鹏城云脑开源生态与数据共享安全

The Open-Source Ecosystem and Data-Sharing Security System of the Peng Cheng Cloud Brain

高文 Wen GAO

鹏城实验室 Peng Cheng Laboratory

北京大学 Peking University

What we are doing @ AI for Good

- 为AI社区提供足够的算力、数据、模型
  Provide computing power, data, and model for AI community
  - 智能算力必须足够强，足以支撑大模型训练
    Enough computing power for pre-training on Large-scale Model
  - 数据必须足够大（多），但要保证数据安全
    Enough data for good model training, with data protection
  - 模型必须开源，支撑AI生态发展
    Model should be in open-source, for AI Ecosystem

提纲 Outline

- PCB-II大算力与大模型
  PCB-II Big Computing Power and Large-scale Model
- 数据安全与DPI
  DPI – Data Security
- 开源开放支撑AI生态建设
  AI Ecosystem Supported by Open-source Methodology
- 总结
  Summary

提纲 Outline

- PCB-II大算力与大模型
  PCB-II Big Computing Power and Large-scale Model
- 数据安全与DPI
  DPI – Data Security
- 开源开放支撑AI生态建设
  AI Ecosystem Supported by Open-source Methodology
- 总结
  Summary

鹏城云脑II：走向智能必不可少的基础设施

Peng-Cheng Cloud Brain (PCB-II): A Must-have Infrastructure

“十四五”国家发改委
重大科技基础设施

新一代人工智能
开放创新平台

响应
国家战略

National
Strategy

服务
本地需求

Local
Demand

支撑深圳AI
重大应用需求

提升大湾区AI研究
基础地位与创新力

理论研究
Theory

- 类脑计算
- 深度学习理论方法

核心技术
Technology

- 大规模异构集群管理
- 新型计算架构与系统

智能应用
Application

- **超大规模AI模型**
- 靶向药物研究

鹏城云脑II：面向AI的专用架构

PCB-II: A Special Designed Architecture for AI

低延时：2 us
Low latency

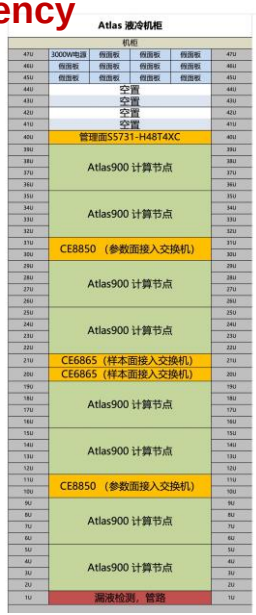
强算力: 1E ops
Big computing power

高存储: 64PB
Huge storage

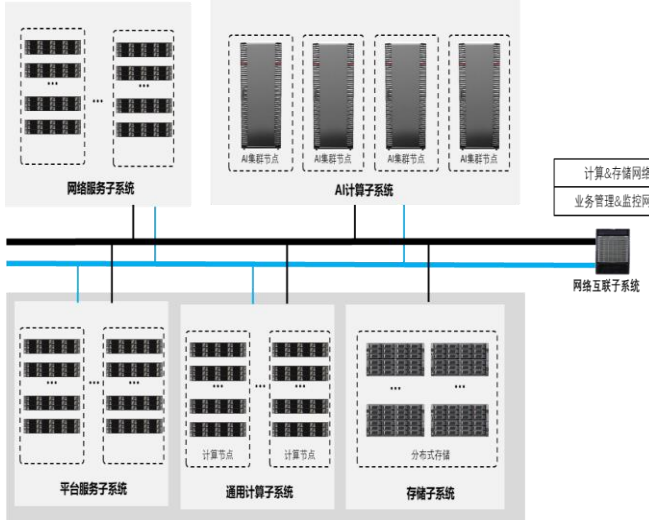
核心芯片

Ascend 910

- 半精度: 256 TFLOPS
  性能为V100二倍
- 整数精度: 512 TOPS
- 工艺 7nm FFC



Atlas900服务器
Atlas900 Server



近千台设备集群
Cluster with a thousand machines

4096颗AI处理器
4096 AI processors

2048颗CPU处理器
2048 CPUs

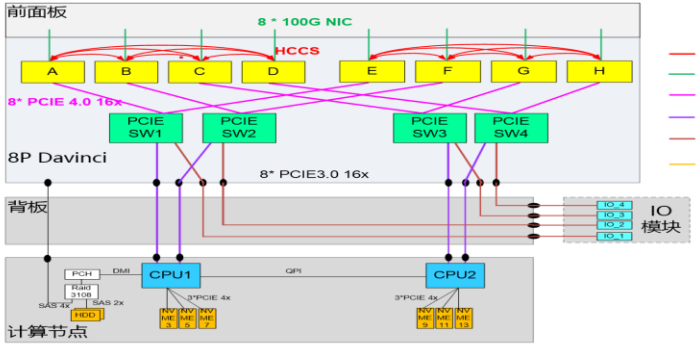
采用特别设计的芯片构建大规模AI算力平台

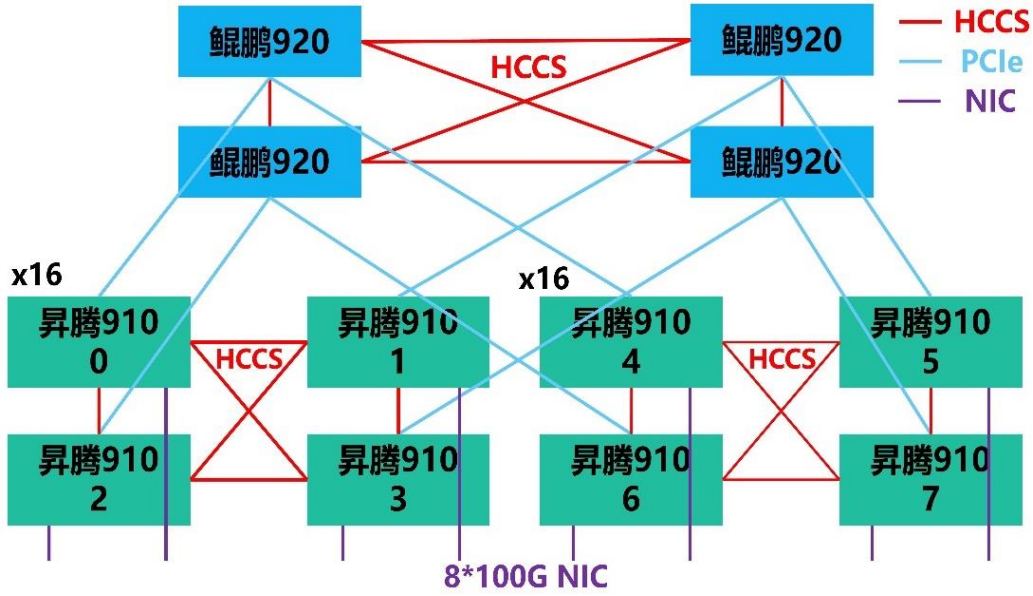
The large-scale AI computing platform is built using self-designed chips

系统设计：节点内CPU和NPU平衡设计

System Design: Balance of CPUs and NPUs in One Computing Node

节点内芯片拓扑增强：
两路x86处理器→四路鲲鹏处理器
Enhanced chip topology in one node: two x86 processors -> four Kunpeng processors





优化集群结点内CPU与NPU配比，显著提升平台的平衡计算性能

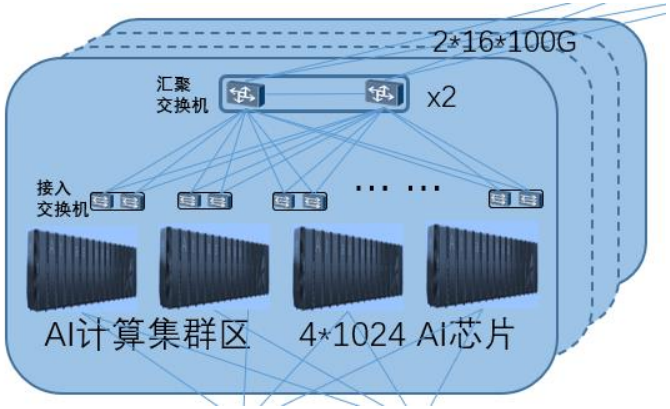
The computing power of the platform is significantly improved via optimizing the ratio of CPUs and NPUs in the cluster

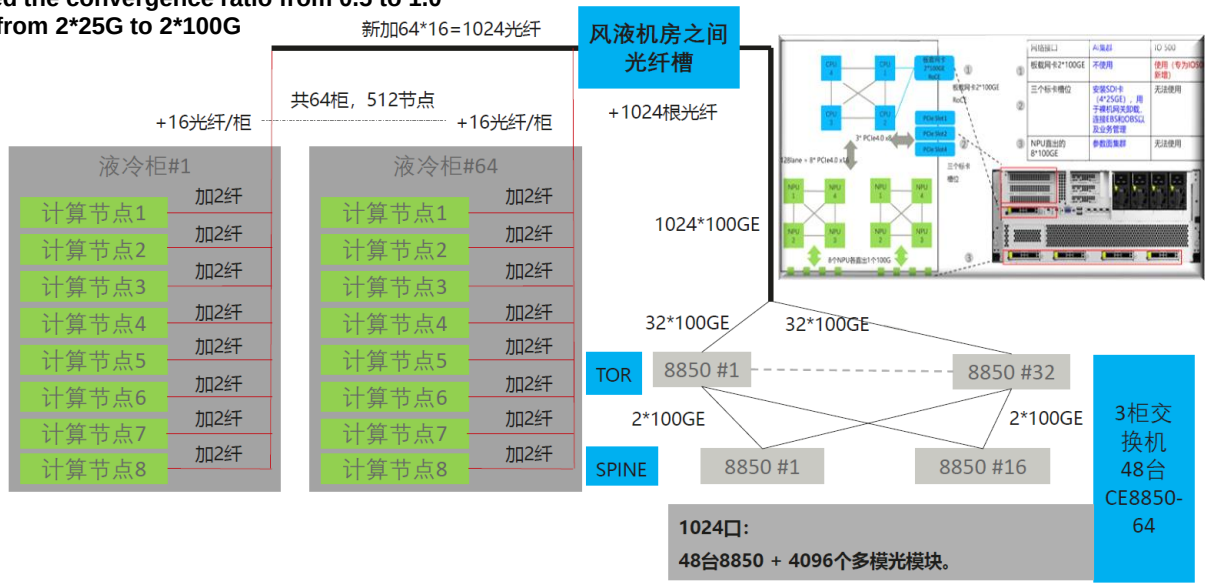
系统设计：集群间网络全联通设计并提速数据面带宽

System Design: Inter-cluster Network Fully Connection and Data Plane Bandwidth Acceleration

集群间网络互联增强：从0.5的收敛比提升至1.0
数据面网络增强：从2*25G提升至2*100G

Enhanced inter-cluster network connection：Increased the convergence ratio from 0.5 to 1.0
Enhanced data plane network bandwidth：Increased from 2*25G to 2*100G





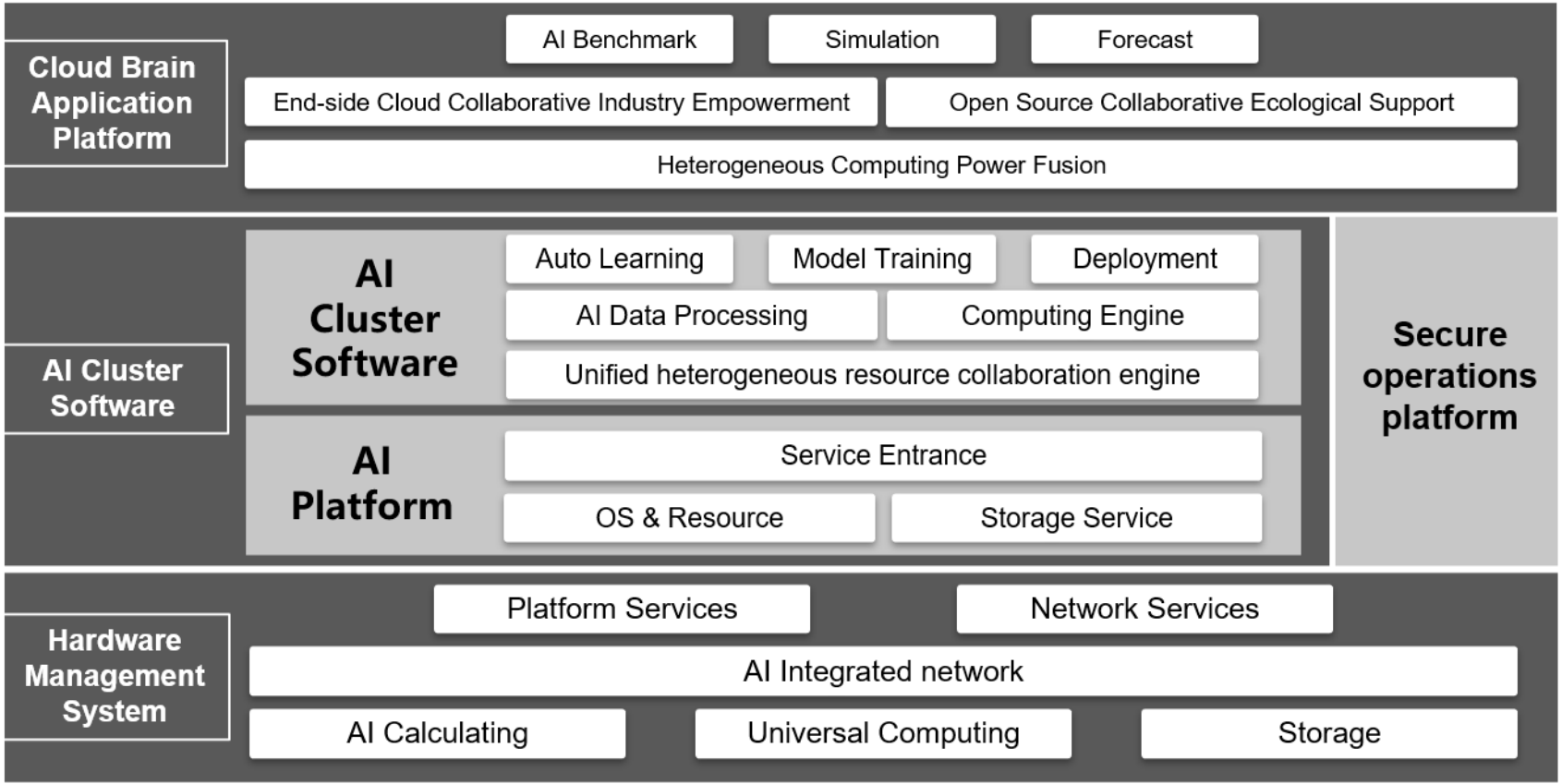
- ✓ 增加48台8850-64交换机
- ✓ 增加4096个多模光模块
- ✓ 增加支持512节点全机互联

增强和优化集群网络结构，显著提升平台的IO吞吐和规模计算性能

The IO throughput and computing power of the cluster is significantly improved via the network structure enhancement and optimization

基础软件架构

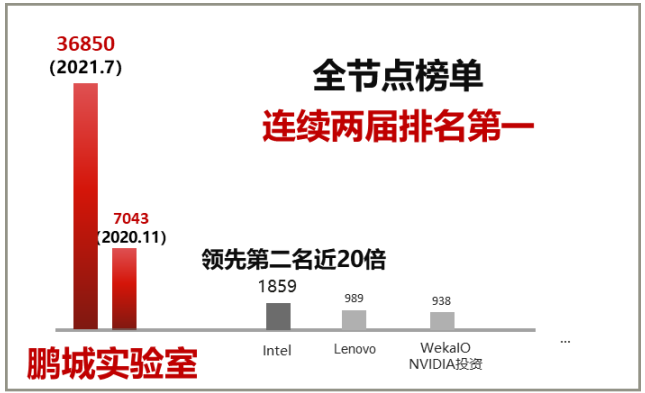
Basic Software Architecture

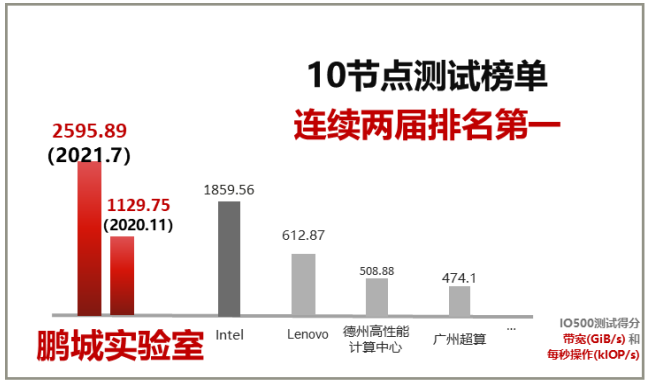


鹏城云脑II：吞吐性能与AI计算能力国际领先

PCB-II: Internationally Leading IO Throughput and AI Computing Capabilities

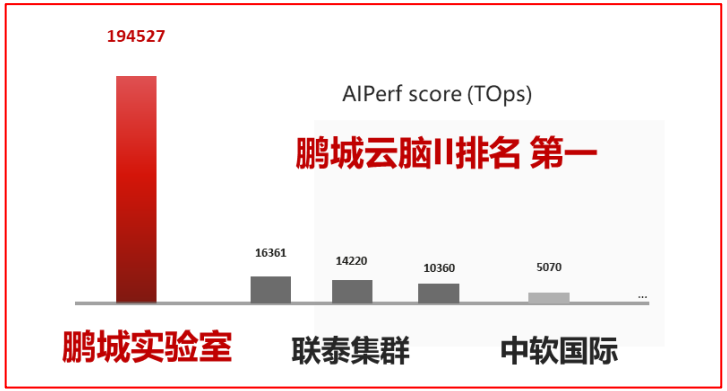
IO500 vl4lo





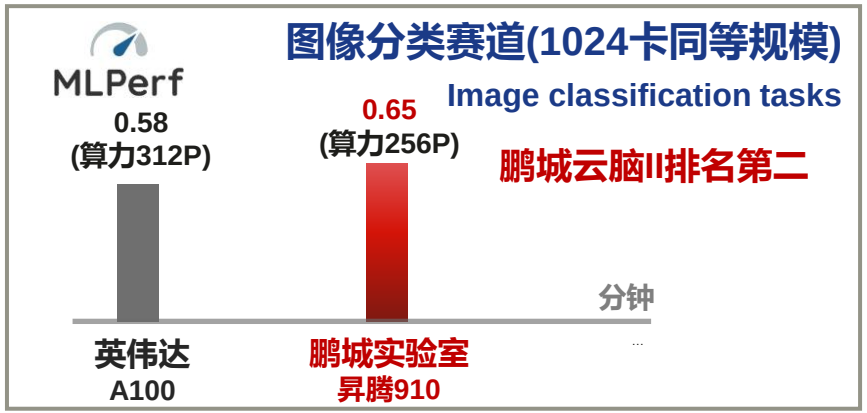
高性能计算存储系统性能排行榜 Storage Benchmark for

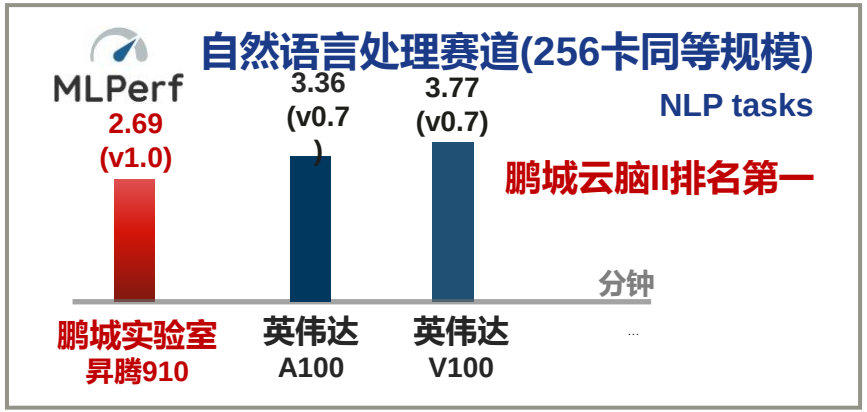
世界人工智能算力500排行榜 The world's top 500 AI computing power

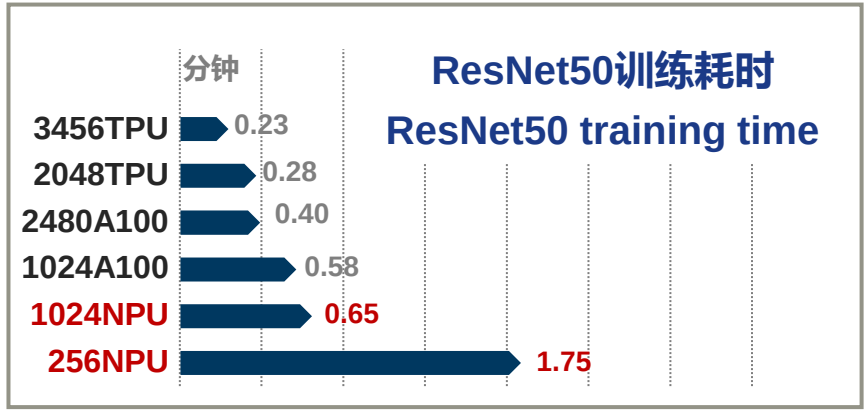


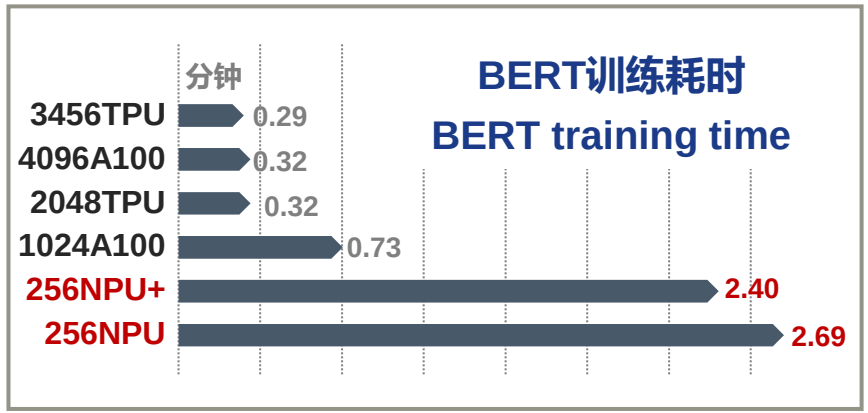
并行训练能力评测： MLPerf性能国际先进

Parallel Training Benchmark: Internationally Advanced MLPerf Performance









2021.6结果 Results on 2021.6

鹏城系列大模型

Peng Cheng Large Model Series: 5 Models

云脑构建了一个超大型NLP训练平台，拥有超过1000亿参数和4K芯片

Cloud Brain builds an ultra-large-scale model NLP training platform with more than 100 billion parameter and 4K chips

鹏程·盘古：训练千亿级开源中文大模型

Peng Cheng Pangu: Training Hundred-Billion Open-Source Large-Scale Language Model

鹏程·大圣
视觉大模型Visual Class

鹏程·通图
图网络大模型Graph Network

鹏程·通言
多语言大模型Multi-language

鹏程·常羲
多模态大模型Multi-model

高质量的语料数据集 High-quality language material data set

通过清洗过滤、加权、质量评估等处理程序处理80TB数据。构建了一组约 1.1TB 的数据，以确保数据无偏差

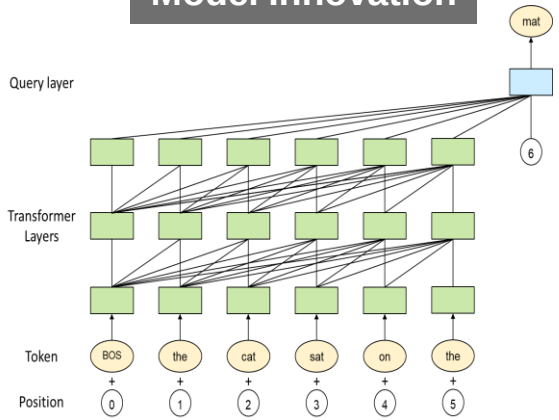
Through cleaning filtering, weighting, quality assessment and other processing procedure of 80TB of data. A data set of about 1.1TB is constructed to ensure that the data is non-biased

+

鹏程·盘古：云脑 II 训练千亿级开源中文大模型

PANGU: Training Hundred-Billion Language

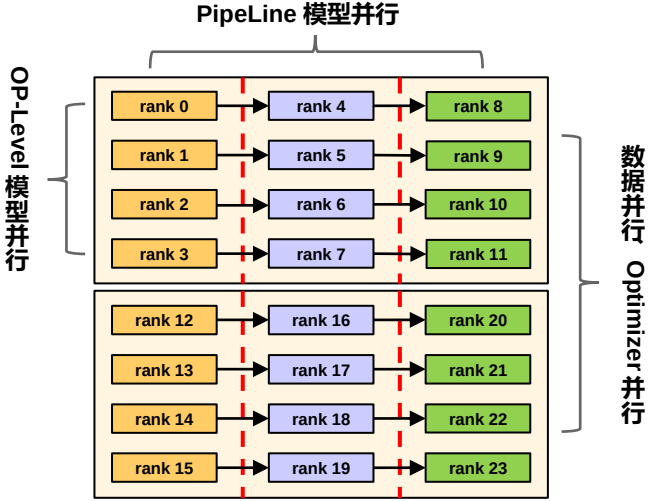
模型创新 Model Innovation



首创随机顺序自回归预训练语言模型ALM

The first random sequential autoregressive pre-training language model ALM

工程创新 Implementation Innovation



全并行方案 Fully Parallel Scheme

开源应用 Open-Source Application

模型	数据开源	算法开源	模型开放	服务开放
英文GPT-3 175B	X	X	X	√
开源清源 CPM 2.6B	X	X	√	X
鹏程•盘古 200B+13B	√	√	√	√

首次实现代码、模型、语料的全开源

Fully open source code, model, and corpus for the first time

训练了全球首个2千亿参数中文预训练语言模型，首次实现了全开源

Train the first two-hundred-billion-scale Chinese pre-training model, and completely open-source

鹏程.盘古模型的性能与开源

Model Performance and Open Source

CPM vs 鹏程·盘古 2.6B 模型对比

CPM（Zhiyuan） vs Pangu 2.6B Comparison

16个下游任务：
零样本学习11个任务领先
单样本学习12个任务超越
小样本学习13个任务超越

16 down-stream tasks:
Zero-shot learning 11 models in the lead
One-shot learning 12 models surpass
Few-shot learning 13 models surpass

各模型在各任务上的性能 CPM2.6B vs 鹏程•盘古 2.6B(7w)

			zero shot		one shot		few shot	
序号	任务	CPM论文	CPM2.6B	鹏程•盘古	CPM2.6B	鹏程•盘古	CPM2.6B	鹏程•盘古
1	CMRC2018	zs: 1.31/13.37	0.590/10.115	**1.212/16.647**	1.709/11.293	**2.485/18.566**	3.107/14.637	**5.684/23.219**
2	DRCD		0/4.618	**0.8/9.990**	0.218/5.171	**2.473/12.483**	0.145/7.143	**5.309/18.29**
3	Dureader		16.63	**21.07**	16.42	**20.18**	17.85	**21.43**
4	CHID	zs: 68.5	68.62	**68.73**	67.91	**68.16**	**66.82**	66.56
5	PD&CFT		35.73 & 38.99	**38.47 & 42.39**	33.3 & 39.73	**38.8 & 41.61**	32.03 & 39.84	**39.07 & 42.05**
6	CMRC2017		24.6	**37.83**	25.4	**38**	23.5	**36.33**
7	CMRC2019		47.69	**61.93**	47.99	**61.54**	47.2	**62.42**
8	CMNLI		49.1	**50.2**	47.56	**49.54**	49.29	**51.17**
9	OCNLI	zs: 44.2	**44.199**	42.61	**44.3**	44	44	**46.78**
10	TNEWS	zs: 70.3	**65.44**	60.95	**69.5**	57.95	**70.17**	63.62
11	IFLYTEK	zs: 70.8	68.91	**74.26**	**79.84**	79.03	**83.99**	80.15
12	AFQMC		**66.335**	59.291	39.7	**64.62**	38.29	**68.999**
13	CSL		**52.3**	50.5	**51.2**	50.9	50.5	**52**
14	WebQA.v1.0		2.48/10.254	**4.431/13.713**	8.399/17.724	**10.218/20.563**	13.095/22.450	**23.710/33.814**
15	CLUEWSC2020		**73.684**	73.355	73.684	**75.328**	70.065	**72.697**
16	C3		49.81	**53.42**	51.43	**52.82**	51.6	**53.64**
	最优计数		**5**	**11**	**4**	**12**	**3**	**13**

鹏程.盘古首次实现代码、模型、语料的全开源

Peng Cheng Pangu firstly and fully releases the code, model and corpus

基于鹏程·盘古的多语言机器翻译

Multi-language Machine Translation

1. 多语种语料语料自动生成 Auto-generation of multiple corpora

训练语料：包括五种语言的340GB多语种语料（中文300G，乌尔都语、阿拉伯语、泰语、印地语各10G），训练盘古α-13B模型

Training corpus: 340GB multi-language corpora for Pangu-α 13B model

阿拉伯语料自动生成样例

Auto-generation of Arabic corpus

input text is:

ماسي في أجزاء، مثل التحكم في المحرك والجسم وإدارة البطارية والتحكم في الترفيه داخل السيارة. ومع ذلك
مع زيادة درجة الذكاء ، يزداد الطلب على الرقائق في السيارات.

output text is:

فإن مجموعة متنوعة من الاشخاص ذوي الخبرة على الحماية من مشكلات كافية جداً . وقال عبد الرحمن جيلاني .
مر مع طبع أقصى حجم الطلاء لجميع الديكورات ، ولكن وكثير من التطبيقات جعل تلك الآليات مهتمة بعملية
ريقة عمل النماذج الخطيرة في صناديقها ، لذلك فمن الممكن أن يتوفر هذا الحد من ذلك للحصول على ترخيص
! ولكن في الواقع ، لديها جديدة أسرع
شاهد المزيد عن هذا المنتج
اكسسوارات / حقائب / رفوف في الهواء الطلق
حقائب / رفوف في الهواء الطلق
اكسسوارات / حقائب / رفوف في الهواء الطلق
تفاصيل التعبئة والتغليف : أداة التنورة / حزمة الرصاص

泰国语料自动生成案例

Auto-generation of Thai language corpus

input text is:

การนอนหลับสนิทเป็นส่วนหนึ่งของการนอนหลับหรือที่เรียกว่า "การนอนหลับสีทอง" ซึ่งมีสัดส่วนเพียง 25% ของเวลานอนทั้งหมด การนอนหลับสนิทสามารถบรรเทาความเหนื่อยล้าในแต่ละวันของผู้

output text is:

มากยิ่งขึ้นตามการนอนหลับอันควรในการช้อมการนอน ทำให้การนอนหลับที่ดีขึ้นกับเวลานอนที่ใกล้กับผู้คนที่มีความตื่นเต้นมากขึ้น
ผู้โดยสารเริ่มทำการฆ่าเชื้อโรคเพื่อพบอัตราความเสี่ยงอันตราย ได้แก่ ความฉลาด น้อยใจสนใจ และคลังกับอารมณ์เพื่อ
1 . 10 แคลอรี่ / ชั่วโมงสำหรับผู้โดยสาร
1 . 10 แคลอรี่ / ชั่วโมงสำหรับผู้โดยสาร
ปกติแล้วห้องแล็บจะจบผู้โดยสารสูงสุด และเป็นห้องแล็บขนาดเล็กที่สำคัญสำหรับกิจกรรมโดยเฉพาะกับการนอนหลับ ทั้งสองสำคัญคือการผลิตที่วีกันหน่อยด้วยการนอนหลับเพียงอย่างเดียวเท่านั้น
คุณจะต้องเลือกผู้ที่มีความหวังดีและคู่รักที่สามารถไปร่วมงานได้ รวมทั้งต้องเลือกพนักงานที่มีประสิทธิภาพที่คุณจะต้องมีการพักผ่อนในหนึ่งสุด
มี 2 ชั้นในเครือ Harrison
ห่มกลางความตื่นเต้นของห้องพักและกลางทุ่งครุ

2. 平行语料自动抽取+小样本学习机器翻译 Auto-extraction of parallel corpus + few-shot learning based machine translation

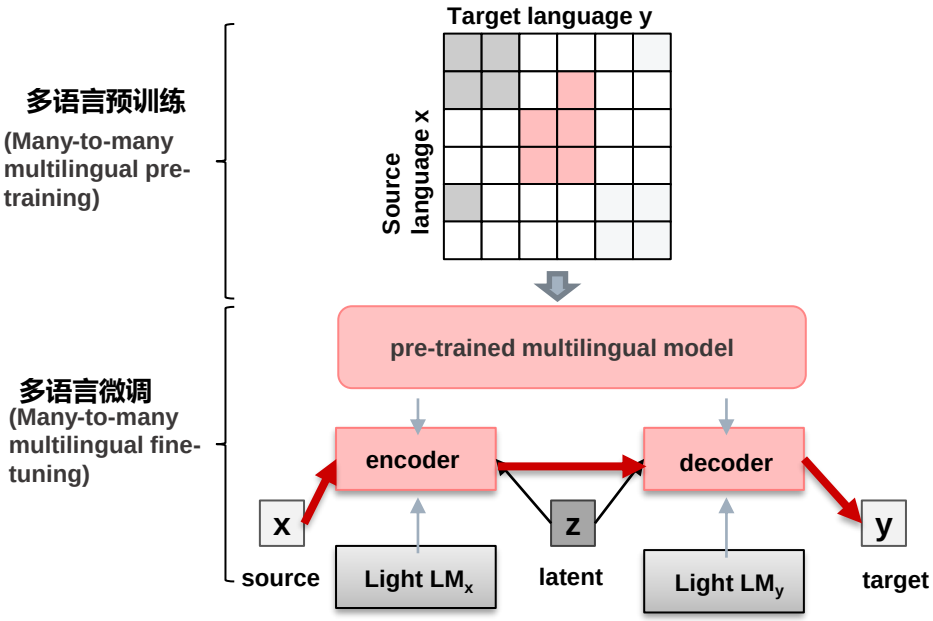
阿拉伯语	中文	谷歌翻译	相似度
انقر هنا لإلغاء الرد.	点击这里取消回复。	点击这里取消回复。	0.950
2018كأس العالم للأندية	2018世界杯	2018 FIFA Club世界杯	0.947
المؤتمر السنوي للمجمع لعام 2019م	开封测控2019年公司年会	2019年综合体年会	0.820
موقع جريدة الصباح 2019	2019今晚开奖现场直播开奖记录一大全	沙巴报纸网站2019	0.841
يرجى تسجيل الدخول للتعليق.	请登录后评论。	请登录后发表评论。	0.890
زمان برگزاری آزمون ها	(二)考试时间	考试时间	0.821
تغییرات شرکت های تجاری	企业类型变更	业务变化	0.859
نقل و انتقال سهام شرکت	公司股权转让	公司股份转让	0.884
مدل جواهرات و زیور آلات	珠宝首饰设计	珠宝和饰品模型	0.800
الأنشطة الثقافية والفنية	艺术、运动类活动	艺术、运动类活动	0.821
ورزش های مناسب کودکان	适合孩子的运动项目	适合儿童	0.895
تكلفة الشحن والتوصيل	运输时间和运输费用	运费	0.837

	Platform	Arabic
Chinese => X	Baidu	14.6
	Pangu-α	**20.4**
X => Chinese	Baidu	34.3
	Pangu-α	**43.7**

鹏程·通言：基于多对多模式的模型

TONGYAN: Many-to-many Multilingual

训练步骤 Train steps | **模型机构 Model Architecture** | **模型架构 Model Improvement**



单双语多任务结合：结合轻量级盘古α模型和单语语料扩充单语语义表示

Single-language multi-task combination: Combining lightweight Pangu α model and monolingual corpus to expand monolingual semantic representation

$$E_{MS} = \log P(y \mid x; \theta_{xy}) + \ell(T) \log P(x \mid x; \theta_x)$$

隐层语义空间扩充：建模跨语言共享语义空间

Expansion of Hidden Semantic Space: Modeling Cross-language Sharing Semantic Space

$$\log P(y \mid x) = \log P(y \mid x; \theta_x) + \xi \log P(y \mid x, z; \theta_{xy})$$

对称学习：通过度量散度，缩小语言间语义鸿沟

Symmetrical learning: narrow the semantic gap between languages by measuring divergence

$$E_{sym} = \log P(y \mid x; \theta_{xy}) + \log P(x \mid y; \theta_{yx}) + \gamma \mathrm{JS}[q(z, x \mid y; \phi_{yx}) \| q(z, y \mid x; \phi_{xy})]$$

对比学习：最小化相似语句间距离，最大化无关语句间距离

Contrastive learning: minimize the distance between similar sentences and maximize the distance between irrelevant sentences

$$E_{ctr} = \sum_{x_i, x_j \in D} \log \frac{e^{sim^+}[R(x^i), R(x^j)] / \tau}{\sum_{y^j} e^{sim^-}[R(x^i), R(y^j)] / \tau}$$

鹏程·大圣：基于视觉与跨模态的模型

DASHENG: Vision and cross-modality

参数量
Parameter quantity

已有：千万级预训练模型
Ten million-level pre-training model

十亿级预训练模型
Billion pre-training model

50亿级预训练模型
5 billion pre-training model

数据量最终达到200亿
The amount of data eventually reached 20 billion

开源数据集
Open source dataset
Object365 ImageNet

隐私数据集
Privacy dataset
交通场景 Traffic
治安场景 security

多模态数据集
Multimodal data set
文本/语音 Text/voice
红外/无线 Infrared/wireless

数据收集
data collection

通用特征提取模型
Feature Extraction

特征图
Feature map

预训练模型
Pre-training model

预训练大模型
Pre-training large model
X-万级 → X-亿级
X- 10,000 → X- 100,000,000

鹏城云脑II

应用场景
Application scenario

蒸馏 重训练
Distillation Heavy training

智能交通
Intelligent transportation

未来社区
Future community

Features/results/video frames 特征/结果/视频帧

提纲 Outline

- PCB-II大算力与大模型
  PCB-II Big Computing Power and Large-scale Model
- 数据安全与DPI
  DPI – Data Security
- 开源开放支撑AI生态建设
  AI Ecosystem Supported by Open-source Methodology
- 总结
  Summary

如何平衡隐私保护与数据挖掘的冲突?

Date Sharing and Mining Security

如何平衡
Balance

数据安全与隐私保护
Data security and privacy protection

数据价值挖掘
Data value mining

大数据场景下隐私保护与数据挖掘的矛盾问题，即如何在保护数据隐私的前提下，最大限度地挖掘大数据价值

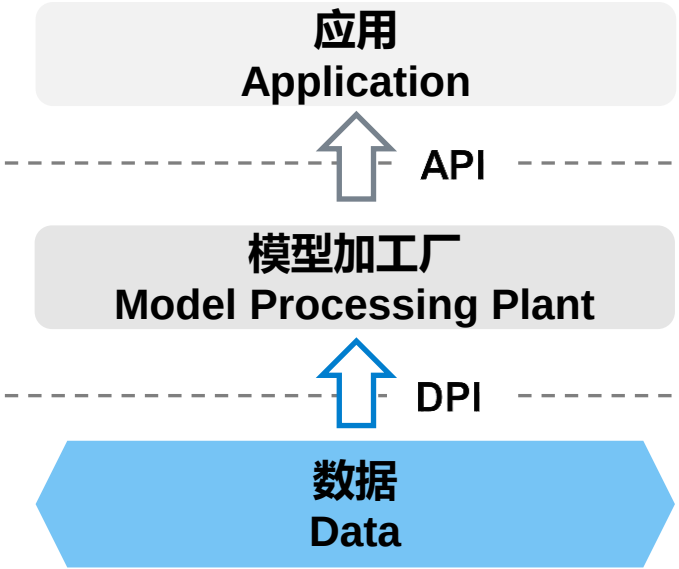
The contradiction between privacy protection and data mining in big data scenarios is to maximize the value of big data under the premise of protecting data privacy

DPI 数据程序接口

Data Programming Interface (DPI)

核心思路：数据不动程序动

Core Idea：Data does not Move, Model Moves



数据需求方 Data demander:

- 外部程序使用数据达到构建模型的目的
  External programs use data to perform model processing.
- **人员不能进入模型加工厂**查看调阅数据
  Personnel cannot enter the plant to access data.

网络靶场技术 Network shooting range technology:

- 外部程序可以在可信计算平台上运行
  external programs can run on the trusted platform.
- 隐私数据可以裸数据的形式放在该平台中
  Private data can be placed in the platform as bare data

数据需求方 Data owner:

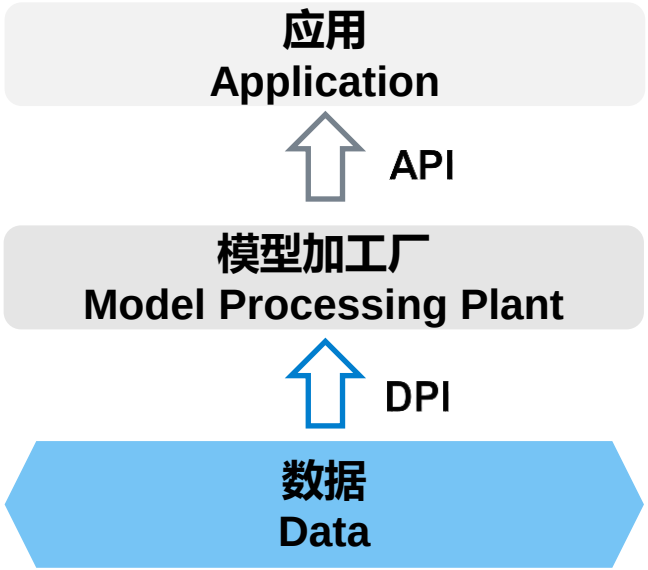
- 信息过滤技术用于建造**防水堡垒**，以确保只有参数等宏观信息才能输出，而不是微原始数据
  Information filtering technology is used to build a **waterproof fort** to ensure that only the macro information such as parameters can be output, but not micro raw data

DPI 运行模式

DPI Operation Mode

核心目标：分享价值不分享数据

Core Objective：Sharing Value, not Sharing Data



默认模式：计算平台在数据调试期间提供**外部替换数据**，供用户测试和调试。用户根据**转换后的样本**进行初步价值挖掘，以确定是否进入模型加工厂挖掘数据

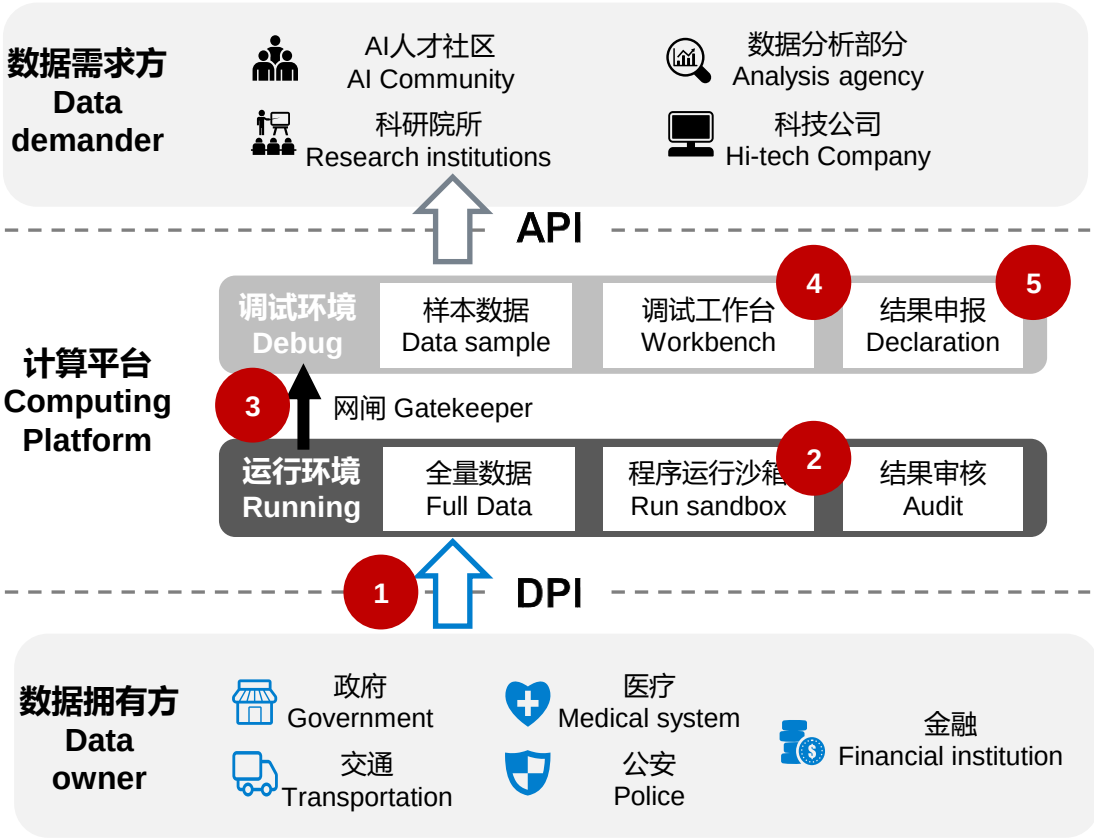
Auxiliary mode: The computing platform provides **external replacement data** during the period of data debugging. The user conducts pro-value mining based on the **transformed samples** in order to determine whether to enter the model processing plant for data mining

扩展模式：计算平台提供远程控制模式，允许数据所有者远程确定**可以授予谁使用数据的权利**。计算平台可以保证获得权利的人只能使用数据生成相应的模型。这达到了**交易使用权而不拥有交易所有权**的目的

Extended mode: The computing platform allowing the data owner to remotely **determine to whom the right to use data can be granted**. The platform can guarantee person granted the right can only use the data to generate model. This achieves the purpose of **trading usage rights without trading ownership**

安全风险分析

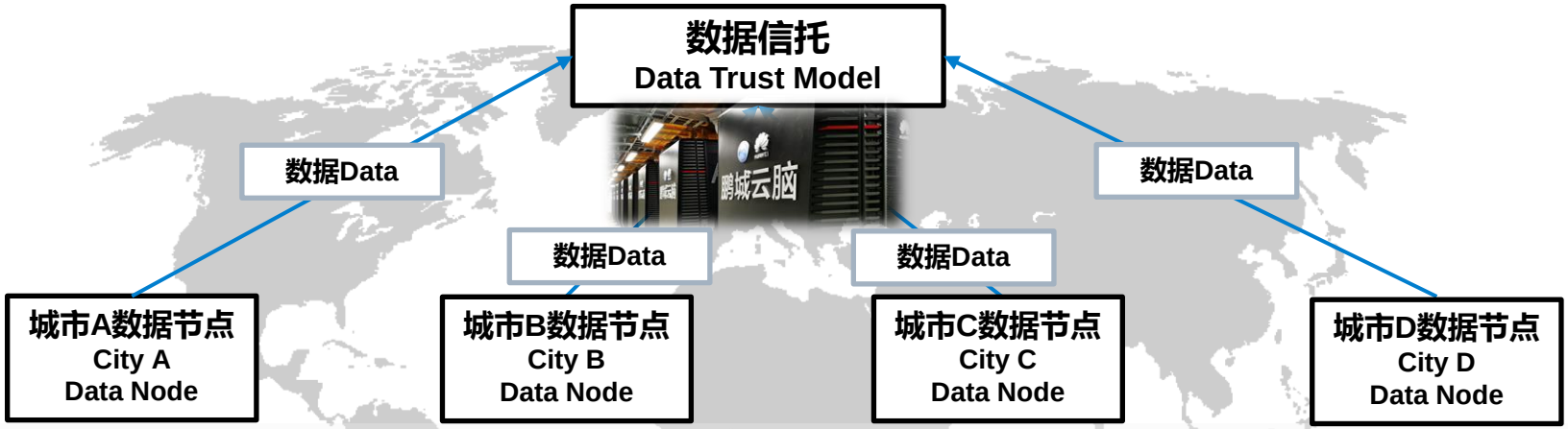
Security risk analysis



1. **在托管数据之前，需要进行安全管理、分类和分级**
   Before the data is hosted, security management, classification and grading are required
2. **防止内部人员和外部代码窃取数据**
   Prevent internal personnel and external code from stealing data
3. **使用网闸交换数据，并使用模拟数据生成技术删除隐私**
   Use a gatekeeper to exchange data, and use simulation data generation technology to removes privacy
4. **在隐私保护的前提下提供调试工具，以获取反馈和调试信息Provide**
   debugging tools under the premise of privacy protection to obtain feedback and debugging information
5. **防止分析结果私下夹带数据并导致泄漏**
   Prevent the analysis result from entraining data privately and causing leakage

打造基于鹏城云脑的数据信托模式

Peng Cheng Data Trust Model



- AI靶场与鹏程云脑对接，创建**数据信任模型**。多个城市的大数据局可以**将政务数据安全托管到鹏程云脑**，通过 AI靶场安全开发，数据需求方只能在调试环境中访问样本数据，形成数据分析程序，并将程序发送到 AI 靶场。经过反复调试的运行环境，并取得较好的效果。向管理层报告审查后，他们可以拿走分析结果，从而实现**"数据不移动，数据可用但不可见"**
- Ai shooting range is docked with Peng Cheng Cloud Brain to create a data trust model. The big data of multiple cities can **safely host government affairs data to Peng Cheng Cloud Brain**, and the government data can be safely opened through the ai shooting range, and data demanders can only be in an shooting range-debugging environment to access sample data to form a data analysis program, and send the program to the AI shooting range. After repeated debugging of the operating environment, and get better results. After reporting to the management for review, they could take away the analysis results, so as to achieve **"data does not move, data available but invisible"**

提纲 Outline

- PCB-II大算力与大模型
  PCB-II Big Computing Power and Large-scale Model
- 数据安全与DPI
  DPI – Data Security
- 开源开放支撑AI生态建设
  AI Ecosystem Supported by Open-source Methodology
- 总结
  Summary

演进中的鹏城云脑 - AI开源生态

Evolving Cloud Brain AI Open-source ecosystem

Open Intelligence启智社区（OpenI）：
是在国家实施新一代人工智能发展战略背景下，新一代人工智能产业技术创新战略联盟（AITISA）组织产学研用通力协作共建共享的开源社区。社区通过构建生态驱动的开源软件、开源硬件和开放数据超级社区，开展“创智”开源创新、“赛智”开放竞赛、“启智”开源培训、“有智”生态建设等赋能活动，沉淀优质内容、凝聚优秀人才、突破核心技术，为中国人工智能的全面发展赋能

iHub鹏城汇智开源托管平台
（简称iHub平台，网址：https://www.ihub.org.cn）定位于面向以汉语为母语的开发者，持续优选汇聚全球人工智能和RISC-V等开源项目与代码，构建我国开源战略资源库，逐步推动我国开源生态的良性发展

启智创新开源社区

OpenI, the Open Source Community with Innovative Governance Model

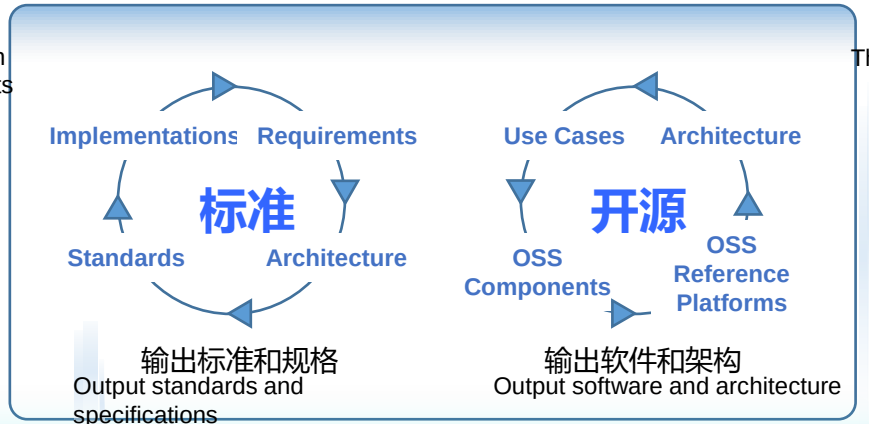
以新一代人工智能为主题的社区共同体，在联盟“一体双翼”的工作架构下，形成以学术与产业融合互动、标准与开源双引擎驱动的开源社区。

With new generation A.I. as theme, AITISA's "duo-driver" as structure, combining research and industry as integration, OpenI will be the open-source community driven by "Standard + Open-source"

以技术专家为主体组成多个技术工作组

One Driver refers to the research groups consist of technical experts

- 标准工作组
- 知识产权工作组
- 投融资工作组
- 开源工作组
- ……



输出标准和规格

Output standards and specifications

输出软件和架构

Output software and architecture

以企业为主体组成多个应用推进组

The other Driver refers to the application groups consist of enterprises

- 智能交通推进组
- 智能医疗推进组
- 智能金融推进组
- 智能教育推进组
- ……

新一代人工智能开源开放平台与开源社区

New generation of open-source A.I. platform and community

开源社区总体架构

General organization structure

从社区纲领、组织架构、成员组成，到知识产权、项目培育、社区激励等方面探索出适应我国人工智能发展特色的开源社区治理模式。

Exploring the suitable governance model for Chinese AI open-source community, from community guiding principle, organization structure, membership structure, intellectual property, project incubation, reward mechanism.

社区纲领
community guiding principle

- ✓ 社区章程
- ✓ 会员管理办法
- ✓ 运营中心管理办法

……

组织架构
organization structure

- ✓ 理事会
- ✓ 技术委员会
- ✓ 秘书处
- ✓ 项目委员会
- ✓ 用户委员会

……

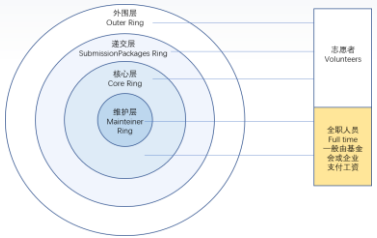
成员组成
membership structure

- ✓ 核心会员
- ✓ 高级会员
- ✓ 会员
- ✓ 联盟成员
- ✓ 开发者
- ✓ 志愿者

……

OpenI 启智

启智社区，确实给力！



知识产权
intellectual property

- ✓ 启智开源许可证
- ✓ 商标

……

项目培育
project incubation

项目培育管道

1. 立项
2. 孵化
3. 毕业

激励机制
reward mechanism

社区贡献评估

启梦行动

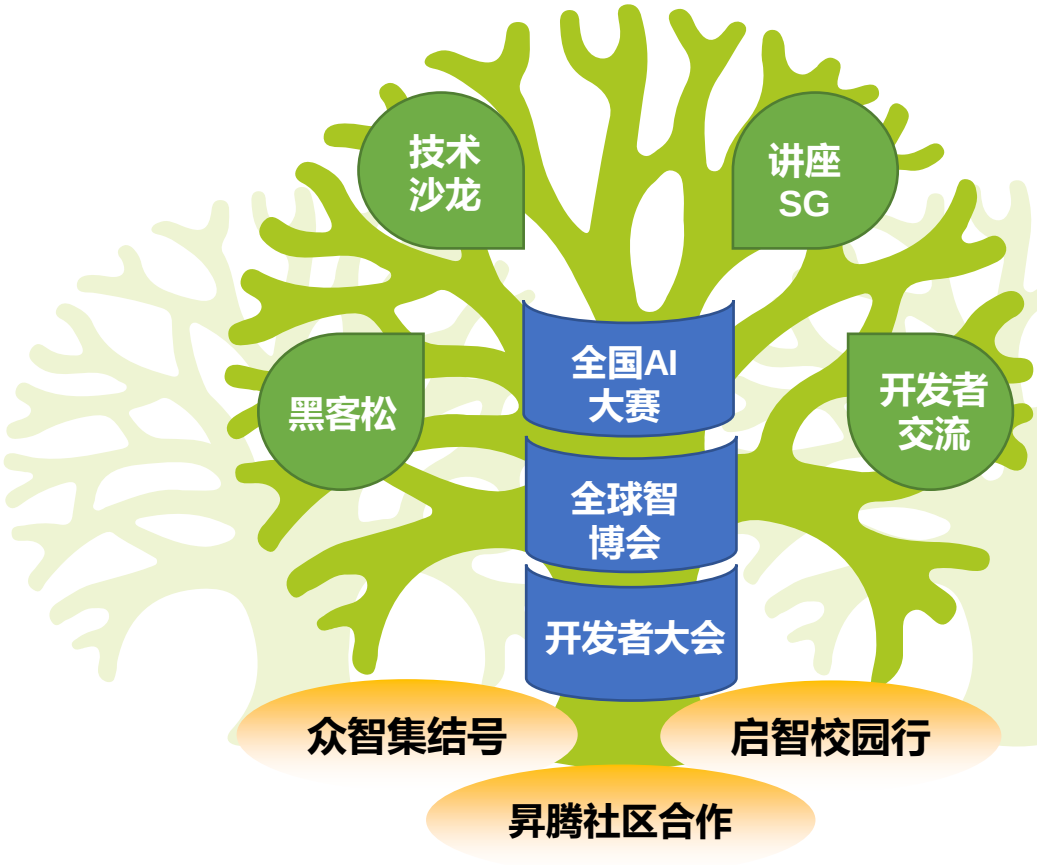
- 2020起三年
- 不低于1000万元
- 激励开发者

社区共同体的有效组织
Structures of the community

社区贡献的长效激励
Incentives for the contribution

开源化的社区运营

Open Source Community Operation



枝芽：多样化运营活动，繁荣开源生态
Branches: various operation activity

主干：打造品牌，提高知名度、曝光率
Stem: building brand, reputation, exposure

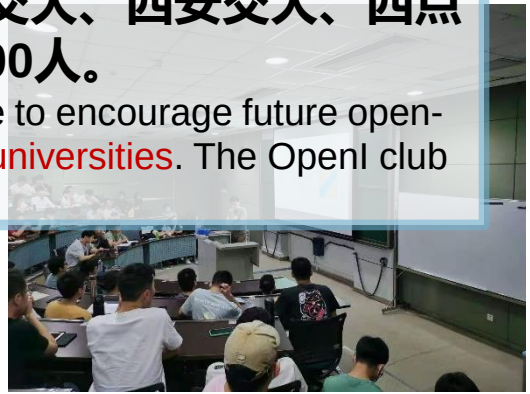
土壤：立足高校培育，广泛孕育人才
Soil: Based on academic facilities, nurturing future talents

土壤：启智校园行，培育未来

Soil: Campus Workshop, Nurturing Future

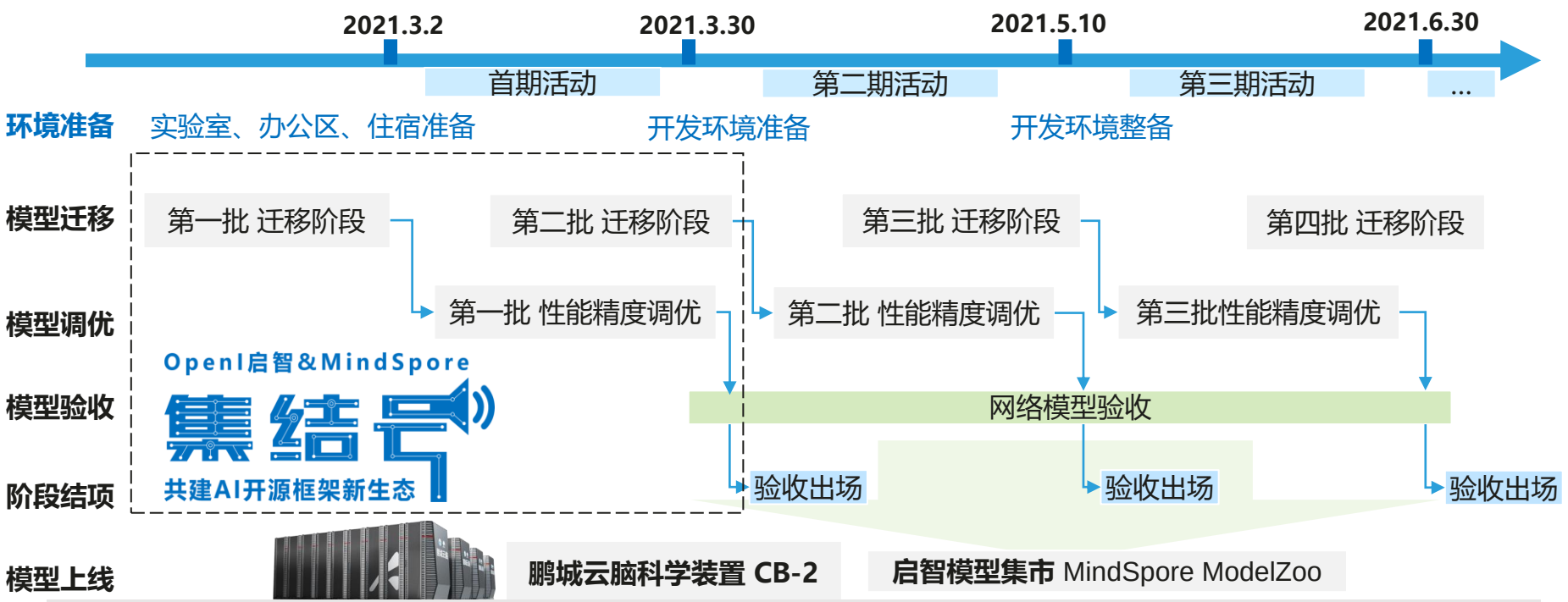
以“培育高校开放精神，培养未来开源人才”为宗旨的启智社区校园行已在哈工大、南方科技大学、国防科技大学、清华、北航、北交大、西安交大、西点、重大等9所高校落地，开源社团超过500人。

Following the principle of “endorsing open-sourcing ideology in college to encourage future open-source talents”, OpenI campus workshop has been held in 9 different universities. The OpenI club has over 500 members.



土壤：启智集结号，众智成城

Soil: OpenI Assembly, Communities Collaboration



“永不落幕持续创新”的集结号系列活动已完成3期活动，为社区贡献70+优质模型，占MindSpore ModelZoo60%以上迁移模型

3 phases of activities have been completed, contributing 70+ high-quality models to the community, accounting for more than 60% of the migration models of MindSpore ModelZoo

土壤：昇腾社区合作，开放共建

Soil: Collaboration with Ascend Community

- 市场活动：以大模型核心开发者为发展对象，开展HAG圈层活动、昇腾人工智能生态大会、HC旗舰大会等多种多样的线下活动扩大技术影响力
- 线上拉新：丰富多彩的线上活动，如：专家精品直播，产品体验官，训练营，各类竞赛等拉新促活留存开发者
- 赋能体系为基石：完善开发者赋能体系，充实大模型知识内容，让开发者在启智社区，昇腾社区“逛”起来，获得成长

与昇腾社区联动，建立首个深圳AI开源大模型开发者兴趣小组

完善生态赋能体系，获取大模型有效开发者，吸纳商业伙伴，构建圈层体系

Linked with the Ascend community to establish the first Shenzhen AI open source large model developer interest group. Improve the ecological empowerment system, acquire effective developers of large models, attract business partners, and build a circle-level system

主干：全球智博会，明星企业+开源创新

Stem: Global AI-Expo, Featured Enterprises + Open-source Innovations

新一代人工智能开放创新平台央视报道，共吸引超过3万名观众到场参观学习，线上直播累计突破800万次，OpenI启智开源展台与技术论坛广受关注，微博连续3天上热搜榜

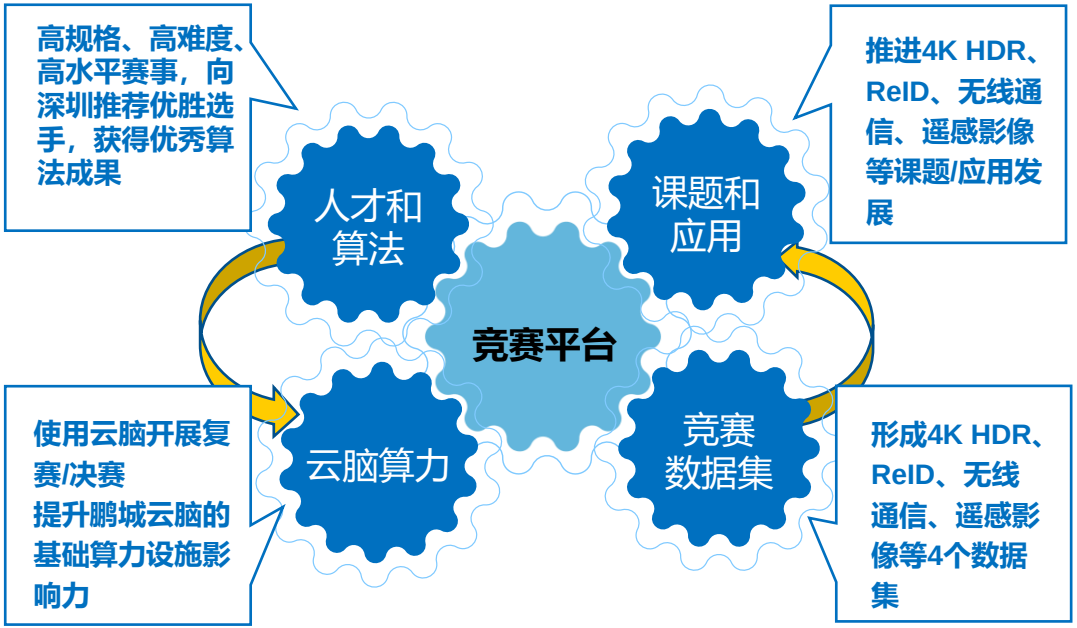
The new generation A.I. Expo is reported by CCTV, attracted over 30 thousand attendees on site , and the number of views on the steaming platform surpassed 8 million. The OpenI Expo gathered extensive attention, and stayed on Weibo trending topic for three consecutive day.

主干：全国人工智能大赛

Stem: National AI Competition(NAIC)

2019、 2020两届“全国人工智能大赛”影响力辐射13个国家和地区， 顶级选手踊跃，累计万人参赛，作品提交4万余次，初步打造AI国际赛事品牌

The 2019, 2020 “national A.I. competition” together involves 13 countries and regions, tens of thousands of topnotch competitors, over 40 thousands submissions, built the branding for a preliminary international A.I. competition.



大规模数据集构建：

AI+无线通信： 紧扣前沿技术,鹏城实验室自有版权600万组信道采样数据；

AI+ReID： 创新了跨多位来源数据的脱敏处理方法，形成全球最大规模高难度数据集，应用广泛；

AI+遥感影像： 100万+语义分割样本数据，数据覆盖不同地区多种地物类型；

AI+4K HDR： 1 万对（HDR和SDR成对）视频序列，200 万帧图像，当前业界最大规模视频重建数据集解决AI视频内容处理的难题。

主干：中国软件开源创新大赛

Stem: China Software OS Innovation Competition

依托鹏城云脑II超过80P资源，本届大赛将支持超过5000名选手角逐赛道二——任务挑战组的35道赛题各项奖项，并选择优胜队伍参加大赛最终决赛，为AI开源生态选拔优秀人才。

Relying on the more than 80P resources of PCB-II, this competition will support more than 5000 players to compete for Track Two-35 questions of the task challenge group, and select the winning teams to participate in the finals of the competition, which is an AI open source ecosystem Select outstanding talents.

枝芽：多样化的运营活动，繁荣开源生态

Branches: Various Community Activities

2018

3月31日

启智开源许可证 1.0 发布（深圳）

7月18日

启智平台在鹏城实验室启动（深圳）

12月5日

OpenI开源社区成立（厦门）

2019

1月16日

启智社区2019年第一次工作会议（深圳）

2月22

OpenI技术委员会2019年第二次会议（北京）

3月7日

启智平台OpenI社区发布首批开源项目（青岛）

5月8日

2019全球智博会（苏州）

5月25日

与Linux和Git之父Linus的炉边会谈

5月31日

启智沙龙：来留仙洞聊聊开源那些事!

2020

7月19日

启智沙龙：数据开放的现状与未来

8月31日

“基于OpenI海参的视频编码大赛

12月21日

2019 启智开发者大会

1月17日

Linux、RISC-V等基金会负责人参访

8月12日

启智社区优秀开发者激励计划“启梦行动”发布

8月15日

新一代人工智能开源技术与生态建设论坛

9月18日

启智社区校园行首场——哈工大（深圳）站

10月31日

哈工大（深圳）校园行第二场

11月7日

Linux & 启智AI开源日

2021

12月2日

2020 启智开发者大会

12月3日

启智&Linux校园行——南科大场

1月3日

启智&MSG北航校园行活动

3月1日

启智&MindSpore首期集结号开发活动

3月13日

启智集结号经验分享——启智校园行活动

3月31日

中国软件开源创新大赛——任务挑战赛

4月6日

启智&MindSpore第二场集结号开发活动

4月14日

启智校园行——西交活动

鹏城汇智：开源托管大集市

Ihub: Open-source Code Hosting Market Place

开源开放资源：算法+数据+开发环境

Open Source: Algorithm + Data + Development Environment

鹏城云脑开源社区服务环境

实训 | 竞赛 | 众包 | 孵化 | 科研 | 协作 | ……

数据、训练、部署、推理全栈开发工具

数据预处理 / 数据标注 → 数据共享 → 模型训练 / AutoML工具 → 训练容器 → 模型压缩 / 模型转换 → 统一标准 → 模型部署 / 在线运行 → 监控分析

持续汇聚开源开放资源

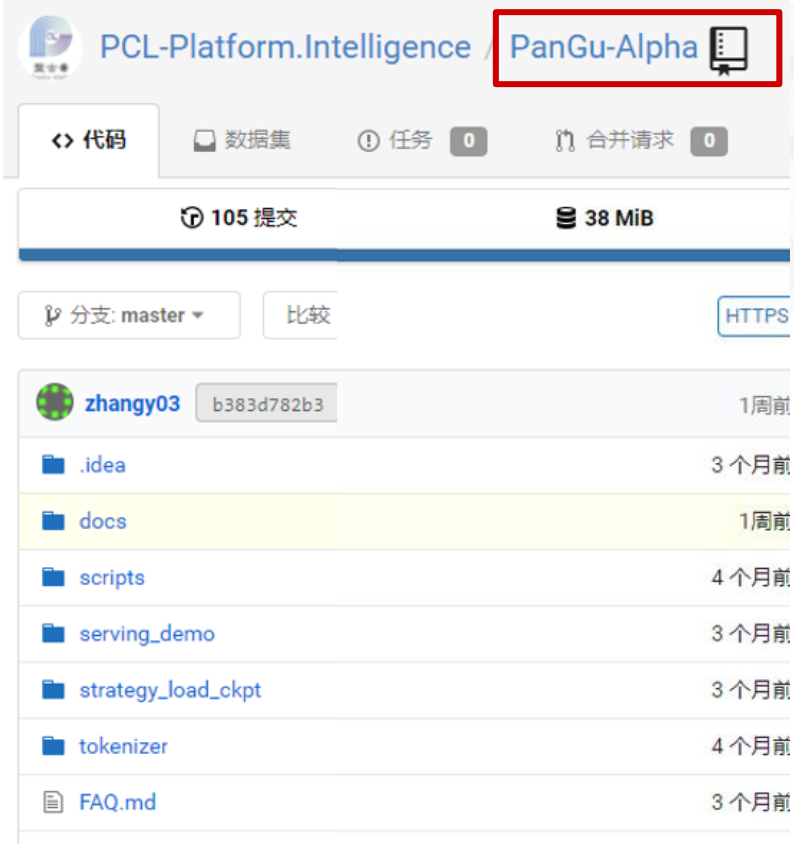
开放数据集
6个应用领域，516个数据集

https://www.ihub.org.cn/

开源代码
25个领域，2万个AI算法镜像

典型开源项目：盘古 & 悟道

Typical Open Source Project: Pangu & WuDao



提纲 Outline

- PCB-II大算力与大模型
  PCB-II Big Computing Power and Large-scale Model
- 数据安全与DPI
  DPI – Data Security
- 开源开放支撑AI生态建设
  AI Ecosystem Supported by Open-source Methodology
- 总结
  Summary

Summary: AI for Good

- 为AI社区提供足够的算力、数据、模型
  Provide computing power, data, and model for AI community
  - PCB-II，足以支撑大模型训练
    PCB-II for pre-training on Large-scale Model
  - DPI, 保证数据安全
    DPI can protect data privacy
  - OpenI，支撑AI生态发展
    OpenI support the better development of AI Ecosystem

THANK YOU

www.pcl.ac.cn

Contacts: gaow@pcl.ac.cn