GOPS 2022
shenzhen

MLOps模型交付标准解读

胡慧　中国信通院云大所工程师

个人简介

胡慧

信通院云大所人工智能部工程师

主要研究领域涵盖人工智政策、标准、产业及生态研究，近年来重点关注人工智能工程化，人工智能关键技术与应用评测重点实验室人工智能工程化委员会人工智能研发运营小组成员，参与编制系列标准《人工智能研发运营一体化（Model/MLOps）能力成熟度模型》，参与大模型技术和应用课题的相关研究。

目录 Contents

/01

整体介绍

Model/MLOps的价值

Model/MLOps缓解AI落地3大痛点，实现高效率、高质量、可持续、低风险的模型生产。

过程

小作坊式生产

AI资产管理散乱无序

过程无规范、无标准化

工具链碎片化

数据来源分散

模型算法“黑盒”

团队

问题难追责

行业知识短板

AI团队各自为营

跨团队协作壁垒多

业务

模型一次性交付，不可持续

业务交付偏离需求，返工率高

模型部署上线耗时且失败率高

过程无度量、模型质量无保障

AI生产管理3大痛点

过程难管理

团队缺协作

交付不持续

01 高效构建

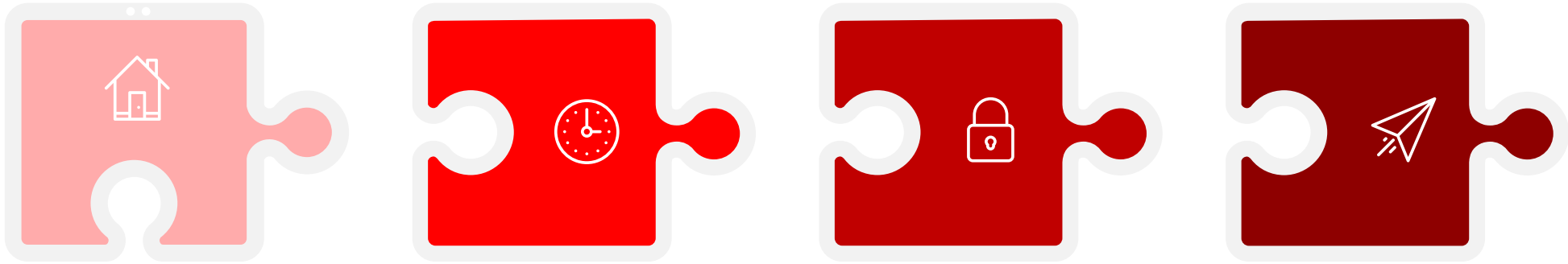
02 持续运转

03 模型治理

Model/MLOps体系

本标准体系的价值

本标准体系为企业提供标准化、可落地的Model/MLOps体系建设指南、定位标尺、改进方向，帮助企业建立和不断完善AI生产管理体系，智能化转型升级快人一步。



规范体系

指引企业Model/MLOps体系的有序发展、规范行业自律。

引领建设

树立行业标杆，为企业建立Model/MLOps管理体系提供实践性指南。

定位标尺

通过能力成熟度层级划分，为企业自我定位提供参考和依据。

全面提升

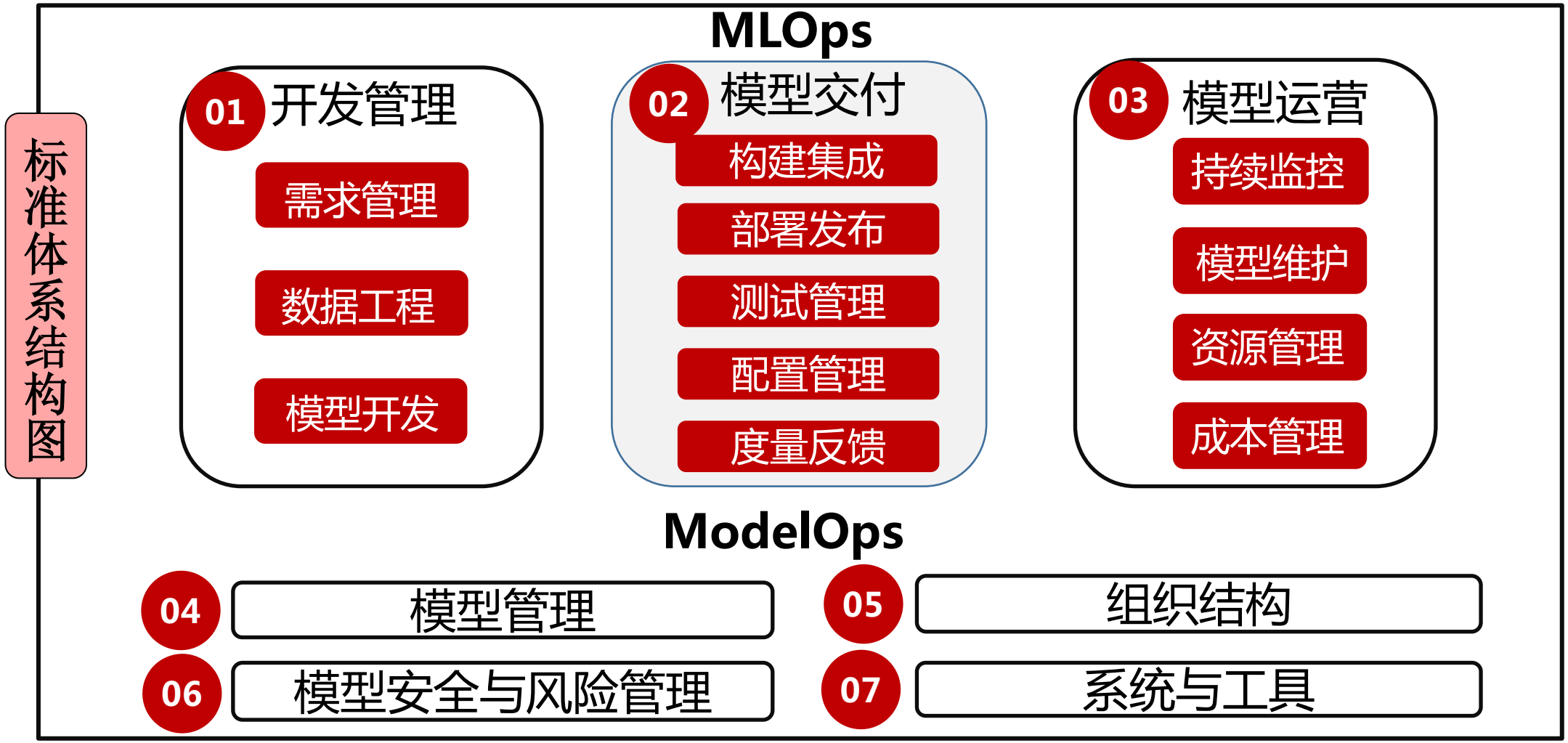
助力企业打造全方位、贯穿模型全生命周期的Model/MLOps体系，加速智能化全面升级。

/02

标准解读

标准体系结构

本系列标准包括了MLOps的过程管理和ModelOps模型治理等内容，共分为7个标准。



开发管理标准（已发布）

GOPS 2022 shenzhen

开发管理标准聚焦AI开发过程管理，助力流程自动化、体系标准化、资产版本化。

01 高效构建

规模化、一体化、标准化的**过程管理**

可通用、复用的模型生产**流水线**

02 持续运转

数字资产的**版本管理**、价值共享

CI、CD、CT、CM

03 模型治理

模型管理　组织结构

合规审计　安全与风险

模型交付标准（征求意见稿）

GOPS 2022
shenzhen

自2022.5启动编制，经过编写、讨论、评审等多轮环节，现发布征求意见稿。

能力域	Model/MLOps能力成熟度模型 第二部分：模型交付											
能力子域	构建集成		部署发布			模型测试		配置管理			度量反馈	
能力项	构建管理	持续集成	模型编排	持续部署	发布管理	测试执行	测试管理	版本管理	变更管理	环境管理	度量指标	质量驱动改进
能力子项	构建过程	模型转换与优化	场景构建	模型管理	流量管理	模型评估	测试数据管理	版本控制	变更过程	环境类型与资源管理	度量指标定义	度量分析
	构建配置	集成过程	模型配置	模型部署	效果评估	测试分层与设计	测试结果管理	分支管理	变更追溯	环境依赖与配置	度量数据管理	反馈改进
	构建计划	集成反馈	集成发布	更新策略管理	模型下线	自动化测试		依赖与模型管理				
				运行监控								

模型交付标准（征求意见稿）

模型交付标准聚焦持续交付，打造CI/CD/CT流水线，敏捷交付、快速上线、持续迭代。

1.构建集成
- 构建管理
- 持续集成

2.部署与发布
- 模型编排
- 持续部署
- 发布管理

3.模型测试
- 测试执行
- 测试管理

4.配置管理
- 版本管理
- 变更管理
- 环境管理

5.度量与反馈
- 度量指标
- 质量驱动改进

01 5个能力子域　02 12个能力项　03 32个能力子项　04 300+条分级要求

01 高效构建
规模化、一体化、标准化的**过程管理**

可通用、复用的模型生产**流水线**

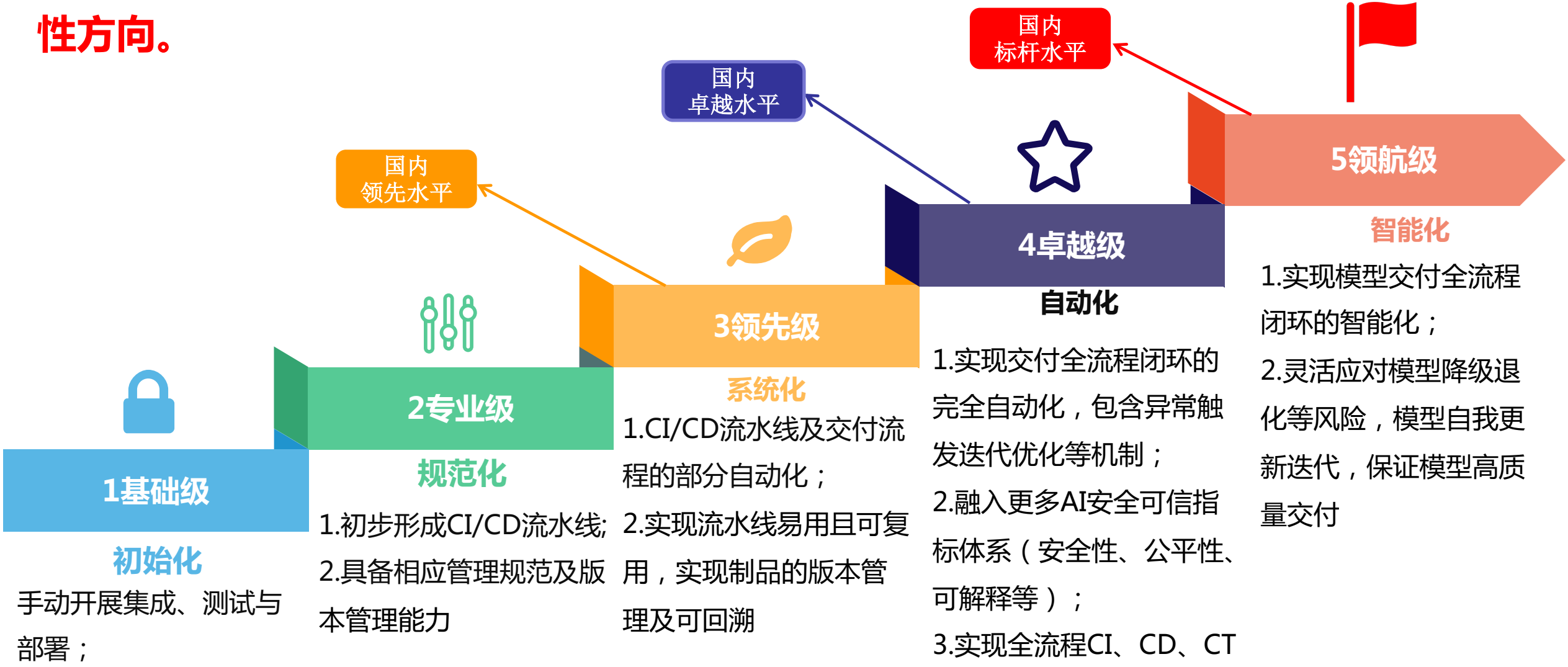
02 持续运转
数字资产的**版本管理**、价值共享

CI、CD、CT、CM

03 模型治理
模型管理　组织结构　合规审计　安全与风险

模型交付标准 能力成熟度划分

以能力成熟度作为衡量标准，帮助企业对标各级别参考水平，为企业发展提供前瞻性方向。



模型交付标准 内容体系

按能力子域-能力项-能力子项的层级，梳理各项定义内涵和能力成熟度分级具体要求。

以部署发布能力子域为例：

5类 能力子域

12项 能力项

30项 能力子项

部署发布		
模型编排	**持续部署**	**发布管理**
场景构建	模型管理	流量管理
模型配置	模型部署	效果评估
	更新策略管理	模型下线
	模型监控	

2 部署发布

部署发布是指将模型和服务代码进行部署和发布的过程。部署发布包括**模型编排、持续部署、发布管理**等3个能力项。

2.2 持续部署

持续部署是将模型包部署至运行环境，并通过更新策略将新版本模型进行持续部署的过程，同时还关注模型运行质量情况，从而对模型进行持续监控与治理。持续部署包括模型部署、部署策略管理、模型监控等3个能力子项。

2.2.1 模型管理

模型管理是指对即将发布的模型进行统一的存储、版本管理和访问控制的过程，以便积累和共享模型资产。

2.2.2 模型部署

模型部署是将模型部署至运行环境，对外提供推理服务的过程。模型运行环境包括云侧服务器、边缘设备和端侧设备等。

2.2.3 更新策略管理

更新策略管理是指对新版本模型设置更新策略，用新版本的产生触发重新部署的过程。

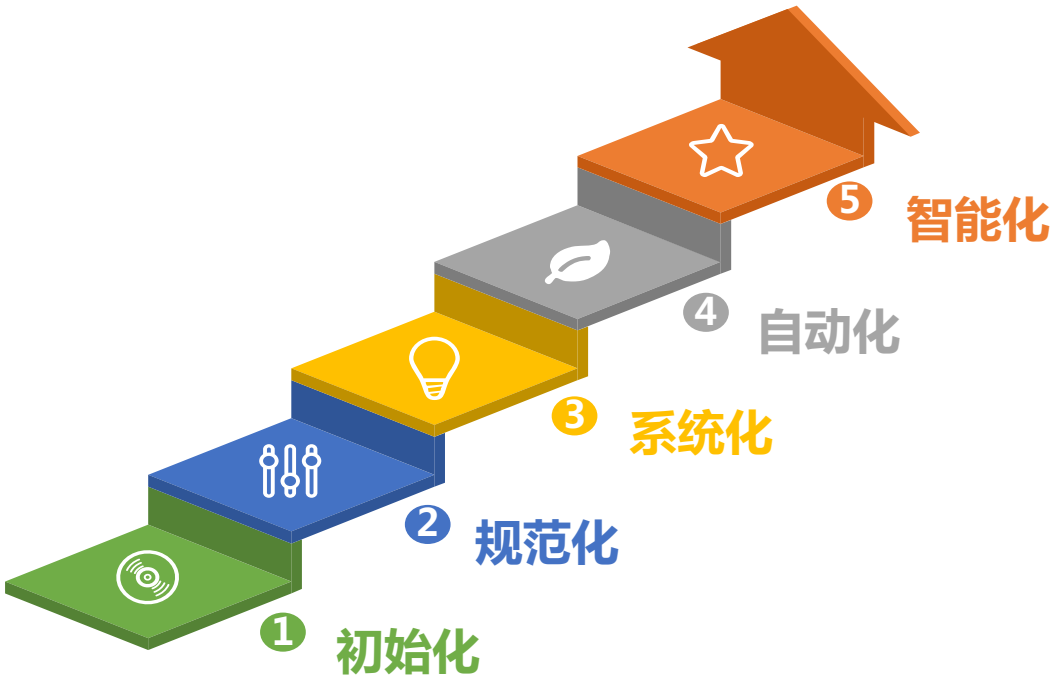
……

模型交付标准 编写维度

编写维度参照能力成熟度分级，从初始化、散乱无序的小作坊模式向系统化、自动化、智能化的工厂流水线模式进化跃迁。

以部署发布能力子域为例：

部署发布是指将模型和预测服务进行部署和发布的过程。

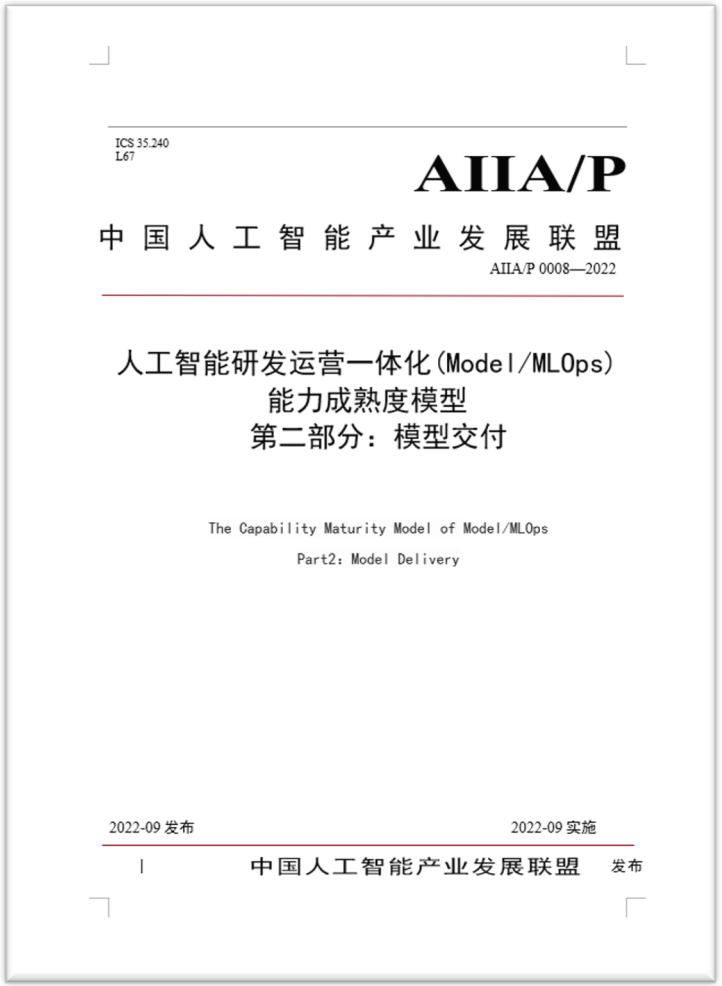


部署发布能力成熟度编写维度			
等级	模型编排	持续部署	发布管理
1级	以人工方式为主初步开展部署与发布工作。		
2级	规范化开展模型部署、更新策略管理、运行监控、流量管理等各项工作；成果物支持版本化管理，初步形成**CD流水线**。		
3级	系统化、全面化开展部署发布；部分自动化执行工作**CD流水线**；具备更新策略执行过程的可追溯和可审核能力;具备流量管理与分配的可视化能力。		
4级	对部署发布过程持续优化；支持**CI、CD流水线**的无缝串接，全自动化执行**CD流水线**；支持高频的模型更新能力。		
5级	模型部署、模型更新策略、模型流量调整的智能化分析、问题反馈和调优机制。		

模型交付标准 参编单位

GOPS 2022
shenzhen

行业30余家单位参与编制，包括金融企业、互联网头部企业、运营商、主要创新企业等。



ICS 35.240
L67

AIIA/P

中 国 人 工 智 能 产 业 发 展 联 盟

AIIA/P 0008—2022

人工智能研发运营一体化(Model/MLOps)
能力成熟度模型
第二部分：模型交付

The Capability Maturity Model of Model/MLOps
Part2: Model Delivery

2022-09 发布　　2022-09 实施

中国人工智能产业发展联盟　发布

CAICT 中国信通院
HUAWEI
Baidu 百度
华佑科技 HUAYOU TECHNOLOGY
中国移动 China Mobile
China unicom 中国联通 让一切自由联通
浦发银行 SPD BANK
中原银行 ZHONGYUAN BANK
中国农业银行 AGRICULTURAL BANK OF CHINA
招商银行 CHINA MERCHANTS BANK
马上消费 WWW.MSXF.COM
蚂蚁集团 ANT GROUP
建信金科 CCB Fintech
恒丰银行 HENGFENG BANK
九章云极 DataCanvas
inspur 浪潮
Testin云测
快手
Alibaba 阿里巴巴
北京一流科技有限公司 ONEFLOW TECHNOLOGY LIMITED
商汤 sensetime
DataExa 一站式AI赋能平台 — 渊亭科技 —
泰豪 TELLHOW
中电金信 GienTech
Starwhale
thoughtworks
MEGVII 旷视
网思科技 创新服务 智慧中国
JDT 京东科技
万达信息
国家电网 STATE GRID 国网电子商务有限公司(国网金融科技集团)

/03

评估测试

谁需要这套评测体系？

这套评测体系帮助各类企业客观定位AI生产过程管理能力，以评促建、以评促改，辅助快速查漏补缺，提升能力和水平，实现降本增效，树立行业标杆，引领产业前进发展。

类型5

AI项目全自动化生产，但离智能化还有一段距离。

类型4

AI项目生产未实现全自动化，模型量化指标不成体系，无异常自动处理机制，尚未实现全流程CI、CD、CT、CM。

类型3

AI项目具备标准化、规范化的管理体系，但自动化水平低下，流水线不可复用，实验不可复现。

类型2

AI项目部署上线周期长，管理混乱，模型交付不可持续 ，MLOps落地刚起步。

类型1

数智化刚起步，尝试AI赋能业务，MLOps规划中，业务、算法、IT团队协作困难。

建议以**5级**为目标。
实现全流程的智能化，灵活应对模型降级退化等风险，模型自我更新迭代，保证模型高质量交付。
领航级

咨询&评测

建议以**4级**为目标。
实现全流程的完全自动化，包含异常触发迭代优化、自动纠偏纠错等机制，并融入更多AI安全可信指标体系（安全性、公平性、可解释等），实现全流程CI、CD、CT、CM。
卓越级

咨询&评测

建议以**3级**为目标。
完善现有的模型生产全流程以达到部分自动化，实现流水线易用且可复用，实现制品的版本管理及可回溯。
领先级

咨询&评测

建议以**2级**为目标。
在组织内建立AI模型全生命周期的规范、制品的版本管理，从小作坊生产逐步向工厂流水线过渡。
专业级

咨询&评测

建议以**1级**为目标。
搭建内部AI开发生产管理体系，手动实现全流程、跨团队的串联，避免后续走弯路。
基础级

咨询

2022下半年工作计划安排

GOPS 2022 shenzhen

2022.7
开发管理**首轮评测**结果公布，证书颁发

2022.8
可信AI峰会：
1.模型交付标准征求意见稿首发
2.MLOps产业观察

2022.9
1.模型交付标准**正式发布**
2.MLOps应用实践蓝皮书编写启动
（MLOps实践优秀案例征集）

2022.10
开发管理**第2轮评测**报名

2022.11
模型交付**首轮评测**报名

2022.12
MLOps应用实践蓝皮书正式发布

中国人工智能产业发展联盟
Artificial Intelligence Industry Alliance

中国通信标准化协会
China Communications Standards Association

联盟评估规范

行业标准

模型交付标准立项成功，后续正在推进中。

标准的地位

Thanks

开放运维联盟
高效运维社区 **荣誉出品**
DevOps 时代

联系人：
胡老师 17371328072（微信同号） huhui@ caict.ac.cn
秦老师 13488684897（微信同号） qinsisi@caict.ac.cn

想第一时间看到高效运维社区的新动态吗？

高维在线