27]

For negative sentences, we also searched the internet for sentences that mentioned an AI keyword and the name of a *non*-AI-related company, again according to ITJuzi. Human annotators manually selected negative sentences from the results of these searches. We also manually selected "false positive" sentences from the same internet search—that is, sentences that the Org Model misidentified within the results of the search as mentioning organizations, when in fact they did not mention organizations at all.

In total, the training set includes 2,816 positive sentences and 1,077 negative sentences. Each sentence is coded as a quadruple of the form *{name of organization, AI-related keyword, [positive or negative], text of sentence}*.[28]

AI Model: model selection and testing

The AI Model is based on the open source Lattice LSTM framework.[29] Amplyfi selected Lattice LSTM based on testing against a dataset of Chinese sentences compiled for a separate project on global investment transactions. These had been manually annotated to distinguish between sentences describing different types of transactions (for example, equity investments and acquisitions). After initial training, Lattice LSTM proved highly precise when labeling the sentences with their transaction types, producing results that aligned with the manual annotation over 90 percent of the time. Because the framework performed so well on this comparable relation extraction task, we were comfortable using it to distinguish positive and negative sentences related to corporate AI activity.

The Lattice LSTM model was trained for 11 epochs using default parameters, yielding an f-score of 0.95. After running the trained model against the entire Web Corpus, Amplyfi also randomly hand-checked 100 model-identified positive sentences, with the following results:

Result	Count
Chinese organizations involved in AI	64
Foreign organizations	10
Incomplete Chinese organization name	3
Probably related to AI but not precise enough to be a named entity	9
Persons, locations, products, AI-terms	10
Unrelated or very truncated terms	4

Human annotation: sampling and method

The AI Model identified about 30 thousand apparent AI-active organizations in the Web Corpus. Because the Org Model and the AI Model are not perfectly precise, we knew that not all of these organizations were actually active in AI; some were not organizations at all. We therefore selected two groups of model-identified organizations for further review by human annotators.

First, Amplyfi ran web searches for the names of the 30 thousand organizations, downloaded a sample of the results, and calculated a provisional “AI score” for every organization mentioned in five or more of the documents in the sample. The “AI score” was defined in each case as the number of documents mentioning both the organization and an AI keyword divided by the number of documents mentioning the organization. We selected for further review each organization with either an AI score above 0.85, or an AI score above 0.5 and mentions in conjunction with an AI keyword in more than 50 documents.

Second, we used provisional results from a separate Amplyfi project, currently under development, in order to identify organizations associated with investment transactions (such as venture capital rounds or mergers).

In total, 3,156 of the apparent organizations identified by the Org Model fell into one or both of these groups and were selected for further review.

These sampling criteria were extremely rough, using provisional methodologies and semi-arbitrary cutoffs. We intended only to streamline human annotators' work by picking out model outputs that were somewhat more likely to correspond to actual AI-related companies. We have no particular reason to think that our sampling methodology was biased in favor of companies that are not included or classified as AI-related in commercial databases. (If this bias existed, our findings would overstate the prevalence of these companies.) However, we did not perform any systematic validation of our method of selecting the sample, so we cannot rule out the risk of bias.

Annotation schema

Chinese-speaking reviewers labeled each record in the 3,156-company subsample with the following information:

- Whether the text in the record was actually the name of a company—that is, a for-profit business that produces goods or services for sale. As noted above, some of the "organizations" extracted by the AI Model were not actually organizations. Others were organizations, but not companies.[30]
- If the text in the record was a company name:
    - The company's website URL.
    - Whether the company had a corresponding record in the Crunchbase dataset.[31]
    - If the company had a Crunchbase record:
        - Whether the company was in the "Artificial Intelligence" Crunchbase industry category.[32]
        - Whether the company's business description included the words "artificial intelligence," "AI," or "machine learning."
    - Whether the company had a corresponding record in the PEData dataset.[33]
    - If the company had a PEData record:
        - Whether the company's PEData labels [标签] included the terms "artificial intelligence," "AI," or "machine learning," either in English or Chinese.
        - Whether the company's business description included the English terms "artificial intelligence," "AI," or "machine learning," or the Chinese terms 人工智能 [artificial intelligence] or 机器学习 [machine learning].

Endnotes

[1] See, e.g., "What Industries are Included in Crunchbase?" Crunchbase, accessed January 12, 2021, https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase-.

[2] See Zachary Arnold, Ilya Rahkovsky, and Tina Huang, "Tracking AI Investment: Initial Findings from the Private Markets" (Center for Security and Emerging Technology, September 2020), https://cset.georgetown.edu/wp-content/uploads/CSET-Tracking-AI-Investment.pdf.

[3] See crunchbase.com; pedata.cn.

[4] See Jacob Devlin, "Multilingual.md," GitHub, https://github.com/google-research/bert/blob/master/multilingual.md. BERT stands for "Bidirectional Encoder Representations from Transformers."

[5] The lists of keywords are available at Daniel Chou, "Using Machine Learning to Fill Gaps in Chinese AI Market Data – Investment Terms," GitHub, https://github.com/georgetown-cset/using-machine-learning-to-fill-gaps-in-chinese-ai-market-data/blob/main/matching_criteria/InvestmentTerms_simp_trad.txt and Daniel Chou, "Using Machine Learning to Fill Gaps in Chinese AI Market Data – AI Terms," GitHub, https://github.com/georgetown-cset/using-machine-learning-to-fill-gaps-in-chinese-ai-market-data/blob/main/matching_criteria/AITerms_simp_trad.txt. We searched Google and Baidu using the format "[AI term] AND [investment term]", trying all permutations involving one word from each list. From the search results for each query, we collected the first 200 documents that included at least one occurrence of each term in the underlying query. We then applied a validation filter to discard documents that were anomalous, for example, in terms of length, punctuation frequency, special character prevalence.

[6] See Ziran Li, Ning Ding, Zhiyuan Liu, Hai-Tao Zheng, and Ying Shen, "Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge," *Proceedings of the 57th Annual Meeting of the Association for Computational* Linguistics, July 28-August 2, 2019, 4377-4386, https://www.aclweb.org/anthology/P19-1430.pdf.

[7] The list of words is available at Chou, "Using Machine Learning to Fill Gaps in Chinese AI Market Data – AI Terms."

[8] See the Appendix for further details. The annotated subsample is available at Daniel Chou, "Using Machine Learning to Fill Gaps in Chinese AI Market Data – SQL," GitHub, https://github.com/georgetown-cset/using-machine-learning-to-fill-gaps-in-chinese-ai-market-data/tree/main/sql.

[9] See the Appendix for further details.

[10] For example, seven different records in the sample correspond to Ping An Insurance and its sub-units – "平安" [Ping An], "中国平安" [Ping An of China], "平安科技," [Ping An Technology] "平安科技（深圳）有限公司" [Ping An Technology (Shenzhen) Co., Ltd.], "

平安医保" [Ping An Medical Insurance], "平安医保科技" [Ping An Medical Technology], " 平安集团" [Ping An Group].

[11] A reasonable estimate of the lower bound might assume, very conservatively, that there are zero AI-related companies outside the 888 companies identified in the AI Model subsample. As noted, 704 of these companies were not labeled nor described with an AI term in either Crunchbase or PEData. Assuming the 79 percent "true positive" rate observed in the 40-company random sample applies generally, there are about 550 companies in the AI Model subsample that are AI-related, but neither labeled nor described with an AI term in either Crunchbase or PEData. The upper bound is harder to estimate, since the AI Model subsample is not necessarily representative of all ~30 thousand entities extracted by our model, but it seems plausible that there could be at least two times as many AI-related companies (i.e., more than a thousand such companies) in this larger group.

[12] See generally "Intro: China's AI Dream," Macro Polo, accessed January 13, 2021, https://macropolo.org/digital-projects/chinai/chinai-intro/; "China's Artificial Intelligence: Next 10 Years" (China Money Network, 2019), https://assets.chinamoneynetwork.com/wp-content/uploads/20190702121154/Next-10-Years-China-Artificial-Intelligence-China-Money-Network.pdf.

[13] It *is* somewhat surprising that Crunchbase, an English-language database, performed more or less the same as PEData, a well-regarded Chinese-language database, in this analysis of Chinese AI activity.

[14] Shiwen Yu, Huiming Duan, and Yunfang Wu, "Corpus of Multi-level Processing for Modern Chinese," Peking University Open Research Data Platform, 2018, https://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/SEYRX5.

[15] NekoApocalypse, "OntoNotes 5.0 Chinese Processing," GitHub, https://github.com/NekoApocalypse/OntoNotes-5.0-Chinese-processing; see Ralph Weischedel et al, "OntoNotes Release 5.0," Linguistic Data Consortium, October 16, 2013, https://catalog.ldc.upenn.edu/LDC2013T19.

[16] Weischedel et al, "OntoNotes Release 5.0."

[17] CClauss, "ChineseNLPCorpus – Boson," GitHub, https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/boson.

[18] InsaneLife, "ChineseNLPCorpus – MSRA," GitHub, https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/MSRA.

[19] ProHiryu, "BERT–Chinese–NER," GitHub, https://github.com/ProHiryu/bert-chinese-ner. BERT vectors for Mandarin are available at Devlin, "Multilingual.md," GitHub, https://github.com/google-research/bert/blob/master/multilingual.md.

[20] Yue Zhang and Jie Yang, "Chinese NER Using Lattice LSTM," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, July 15-20th, 2018, 1554-1564, https://www.aclweb.org/anthology/P18-1144.pdf; Jiesutd,

"LatticeLSTM," GitHub, https://github.com/jiesutd/LatticeLSTM; Zjy-ucas, "ChineseNER," GitHub, https://github.com/zjy-ucas/ChineseNER.

[21] Erik F. Tjong Kim Sang and Fien De Meulder, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," *Proceedings of the Seventh Conference on Natural Language Learning*, 2003, 142-147, https://www.aclweb.org/anthology/W03-0419.

[22] Tjong Kim Sang and De Meulder, "Introduction to the CoNLL-2003 Shared Task."

[23] ProHiryu, "BERT–Chinese–NER – tf_metrics.py," GitHub, https://github.com/ProHiryu/bert-chinese-ner/blob/master/tf_metrics.py. Specifically, we used this metric when comparing different training or parameter sets with the Chinese BERT model's. We used ConLL to compare the performance of the Chinese BERT model with other models.

[24] See generally Daniel Jurafsky and James H. Martin, "N-gram Language Models," *Speech and Language Processing*, draft of December 30, 2020, http://web.stanford.edu/~jurafsky/slp3/3.pdf.

[25] Itjuzi.com. We used the Bing API for web searching. See "Bing Search API Documentation," Microsoft, https://docs.microsoft.com/en-us/azure/cognitive-services/bing-web-search/. The Web Corpus was not ready when the AI Model was trained, so we were not able to use it for training.

[26] The list is available at Chou, "Using Machine Learning to Fill Gaps," https://github.com/georgetown-cset/using-machine-learning-to-fill-gaps-in-chinese-ai-market-data/blob/main/matching_criteria/AITerms_simp_trad.txt.

[27] Because human annotation confirmed that the search returned mostly positive results, we knew that the additional sampled sentences would mostly be positive. We estimate that 40 to 100 false positives were introduced at this stage. On distant supervision generally, see Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky, "Distant Supervision for Relation Extraction without Labeled Data," *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, August 2-7, 2009, 1003-1011, https://www.aclweb.org/anthology/P09-1113.pdf; Peng Su, Gang Li, Cathy Wu, and K. Vijay-Shanker, "Using Distant Supervision to Augment Manually Annotated Data for Relation Extraction," *PLoS One*, July 30, 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6667146/.

[28] For example: *{ 科大讯飞, 语音识别，true, "无论是科大讯飞、三角兽这样致力于语音识别等技术的科技公司，还是致力于人脸识别、图像识别、视频分析、无人驾驶等技术的公司等等，如雨后春笋出现在大家面前。" }*

[29] Li et al, "Chinese Relation Extraction with Multi-Grained Information."

[30] By way of illustration, annotators were specifically instructed not to label the following types of text as company names: abstract concepts; specific technologies, products, or apps; investment firms (for-profit organizations that make money by investing, rather than producing goods or services for sale); research institutions, including universities and organizations within universities; departments, labs, and subunits within companies; government agencies, departments, and offices; and publications. If annotators could not determine whether a given record was a company name or not after running web searches, the record was marked as "unclear" and excluded from further analysis.

[31] Annotators accessed Crunchbase and PEData through their public web interfaces.

[32] "What Industries are Included in Crunchbase?"

[33] Annotators accessed Crunchbase and PEData through their public web interfaces.