



# 大模型及天文科学数据处理思考

崔辰州

国家天文科学数据中心

NAOC, CAS

感谢：齐法制、陶一寒、张正德、阿里云、达摩院、之江实验室、Szalay……

# Disclaimer

- I'm not a ML or AI expert



资料提供：齐法制、陶一寒、张正德、阿里云、达摩院、之江实验室、Szalay……

# ChatGPT简要技术回顾

- 为何出圈：不再是传统意义的人机对话系统，而是以自然语言为交互方式的通用语言处理平台
  - **技术有突破**：基础数据+核心模型+优化算法
  - **应用接地气**：几乎可以完成所有语言相关的功能
  - **效果超预期**：相比于先前系统效果产生质的飞跃
- 超出预期的交互体验
  - (1) 通用的意图理解能力，(2) 强大的连续对话能力，(3) 智能的交互修正能力，(4) 较强的逻辑推理能力



# ChatGPT简要技术回顾

- ChatGPT是以产品思维驱动的重大集成创新成果
- ChatGPT是OpenAI自2018年以来坚持生成式AI、长期技术积累，量变产生质变的重大成果，是迈向AGI的阶段性成果
- ChatGPT的集成技术组成
  - ChatGPT的基础模型架构**生成式解码器GPT**来源于**Google于2017年**提出的**Transformer**
  - ChatGPT的**指令学习**技术同样来源于**Google于2021年**提出的指令学习模型**FLAN**
  - ChatGPT基于**人类反馈的强化学习**技术来源于**OpenAI自己2017年**提出的优化算法**PPO**和**2021年**提出的基于**人类反馈的自动摘要**方法

# 大模型的通用性局限

① 事实检索性和复杂  
计算性任务效果差

朱自清散文《背影》的第一句是什么？

$$12^{12} = ?$$

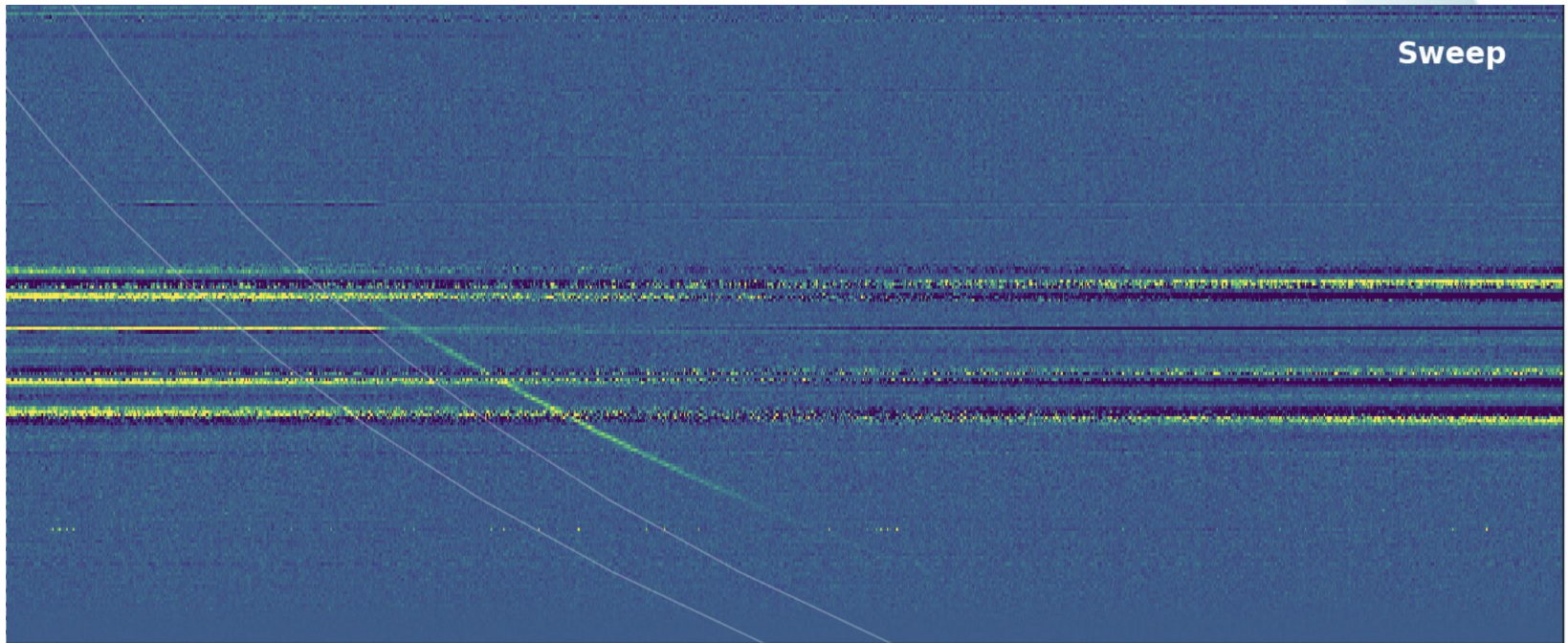
② 实时性、动态变化  
性的任务无法实现

明天北京的天气怎么样？

③ 超出模型训练数据  
类型的任务无法完成

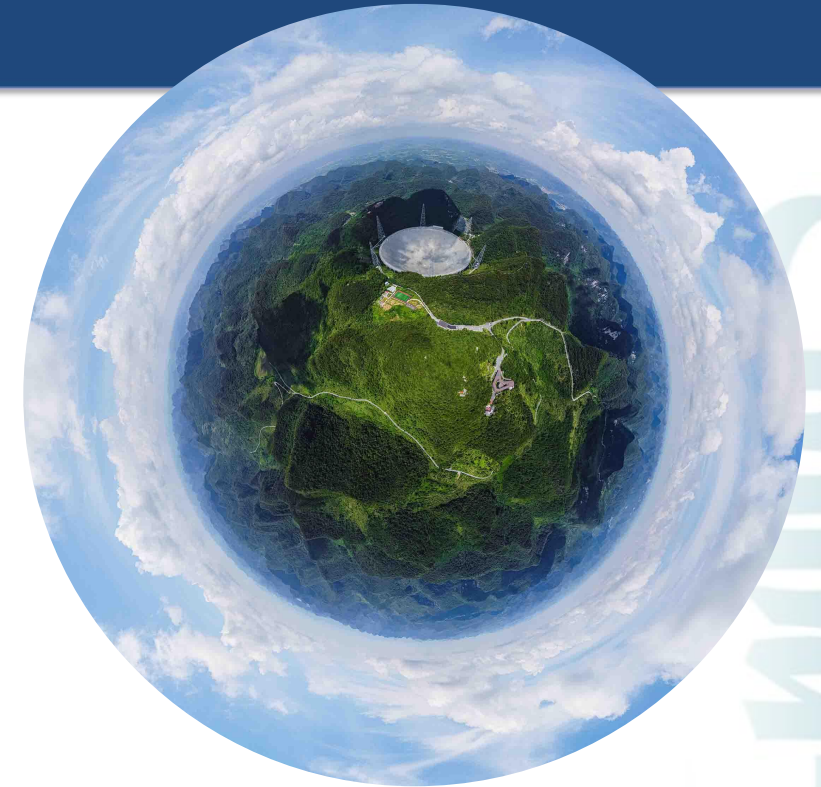
这个图片中的路牌上写的是什麼？

# ASTRONOMY

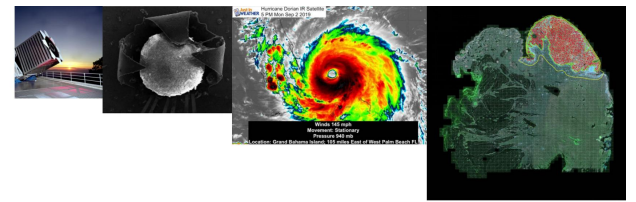


# Astronomy in the bigdata era

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected



	接收机端	科学数据
FAST-19(now)	38GB/sec	3GB/sec 90PB/year
FAST-PAF(2027)	720GB/sec	未知



## The Impact of Astroinformatics

Alex Szalay

Institute for Data Intensive Engineering and Science  
**idies**

JOHNS HOPKINS  
UNIVERSITY

# Today's Science Environment

- For a long time science was bimodal, small PI projects vs Big Science
- It is changing again today – more in the **middle**
  - NSF Mid-Scale projects, NIH U01, private collaborations, public-private partnerships (Sloan Digital Sky Survey, PFS, Human Genome ...)
  - Typically: create a unique instrument, use cutting edge technology, take risks, push budgets to the limit (and beyond) to maximize science, generate large amounts (petabytes) of data
  - Enormous fresh energy liberated!
  - At the “sweet spot” for science
  - Do computations on a shoestring
  - Generally: computations will be done opportunistically
  - AI is emerging everywhere

*In this new model there is much more focus on generating (and sharing) novel data*



# Agility vs Tenacity

- We see extremely agile changes in the industry (Google, Facebook, Amazon, Microsoft)
- Industry attention span for science is spotty, lasts a few years at most
- Universities cannot compete with the industry in agility – faculty hires are for 40 years...
- **But we can compete in tenacity and unique high-value data!**
- There are more and more midscale projects at Universities (few \$M)
- Using AI to interpret the data, discover new patterns, new phenomena
- This requires the data sets to be “AI-ready”

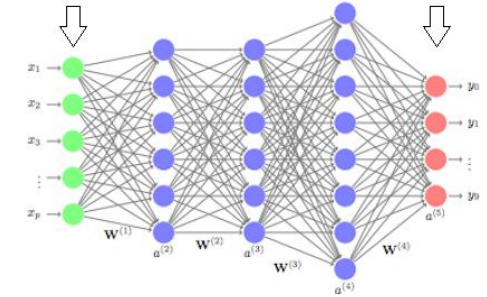
*The big breakthroughs will always come from a unique data set (Human Genome, SDSS, ImageNet) -- combined with a disruptive idea*

# Computing and Data in the Era of AI

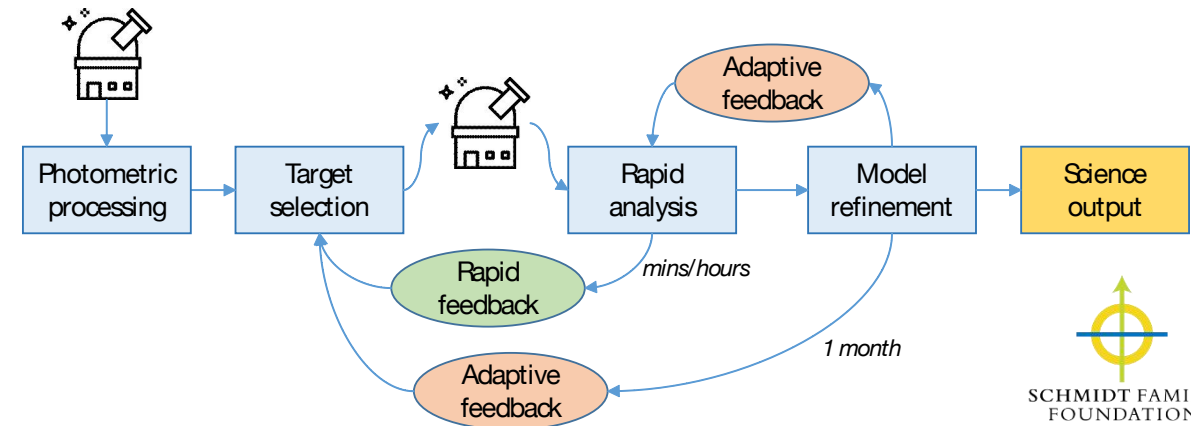
- Hardware used to be the capital investment, scientists mostly wrote their own software
- Today it is the opposite:
  - Hardware became a commodity, available everywhere
  - Today: data sets and software are becoming the capital investment
- Software technologies change every two years, in AI faster
- Code carpentry emerging
- Digital twins
- Codes need to be ARCO (Analysis Ready and Cloud Optimized)
- Software written by graduate students is not good enough
- We need serious software professionals of industrial caliber

# What is Next: Experimental Design by AI?

- Instead of more data, we need **MORE RELEVANT DATA**
- Need to dramatically improve experiment design....
- Artificial Intelligence in large-scale experiments:  
    use AI **before** we collect the data
- Example: Next Generation Astronomical Surveys
  - *Observing spectra is 1,000 times more expensive than imaging*
  - *Use feedback from observed targets and improve target selection via reinforcement learning (active learning)*
- **Fifth Paradigm:** when algorithms make the decisions about our experiments?



Put the telescope (or the microscope)  
in the reinforcement loop!



*Supported by the Schmidt Family Foundation  
at JHU and Princeton: Use AI Feedback for the PFS  
project*



***If we let AI drive our cars, it can drive our microscopes and telescopes...***

# Virtual Observatory (VO)

- Virtual Observatory (VO) is a data-intensively online astronomical research and education environment, taking advantages of advanced information technologies to achieve seamless, global access to astronomical information.
  - 虚拟天文台是通过先进的信息技术将全球范围内的研究资源无缝透明连接在一起形成的数据密集型网络化天文研究与科普教育平台。
- The Virtual Observatory (VO) aims to provide a research environment that will open up new possibilities for scientific research based on **data discovery, efficient data access, and interoperability.**
- International Virtual Observatory Alliance

--"facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory."



# China-VO

- Chinese Virtual Observatory (China-VO) is the national VO project in China initiated in 2002 by Chinese astronomical community leading by National Astronomical Observatories, Chinese Academy of Sciences.
- China-VO became a member of the IVOA with the recommendation of Dr. Jim Gray

## R&D Focuses

- China-VO Platform
- Unified Access to On-line Astronomical Resources and Services
- VO-ready Projects and Facilities
- VO-based Astronomical Research Activities
- VO-based Public Education





# Discover IVOA with ChatGPT

Adrian Damian  
Bologna 2023

Chinova-170

## ChatGPT

- Why ChatGPT?
  - Large Language Models (LLM)
  - No time/resources to try the other available ones: Bing, Bard, Copilot, etc.
- Why discover IVOA?
  - Multiple uses of ChatGPT
  - "AI is going to give every student a personalized tutor" - Sal Khan the founder of Khan Academy
- "Methodology" - series of conversations with ChatGPT about IVOA related subjects: standards, libraries, examples etc.

## General Observations

- Good general grasp of the terms/terminology
- Interesting view of the IVOA:
  - TAP is always associated with IVOA
  - Main applications (Aladin, TOPCAT, STILTS, PyVO) are often mentioned
  - Great exploration tool

### HOWEVER...

- Knowledge cutoff - September 2021.
- It doesn't reference the source of the information
- Main Problem: reliability.

**"Our biggest concern was around factuality, because the model likes to fabricate things." - John Schulman (OpenAI co-founder)**



## Incorrect Answers

What is pyvo?

Answer:

“Pyvo is a Python package that provides a simple and efficient interface for working with data from the Czech beer culture, including beer styles, breweries, and beer events. It includes data from the Czech Beer Guide, Open Brewery DB, and the Czech Beer Culture website.

With Pyvo, users can easily access and manipulate data related to Czech beer, including information on beer styles, breweries, and events. This package can be useful for beer enthusiasts, researchers, or anyone interested in exploring data related to Czech beer culture.

Pyvo can be installed using pip, the Python package manager, and it is open-source software, released under the MIT license.”

## How to Improve?

- ChatGPT feedback (requires a specialist)
  - Not clear how fast the feedback is incorporated but progress is visible
- Better prompts
  - Avoid the GIGO effect.
  - Prompt engineers?
  - Idea is to give context, ask ChatGPT to assume the identity of a profession,
- Better data to train on
  - Use a ChatBot trained with custom data
  - Explosion of tutorials/websites etc.

## Final Thought

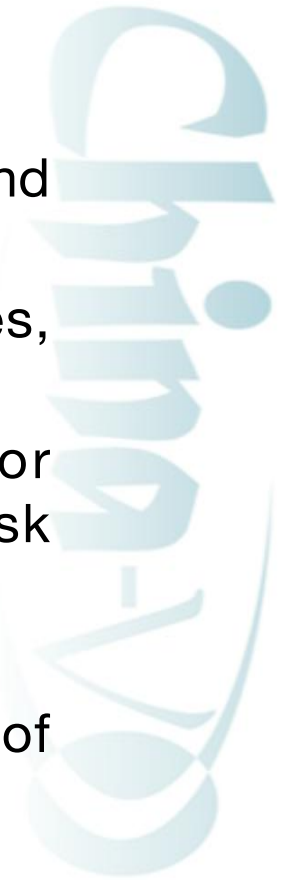
ChatGPT and all the new related AI tools and ChatBots offer a novel/disruptive way for exploring and consuming information and we will need to adapt to that.





# Potential of foundation models for Astronomy

- Foundation models mark the beginning of a new era in machine learning and artificial intelligence.
- How can it serve Astronomy
  - Apply LLMs to Astronomy papers, integrate domain knowledge for searching and question answering tasks
  - General representation for different data types (images, spectra, time series, catalogue etc.) for astronomical data analysis tasks
  - To serve as research assistant for more complicated task, e.g. plan for observation, generate report and figures. automatically decompose the task (AutoGPT)
- From VO perspective
  - Make the data ready - Standardize the representation for astronomical data of different type and from different sources?



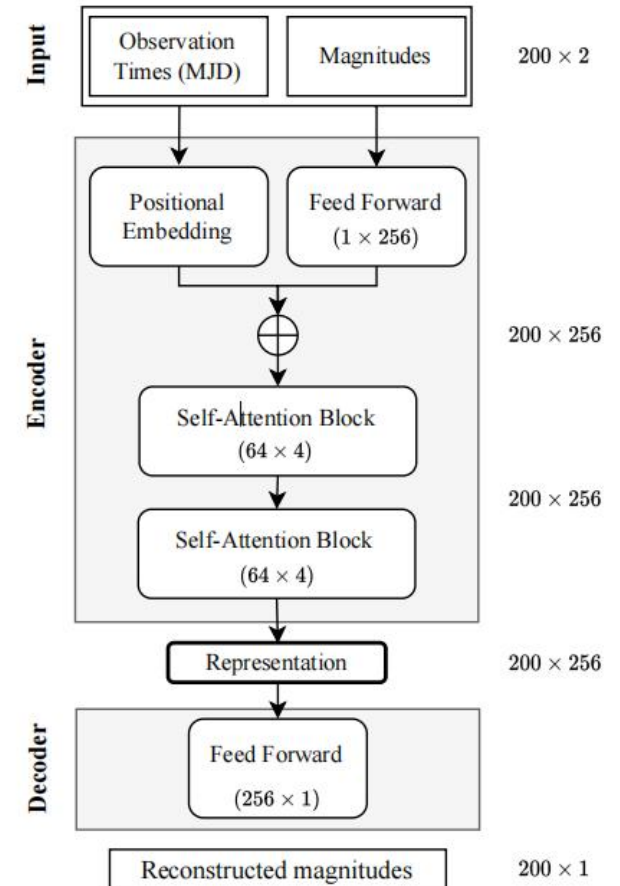
# Related work in Astronomy

- **ASTROMER**: A transformer-based embedding for the representation of light curves
  - pre-trained on millions of light curves from different surveys (MACHO, OGLE, ATLAS)
  - representation to create informative light curves embeddings
  - finetuned for solving downstream tasks, e.g. classification of variable stars, predicting physical parameters



<https://www.stellardnn.org/projects/astromer/index.html>

C. Donoso-Oliva et al. ASTROMER: A transformer-based embedding for the representation of light curves.

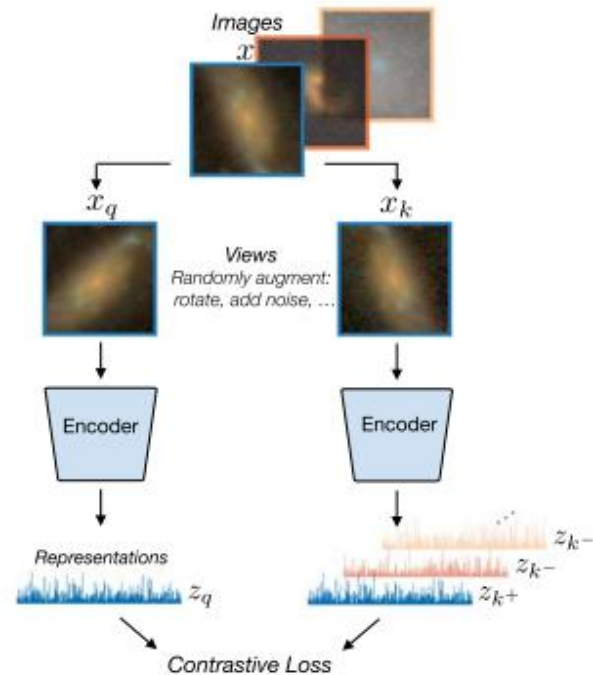


# Related work in Astronomy

- Self-supervised Representation Learning for Astronomical Images
  - multiband galaxy photometry from the Sloan Digital Sky Survey (SDSS) to learn image representations

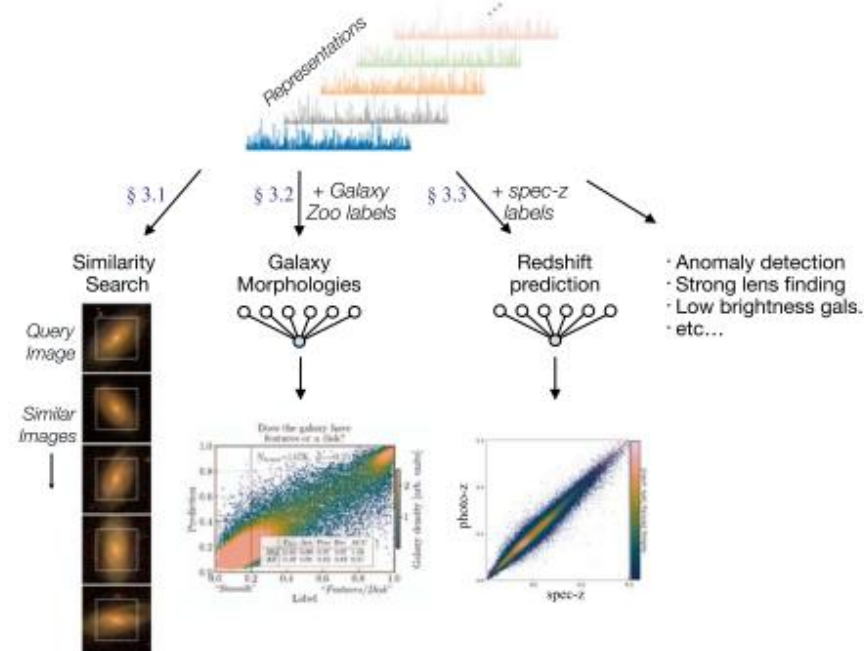
## 1. Self-supervised contrastive representation learning

Learn representations in an unsupervised manner



## 2. Downstream tasks

Use representations for a variety of applications



# Related work in Astronomy

- Towards Galaxy Foundation Models with Hybrid Contrastive Learning
  - 552k labelled and 1.34m unlabelled galaxies from five telescopes and four Galaxy Zoo campaigns
  - ML-friendly galaxy datasets for major Galaxy Zoo (<https://github.com/mwalmsley/galaxy-datasets>)

Name	Method	PyTorch Dataset	Published	Downloadable	Galaxies
Galaxy Zoo 2	gz2	GZ2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	~210k (main sample)
GZ Hubble*	gz_hubble	GZHubble	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	~106k (main sample)
GZ CANDELS	gz_candels	GZCandels	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	~50k
GZ DECaLS GZD-5	gz_decals_5	GZDecals5	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	~230k (GZD-5 only)
GZ Rings	gz_rings	GZRings	<input type="checkbox"/>	<input checked="" type="checkbox"/>	~93k
GZ DESI	gz_desi	GZDesi	<input type="checkbox"/>	WIP	WIP
CFHT Tidal*	tidal	Tidal	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1760 (expert)

Mike Walmsley et al. Towards Galaxy Foundation Models with Hybrid Contrastive Learning

# Call for AI Proposals

	G	H	I	
1	1、姓名	2、所在单位	3、研究方向	4、对天文领域大模型的设想和需求（包括拟开展应用的科研场景、功能设想、所使用的相关科研数据等）
2	沈世银	上海天文台	星系天文学	结合图像，光谱数据的星系物理（结构，形成和演化）性质研究；结合光谱及时域数据的星系中心活动星系
3	宋子豪	之江实验室	谱线分析	天文通用观测数据的目标识别分割，观测光谱发射线认证
4	王赫	中国科学院大学·国	引力波天文学	期待阿里可以开发的大模型，特点特性在于1 广泛支持多模态数据上，比如要适用性非常广，模态类型丰富
5	王楠	中科曙光	人工智能基础设施	大模型大规模训练扩展，相比小模型的硬件加速比和硬件利用率
6	姚嵩	上海天文馆（上海科	光学夜天文观测，天文	天文学概念解释，知识图谱梳理，专业文献整理、解读，论文优化，常用公式和算法介绍以及程序编写，科
7	李观	中国科学院计算机网	可视化，可视分析	无
8	王磊	上海天文馆	星系团	天文科普场馆智能助手。内容包括:天文知识问答，天文馆展项讲解
9	左肖雄	国家天文台	天文信息技术	数据分类，观测助手
10	朱文白	中国科学院国家天文	射电天文；射电望远镜	暂无
11	钱磊	中国科学院国家天文	射电天文	公式推导，自动生成标准格式实验报告，识图
12	崔辰州	国台	VO	尚不是很清楚
13	邵务俊	国家天文台	天文信息技术	文献数据挖掘
14	张颜荣	广州大学	脉冲星	脉冲星搜寻，消除射电干扰
15	谢海闻	长春人造卫星观测站	望远镜控制	迅速了解前沿天文研究对观测数据的需求，以推进望远镜技术的研究
16	王赫	中国科学院大学	引力波天文学	1. 支持多模态（如时域、频域、时频域等）2. 开放 API 接口，允许科研人员开发科研插件（可接入不同领域
17	梁博	国科大杭州高等研究	引力波参数反演（主要	利用大规模模型进行数据的压缩，在GW领域其实光数据集的生成可能都需要100W，哪怕利用目前最好的单
18	田海俊	杭州电子科技大学	天文信息学、天体物理	1. 基于欧空局盖亚（Gaia）卫星数据，系统性搜寻银河系内星团、星协、星流等性质相识的子结构。2. 基于
19	支挥	中国科学院国家天文	望远镜控制	基于大模型的目标源提取（图像分割）其余不太了解，想学习一下
20	吴明磊	山东工商学院	恒星低质量光谱	想了解如何获取恒星光谱的连续谱
21				



# NAOC – Alibaba Cloud Partnership



NAOC · 阿里云

## 计算·探索未知的价值

Computing Decodes the Unknown World

2016年10月13日·中国科学院国家天文台与阿里云发布  
共建天文大数据联合研究中心

### 成立天文大数据研究中心

飞天系统+人工智能

- 海量数据上云
- 数据深度挖掘
- 虚拟天文台共享平台

#### "天文数字"有多大?

- FAST: 500米口径球面射电望远镜, 每天: 50TB
- LAMOST: 郭守敬望远镜, 每年: 10TB
- SKA: 平方公里阵列射电望远镜, 每秒: 50TB

1GB = 1,000 Megabytes (MB)  
1TB = 1,000 Gigabytes (GB)  
1PB = 1,000 Terabytes (TB)  
1EB = 1,000 Petabytes (PB)  
1ZB = 1,000 Exabytes (EB)

美国学者预测:  
到2025年天文数据采集量为每年25ZB (250kTB)

#### 天文大数据挖掘与研究

- 计算和大数据挖掘能力提升
- 促进数据密集型从量变到质变

天文数据公共资源

阿里数据目标: 数据融合, 智能化算法, 智能挖掘

#### 阿里云与数据密集型学科

- 天文
- 地理
- 气象
- 医疗
- 量子

#### 从中国到世界

中国望远镜 + 中国计算能力 + 互联网

服务世界天文用户  
促进基础学科发展

# 行业GPT技术框架

**行业数据**

第二章 危害公共安全罪

第一百一十四条 【放火罪、决水罪、爆炸罪、投放危险物质罪、以危险方法危害公共安全罪、劫持航空器罪、暴力危及飞行安全罪】

第一百一十五条 【放火罪、决水罪、爆炸罪、投放危险物质罪、以危险方法危害公共安全罪】

广东省台山市人民法院  
刑事判决书

广东省台山市人民检察院  
起诉书

公安机关办理刑事案件程序规定

第一章 刑法总论 / 17

第一节 刑法的概念、渊源与分类 / 17

第二节 刑法的性质、机能与目的 / 22

第三节 刑法的制定、修改与根据 / 27

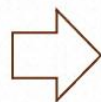
第四节 刑法的规范、体系与解释 / 30

第二章 刑法的基本原则 / 32

...

网页  
文档  
图表  
书籍

数据预处理



**行业知识**

预训练

1 2 3 4 ... 2046

连续文本

2021年8月至11月期间，被告人陈艳娟伙同陈小冰（另案处理），隐瞒自己真实姓名、年龄、离异且有女儿等事实，与被害人阮某晶交往、“结婚”（仅摆酒，无婚姻登记），骗取被害人阮某晶及其父母信任，从而先后以考取机动车驾驶证、购买喜饼、金器、结婚用品、手机及收取礼金、办理出国体检、没钱花、交房租等为由骗取被害人阮某晶及其父母的财物共计106399元和价值17395元的金器首饰。

连续文本

2021年8月至11月期间，被告人陈艳娟伙同陈小冰（另案处理），隐瞒自己真实姓名、年龄、离异且有女儿等事实，与被害人阮某晶交往、“结婚”（仅摆酒，无婚姻登记），骗取被害人阮某晶及其父母信任，从而先后以考取机动车驾驶证、购买喜饼、金器、结婚用品、手机及收取礼金、办理出国体检、没钱花、交房租等为由骗取被害人阮某晶及其父母的财物共计106399元和价值17395元的金器首饰。

问答数据构造



**知识问答**

Instruction tuning

Pretrained LM → Instruction-tune on many tasks: B, C, D, ... → Inference on task A

Model learns to perform many tasks via natural language instructions

Inference on unseen task

调用本地工具

调用本地工具

检索工具

数据库 知识图谱 搜索引擎

本地知识库

指令和问答

Instruction	以下是一起案件事实描述，请问被告人触犯了什么罪行？
Input	2021年8月至11月期间，被告人陈艳娟伙同陈小冰...
Output	诈骗罪

指令和问答

Instruction	以下是一起案件事实描述，请问被告人触犯了什么罪行？
Input	2021年8月至11月期间，被告人陈艳娟伙同陈小冰...
Output	诈骗罪

调用本地工具



**信息准确**

工具调用 tuning

Pretrained LM → Instruction-tune on many tasks: B, C, D, ... → Inference on task A

Model learns to perform many tasks via natural language instructions

调用本地工具

检索工具

数据库 知识图谱 搜索引擎

本地知识库

指令和问答

Instruction	以下是一起案件事实描述，请问被告人触犯了什么罪行？
Input	2021年8月至11月期间，被告人陈艳娟伙同陈小冰...
Output	诈骗罪

检索工具

数据库 知识图谱 搜索引擎

本地知识库

持续提升

使用反馈



基于反馈学习 RLHF

Pretrained LM → Instruction-tune on many tasks: B, C, D, ... → Inference on task A

Model learns to perform many tasks via natural language instructions

调用本地工具

检索工具

数据库 知识图谱 搜索引擎

本地知识库

指令和问答

Instruction	以下是一起案件事实描述，请问被告人触犯了什么罪行？
Input	2021年8月至11月期间，被告人陈艳娟伙同陈小冰...
Output	诈骗罪

专家反馈

早直播是一种节水、节肥、省工、省时、省种子的新型种植方法，具有以下优点：

1. 节水：早直播不需要灌溉，可以直接播种，减少了水资源的浪费，避免了土壤过度湿润和水分蒸发，避免了土壤中水分的积累，减少了水分的蒸腾和土壤的湿度，从而减少了水的消耗。
2. 节肥：早直播不需要施肥，因为它是一种无土栽培方法，不需要施用基肥和追肥，避免了肥料的浪费和土壤的污染。
3. 省工：早直播不需要喷洒农药，因为它不需要人工除草，节省了农药的浪费和环境的污染。
4. 省工：早直播可以减少人工的投入，因为它不需要灌溉和施肥，从而减少了人工的投入。
5. 省种：早直播可以提高播种效率，因为它不需要灌溉和施肥，从而减少了种子的浪费。
6. 省种子：早直播可以减少种子的消耗，因为它不需要灌溉和施肥，从而减少了种子的消耗。

水稻早直播是一种水稻种植的方式，它相对于传统的水稻种植方式有以下优点：

1. 节省水资源：早直播不需要灌溉，不需要消耗大量的水资源，这对于水资源短缺的地区来说非常重要。
2. 提高种植效率：早直播的种植效率比传统的水稻种植方式更高，因为它不需要进行灌溉和施肥，种植时间也更短，可以更快地种植和收获。
3. 减少病虫害：早直播的种植方式不需要进行灌溉，因此也不会滋生病虫害，这对于减少农药和化肥的使用量，提高农产品质量有很大的好处。
4. 适应性更广：早直播的种植方式“适应性更广”，因为它不需要进行灌溉，所以可以在更加干旱和贫瘠的土地上种植水稻，这对于提高水稻的产量和稳定性有很大的好处。



# 大模型协作



专家团队  
技术团队



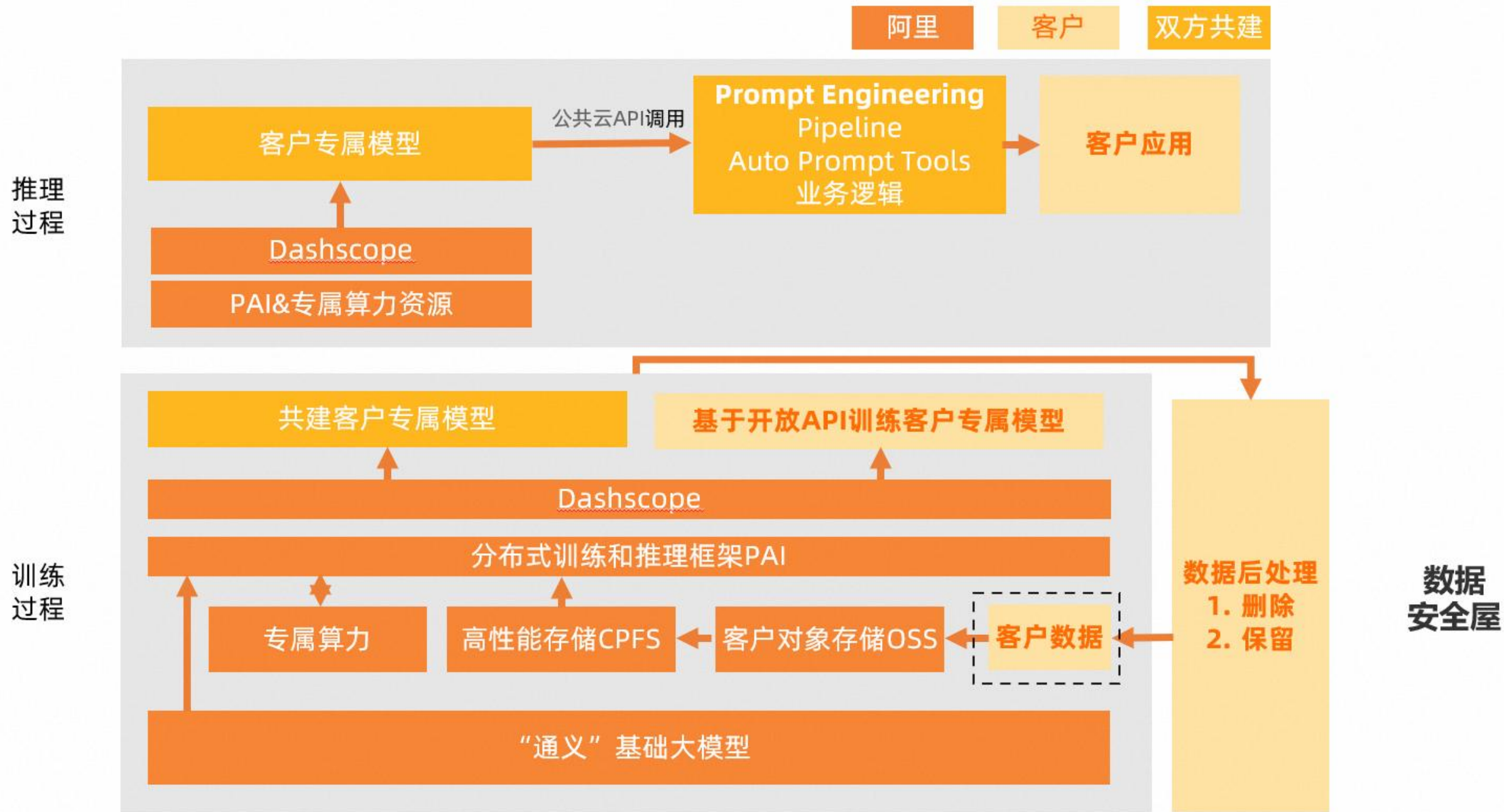


# 资源需求

- 算力资源，8张A800卡，支撑模型训练，数张A10卡，支撑模型服务
- 专家团队，提供业务咨询和流程设计
- 专家团队，提供和协助处理文档/问答/反馈数据
- 专家团队，确定使用场景
- 应用开发团队，本地系统对接和前端服务开发等



# 客户专属模型&行业模型数据安全保障



# 发展高能物理领域大模型的路线

应用层面

文本/图像任务涌现

领域知识问答、学科代码、学科图像处理

科学任务涌现

Jet分类、Shower模拟等复杂多样的科学任务

系统涌现

提出问题、验证问题、发现知识

算法层面

HEP文本和图像大模型

HEP科学数据大模型

BES III



SNS

...

数据层面



领域文本数据



领域图像数据

领域科学数据

JetClass数据集

JetImages数据集等

2028

算力层面

8\*A100 (80G) GPU

GPU → NPU


NPU、DCU

AI科学家



2035

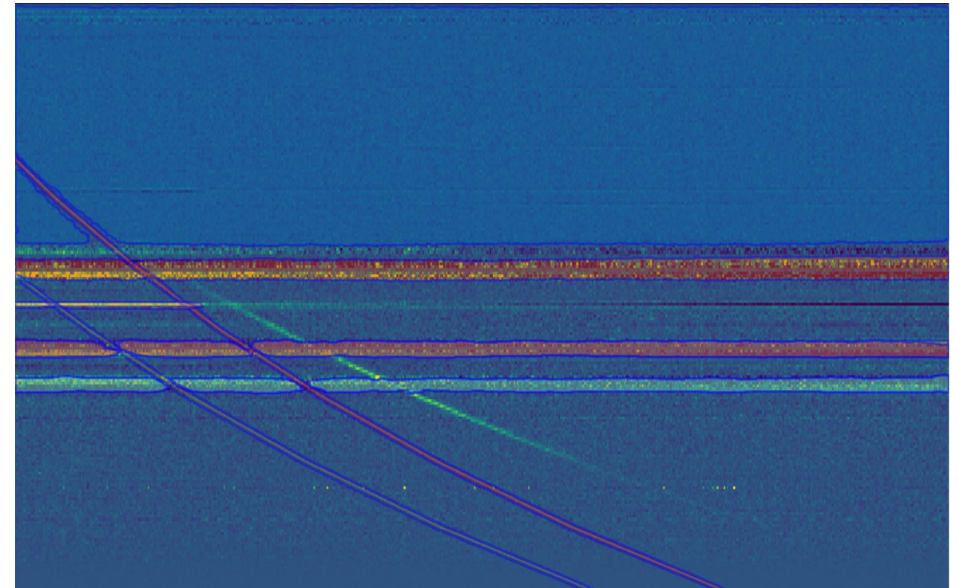
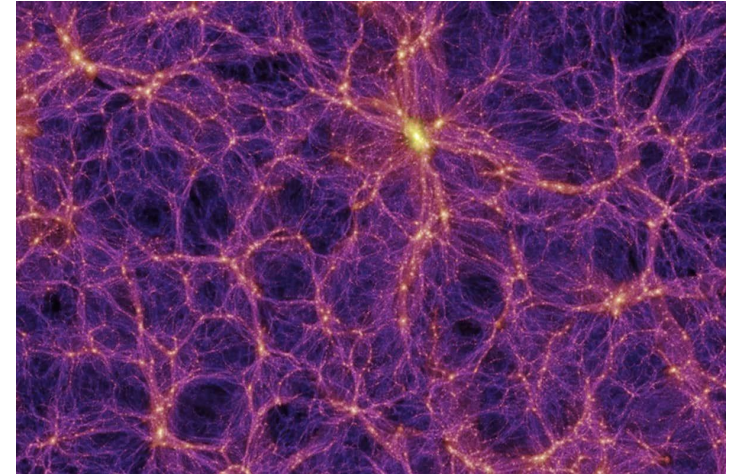
$$\begin{aligned} \mathcal{L}_{\text{StandModel}} &= \mathcal{L}_{\text{gauge}} + \mathcal{L}_{\text{Higgs}} \\ &+ \mathcal{L}_{\text{fermion}} + \mathcal{L}_{\text{Yukawa}} \end{aligned}$$

文本、图像和科学多模态大数据融合   
小世界生成数据回路



# Foundation models for astronomy

- Large Language Model(ChatGPT, LLaMA)
  - Chatbot for education
  - Research assistant (papers, proposals)
  - Citizen science (astronomical objects labeling)
  - Autonomous agents(“find pulsar signals in dataset abc-203”)
  - Time series data processing(light curve)
- AIGC(GAN, Diffusion Models)
  - Simulations: generate galaxy photos for public science; generate FAST observing data for verifications?
  - Faster and cheaper cosmological simulations of galaxy formation
- Few-shot learning(SAM, I-JEPA)
  - FRB/PSR search
  - HI Source Finding
  - Foreground Removal



# 大模型通用能力的拓展

- **大模型与外部信息交互：**
  - **模型之间的交互：** 不同的大模型专注不同领域的的能力，通过模型之间的交互可以实现所有能力的整合，从而拓展大模型的通用能力
  - **模型与工具的交互：** 检索、数学计算等任务无需通过大模型实现，若能让大模型学会使用工具，则可以极大地拓展大模型的通用能力
  - **模型与环境交互：** 大模型通过学会与虚拟环境或者物理环境进行交互，实现大模型从感知到行为的通用能力



# 科研助手 vs. 原始创新

# What's the next?

虚拟现实

区块链

元宇宙

人工智能

.....



China-VO

China-VO

Who is the final winner?

IT giants, Scientists, Nvidia, ...?

Thank You!