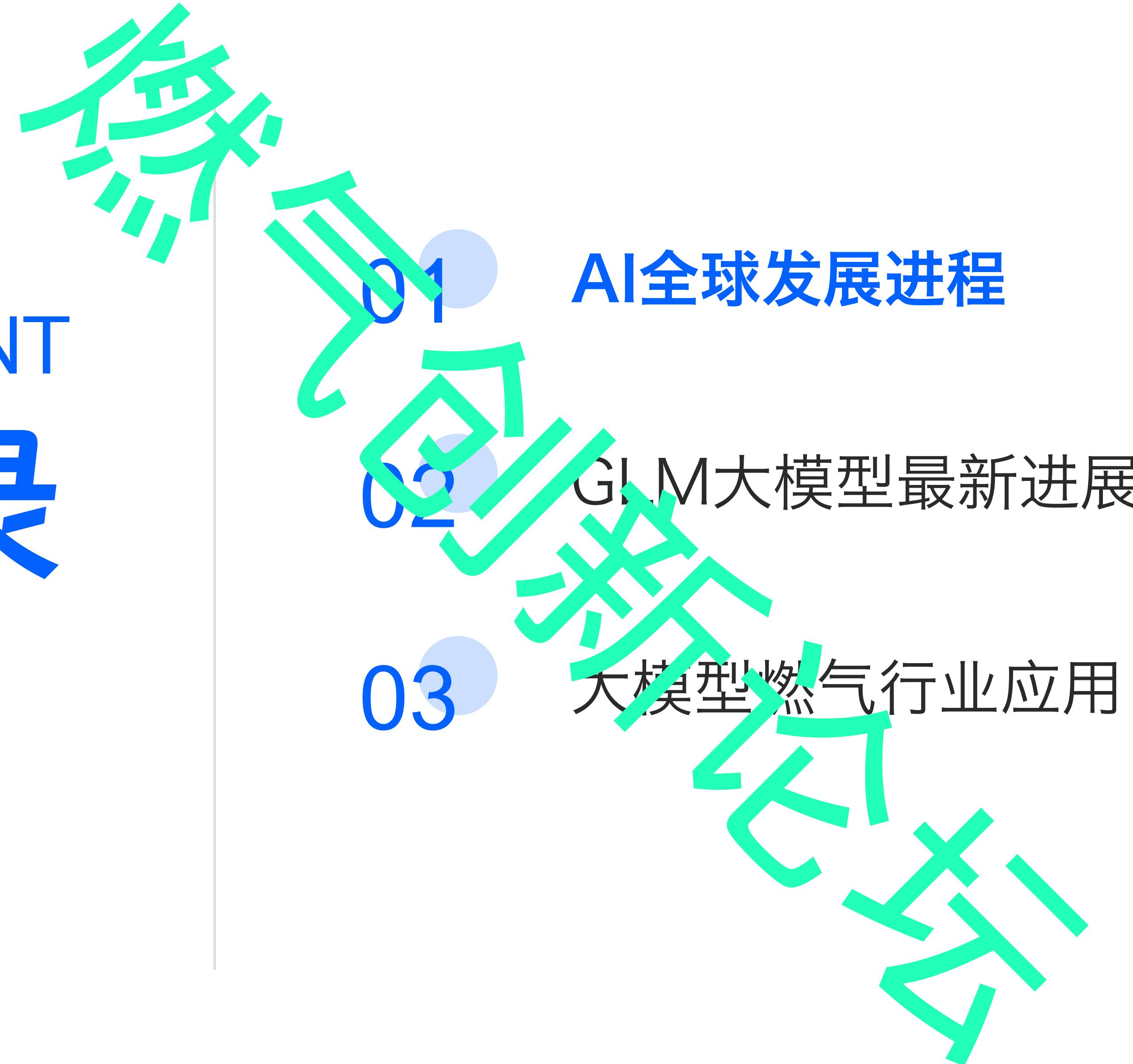


大模型：开启燃气行业创新引擎

张帆 / 智谱 AI

智谱

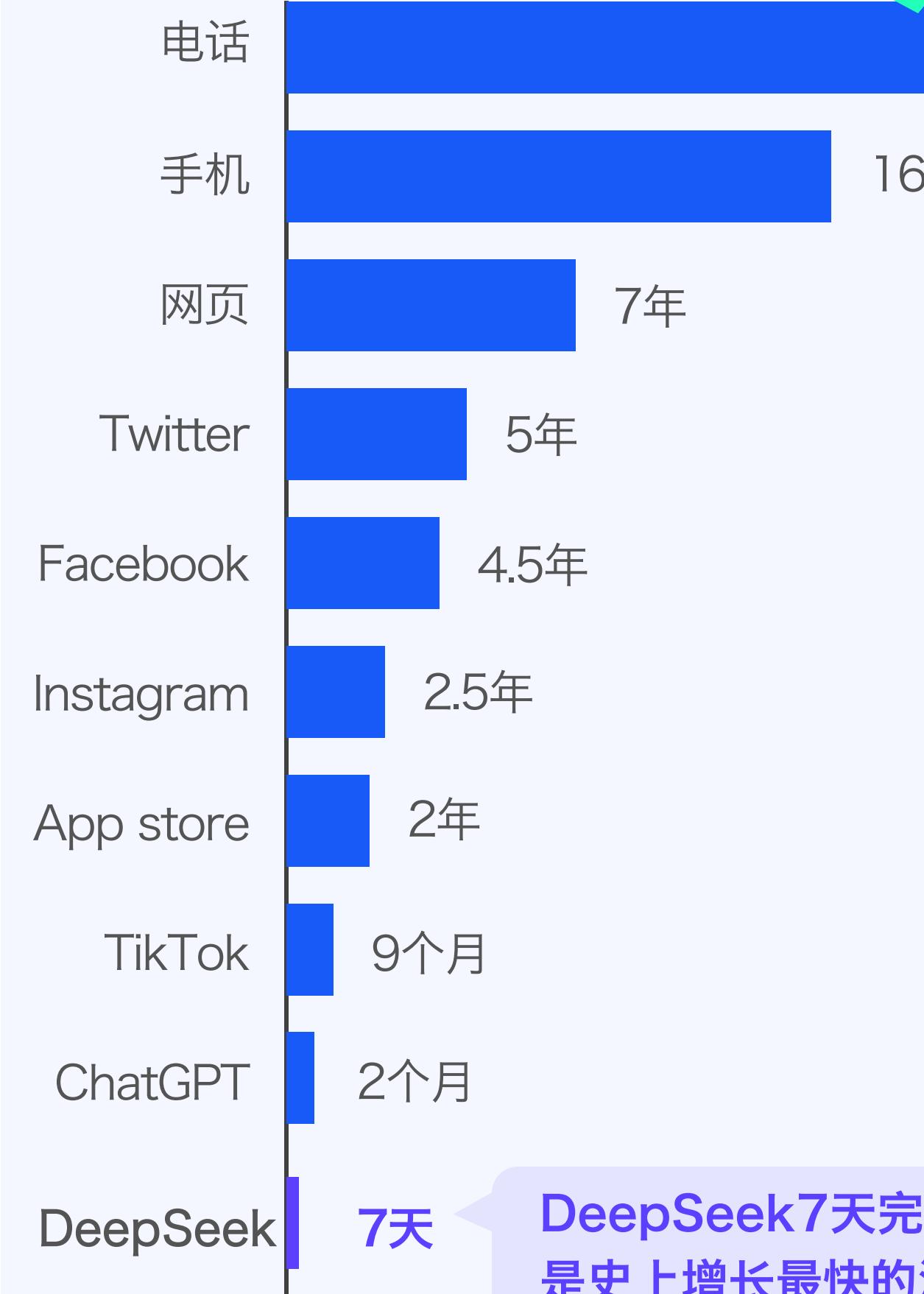
CONTENT
目 录



生成式AI的发展速度远超科技史上任何变革性技术

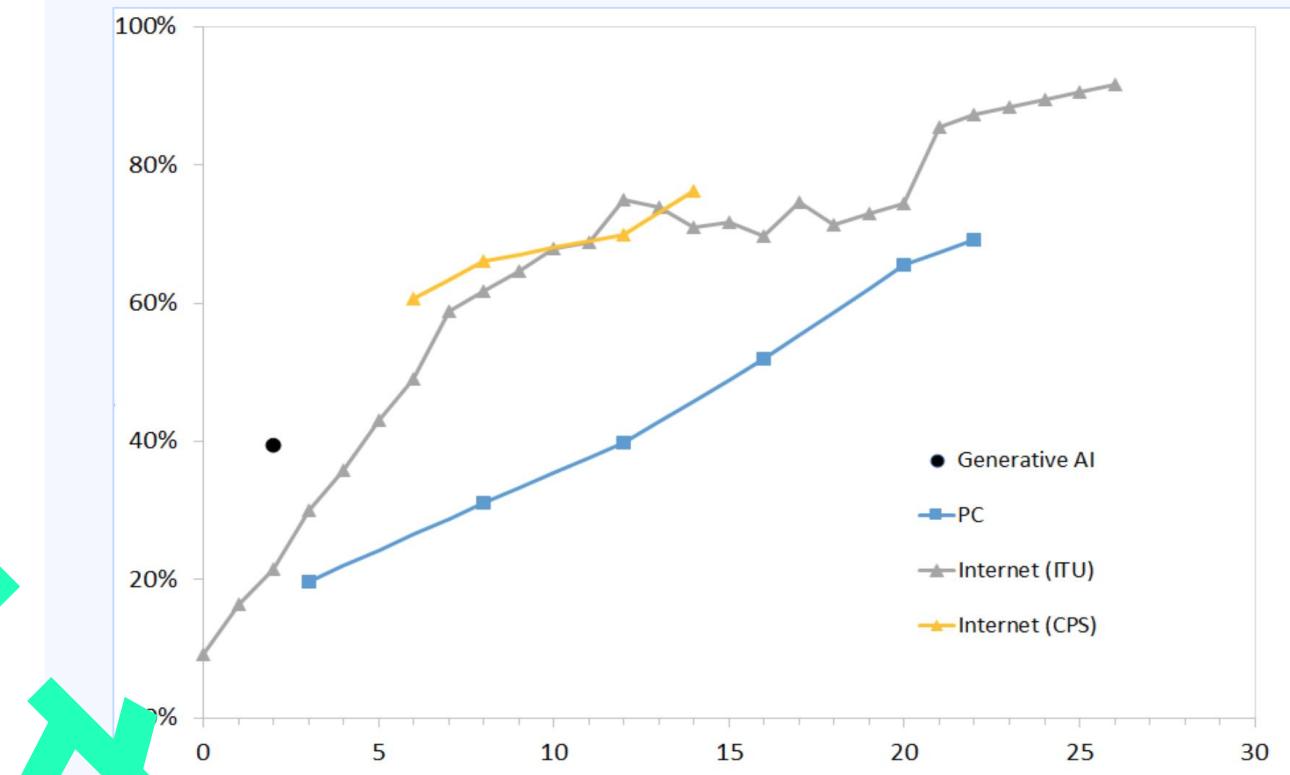


全球各类消费级应用增长1亿用户所用时间



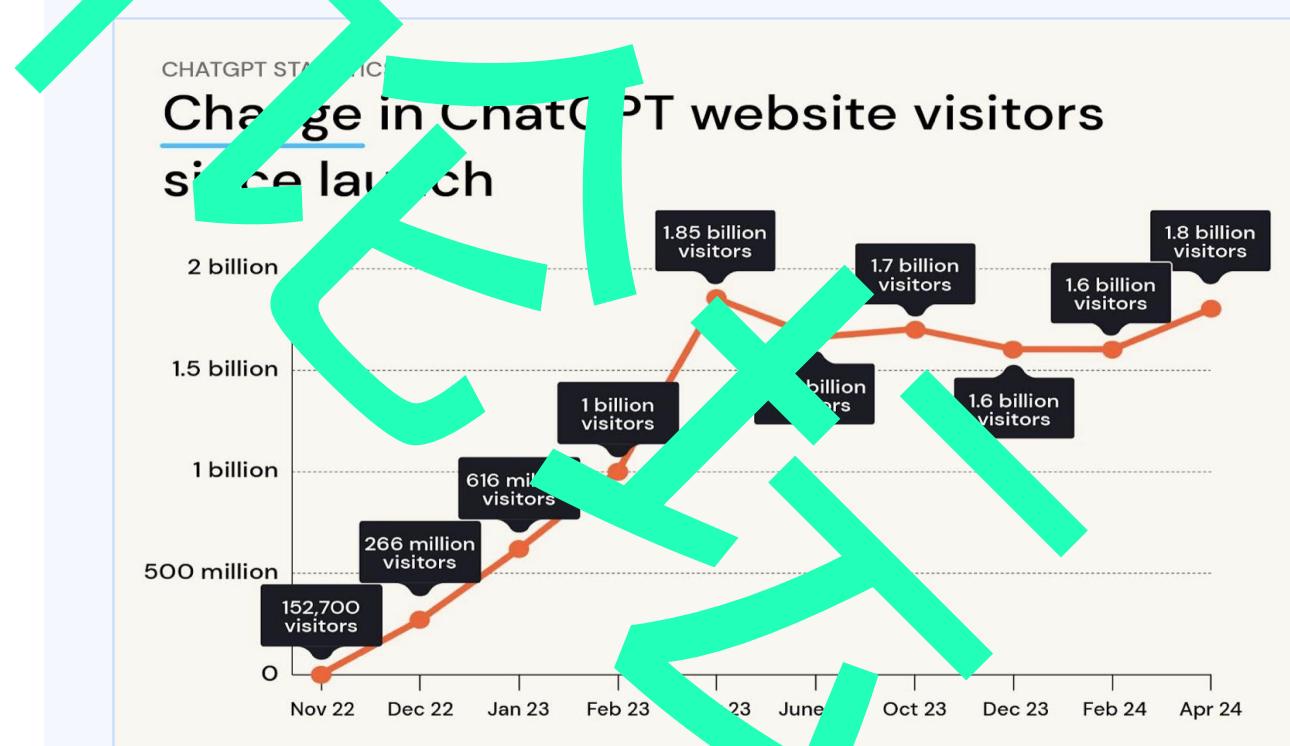
DeepSeek 7天完成了1亿用户的增长，
是史上增长最快的消费级应用

生成式AI进入大众市场的速度远超PC和互联网



- 个人电脑在进入大众市场3年后的采用率为**20%**
- 互联网的大众采纳率在**2年后**达到**20%**
- 生成式AI的采用率在**2年内**达到近**40%**，远超个人电脑和互联网

生成式AI应用的用户规模迎来了爆发式增长



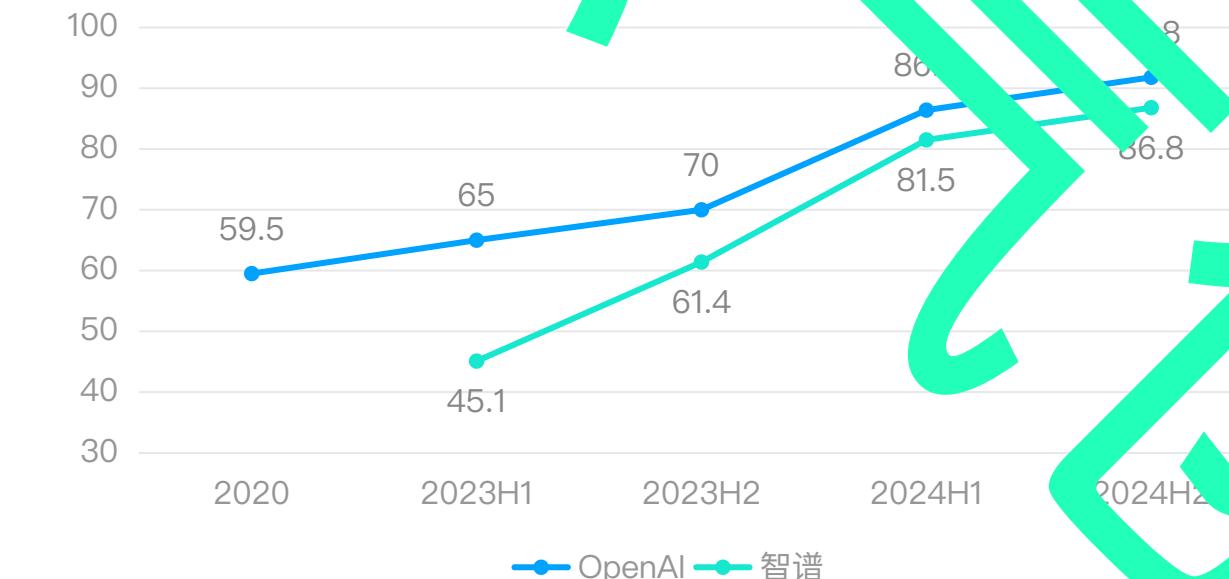
- ChatGPT的用户规模从**15万**(2022年)飙升至**1.8亿**(2024年)，**2年内翻1000倍**
- 在美国，**53%**用户已使用过ChatGPT，**24%**的用户已使用过Google Gemini

大模型在自然语言理解、模态丰富度、复杂推理等多维度能力显著提升 智谱

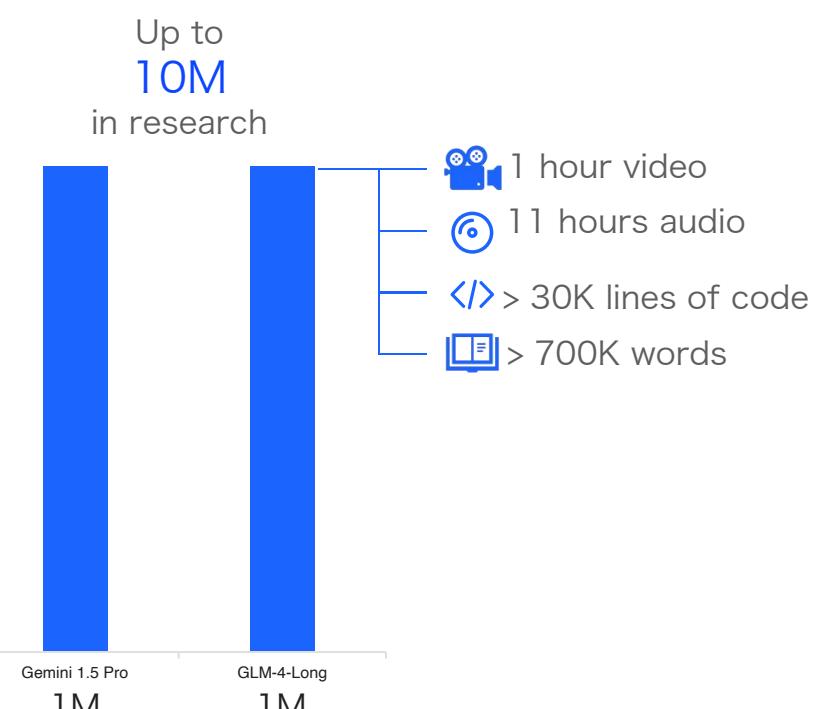
大规模多任务语言理解能力已接近人类专家水平

- OpenAI的大模型在不到4年内，通用能力提升至人类专家水平90%以上(o1)
- 智谱大模型通用能力评测在1年内从45提高到86，达到o1的95%

OpenAI和智谱大模型MMLU评测



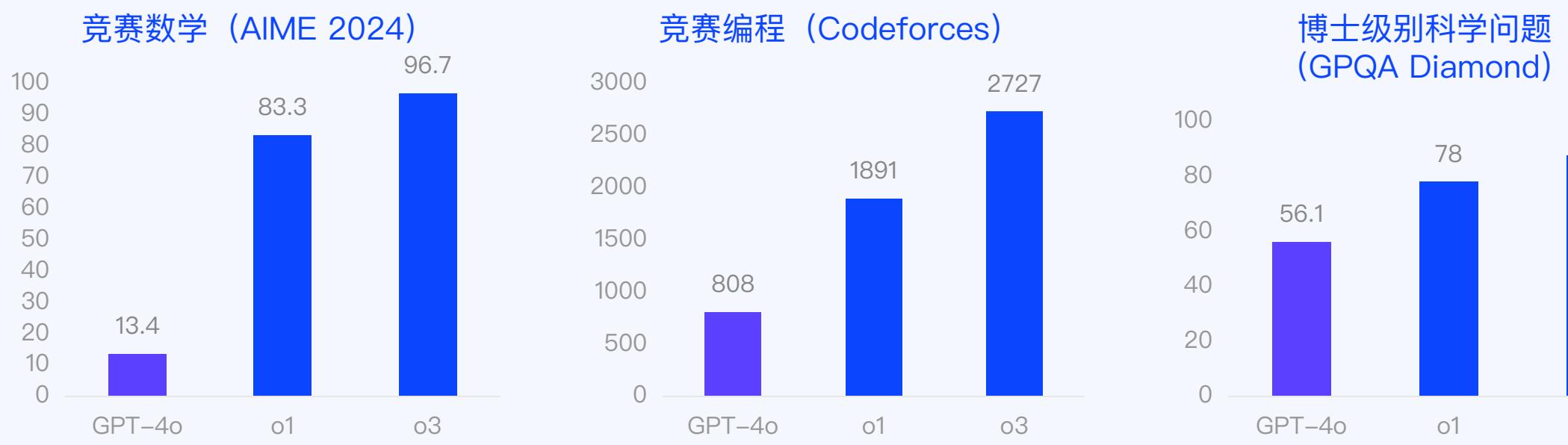
上下文长度不断突破，记忆力增强



Gemini 1.5 Pro支持100万token上下文

智谱GLM-4-Long支持高达1M（约150-200万字）上下文长度

复杂推理能力超越大部分人类水平



从GPT-4o到o3，大模型在数学、代码、专业科研等场景的复杂推理能力，在半年左右实现质的飞跃，超越大部分人类水平。

单一模态到多模态，从文本理解到物理世界互动



多模态理解模型
支持语音及视频通话，能够感知人类情绪状态



DeepSeek现象级爆发



DeepSeek-V3发布时间：2024年12月26日

DeepSeek-R1发布时间：2025年1月20日

通过一系列算法和工程创新，大幅降低了模型训练和推理成本，并在基座模型和推理模型的性能上追赶上国际一流水平。

技术上大幅缩减了开源模型和闭源模型的差距，加速技术迭代；应用上有助于推动AI应用的爆发，加速了大模型的渗透。

成本
低

DeepSeek-V3

成本大降，动摇大模型“狂卷算力模式”

MoE架构创新

与传统MoE架构不同，DeepSeekMoE使用了更细粒度的专家，并将部分专家设为共享专家，使模型能够高效地处理复杂任务。

MLA机制

创新MLA（多头潜在注意力）机制，显著降低了推理过程中的KV缓存，相比标准MHA减少93.5%。从而降低内存消耗，并保持性能。

FP8混合精度训练

原生支持FP8混合精度训练，与传统的FP16或FP32相比，显著减少了GPU内存的使用和计算开销，从而降低训练成本。

Training Costs	Pre-Training	Context Extension	Post-training	Total
in H800 GPU Hours	2664K	119K	5K	2838K
in USD	\$5.328M	\$0.238M	\$0.01M	\$5.576M

Table 1 | Training costs of DeepSeek-V3, assuming the rental price of H800 is \$2 per GPU hour.

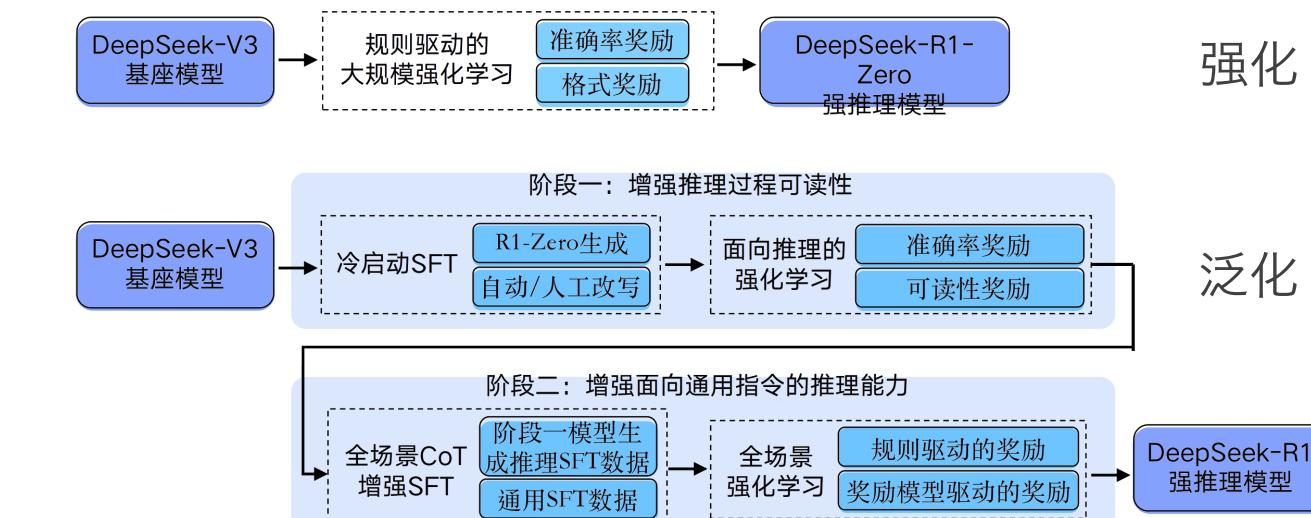
成本：DeepSeek-V3的训练费用为557.6万美元，远低于行业平均水平。

性能：DeepSeek-V3在性能上比肩GPT-4o以及Claude-3.5-Sonnet不分伯仲。

DeepSeek-R1

「强化学习」驱动模型推理能力提升，性能比肩OpenAI o1 正式版

DeepSeek-R1 训练流程



首次验证了纯强化学习方法增强模型推理能力的可行性。无需大量标注数据意味着更低的训练成本。

	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759.0
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717.0
o1-mini	63.6	80.0	90.0	60.0	53.8	1820.0
QwQ-32B	44.0	60.0	90.6	54.5	41.9	1316.0
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954.0
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189.0
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481.0
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691.0
DeepSeek-R1-Distill-Llama-14B	50.4	80.0	89.1	49.0	39.6	1205.0
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633.0

证实了知识蒸馏能够高效低成本地提升小模型推理能力。

趋势一：推理模型的出现为大模型带来更强大的推理能力

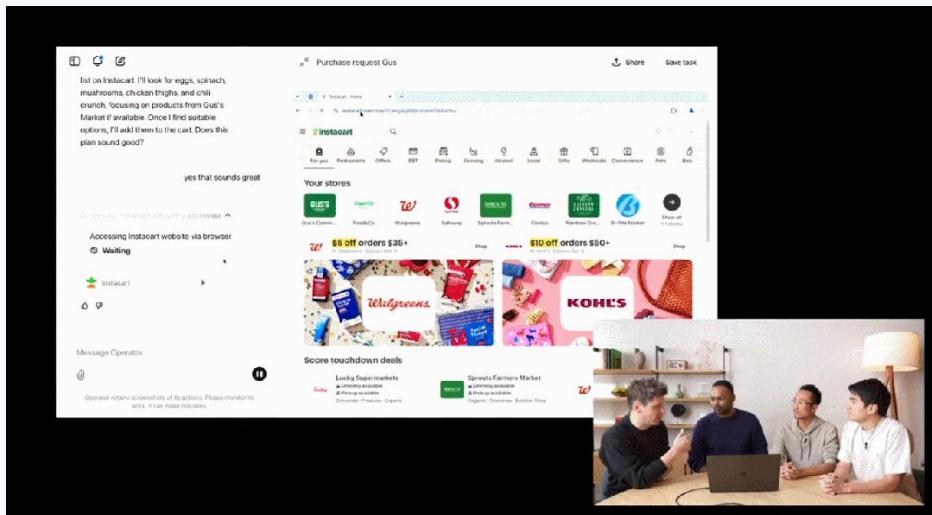


趋势二：智能体迎来大规模爆发

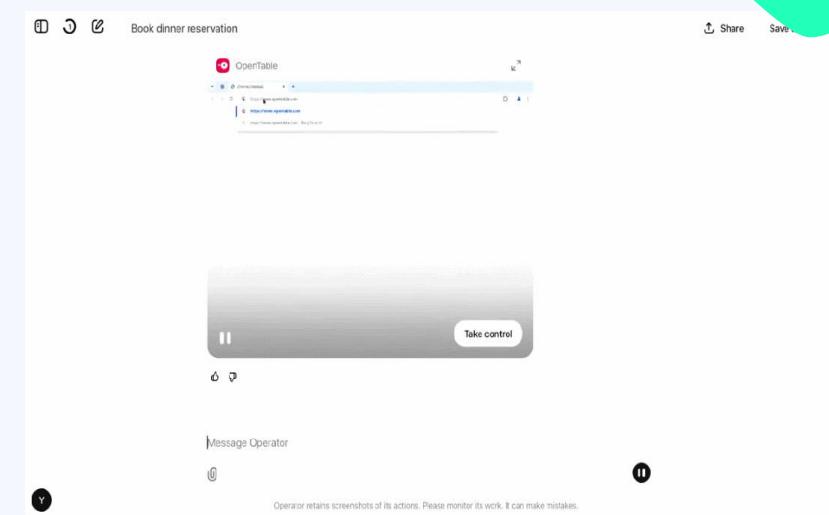


OpenAI

2025年1月23日，OpenAI正式发布其首款 AI智能体 Operator，能够像人类一样通过视觉理解和操作计算机界面，使用鼠标、键盘执行任务。



示例1：预定晚餐



示例2：购物平台下单

“Operator 是我们进入 Level 3 智能体的开端。它不仅标志着 OpenAI 技术方向的飞跃，也是我们实现 AGI 愿景的重要一步。”

——Sam Altman, OpenAI CEO

AI Anthropic

Anthropic计划在2025年推出一款名为“Virtual Collaborator”的AI智能体

办公室的“全能助手”

自主运行在电脑上

编写、编译代码

验证工作成果

通过Slack
和同事沟通

在Google
Docs上协作

定期向用户
汇报进度

“我们将在未来几年内接近实现功能强大、甚至可能达到人类水平的AI系统。虽然还存在不确定性，但是从过去6个月的发展来看，这种不确定性已经大大降低。”

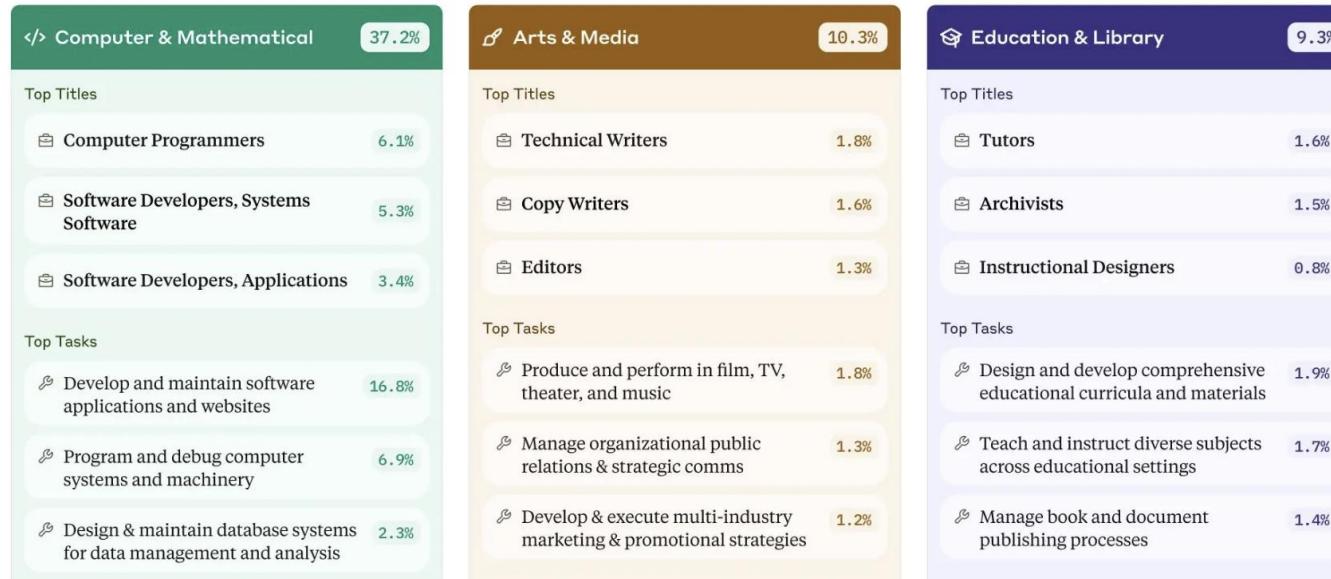
——Dario Amodei, Anthropic CEO

趋势三：编程是AI应用最多的领域，IT相关职位将迎来革新

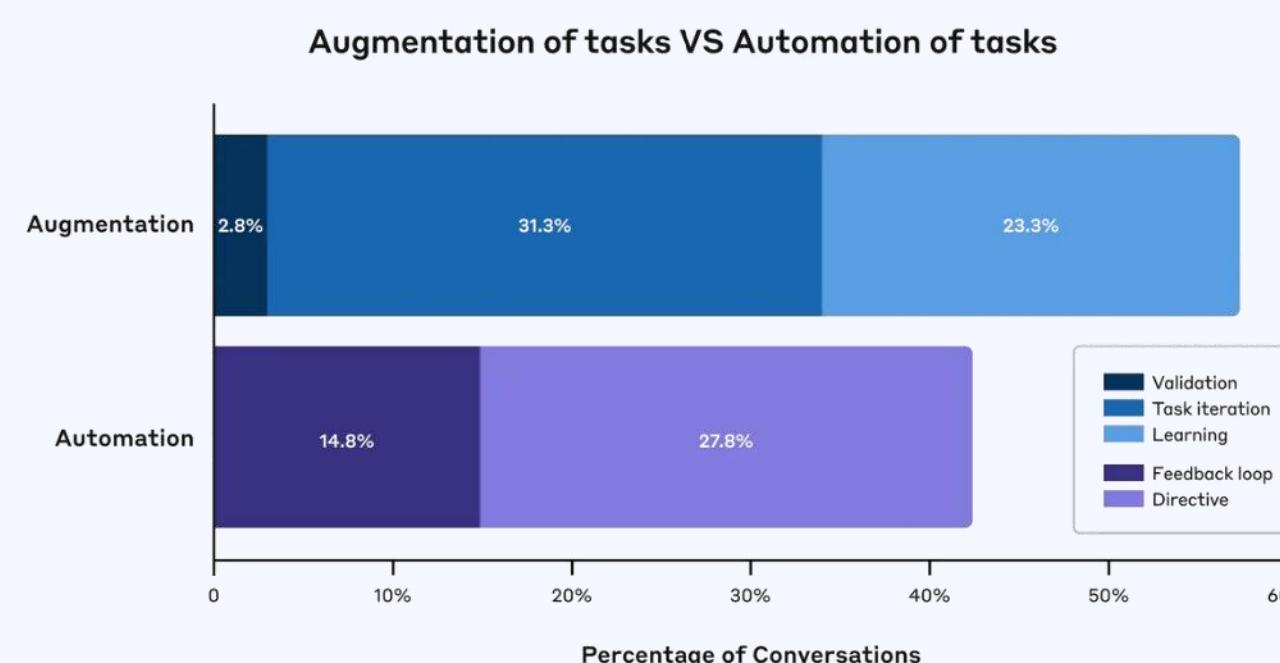
在一项Anthropic的研究中显示



AI使用最多的领域是“计算机和数学”相关，占比37.2%，主要是改代码、修bug、搞网络故障等场景。



AI的增强模式是在“扩饭碗”：57%的任务是“增强型”(augmentation)，比如AI帮你验证工作、学习新技能、头脑风暴；43%的任务是“自动化”(automation)，比如让AI直接写个文档、格式化文件。



1. 将来会有更多而不是更少的工程师

“外面有很大的对工程的需求，而很多这样的问题可以用代码解决，所以我认为如果把每个开发者的能力用AI成倍增加，将来会需要更多而不是更少的开发者。对于工程而言，决定要做什么，解决什么问题永远是最核心的，而代码工具可以帮工程师去执行，让工程师更聚焦于高价值问题上。”

— Michael Truell, Anysphere CEO

2. 软件工程师角色将会改变

“五年或十年后的软件工程师的角色会看起来有点像今天的**技术架构师和产品经理的混合**。你所做的会是面对你或你的业务中的问题，并且思考然后分解解决方案应该是什么。”

— Michael Truell, Anysphere CEO

3. 编写评估标准将成为产品经理的一项核心技能

“到了2024-2025年，构建一个智能的产品经理的工作越来越像**负责产品界面的产品经理和API产品经理**。比如我们最近发布了代码分析功能，现在Cloud可以分析CSV文件并为你生成代码。产品经理可能完成了80%的工作，然后不得不把项目交给一些懂得如何编写评估标准、进行模型微调和优化提示词的产品经理。我们发现这其实就是一个角色，因为你的产品质量现在完全取决于你对评估和提示方面的水平。所以产品经理这个角色的定义正在发生融合。”

— Mike Kreiger, Anthropic CPO

“OpenAI专门设置了一个训练营，**让每个产品经理学习如何编写评估标准**，理解好的和差的评估之间的区别。当然我们还在不断完善这个过程，需要持续迭代和改进，但这确实是开发优质AI产品的关键所在。”

— Kevin Weil, OpenAI CPO

| 定义AGI目标：AI能力分级

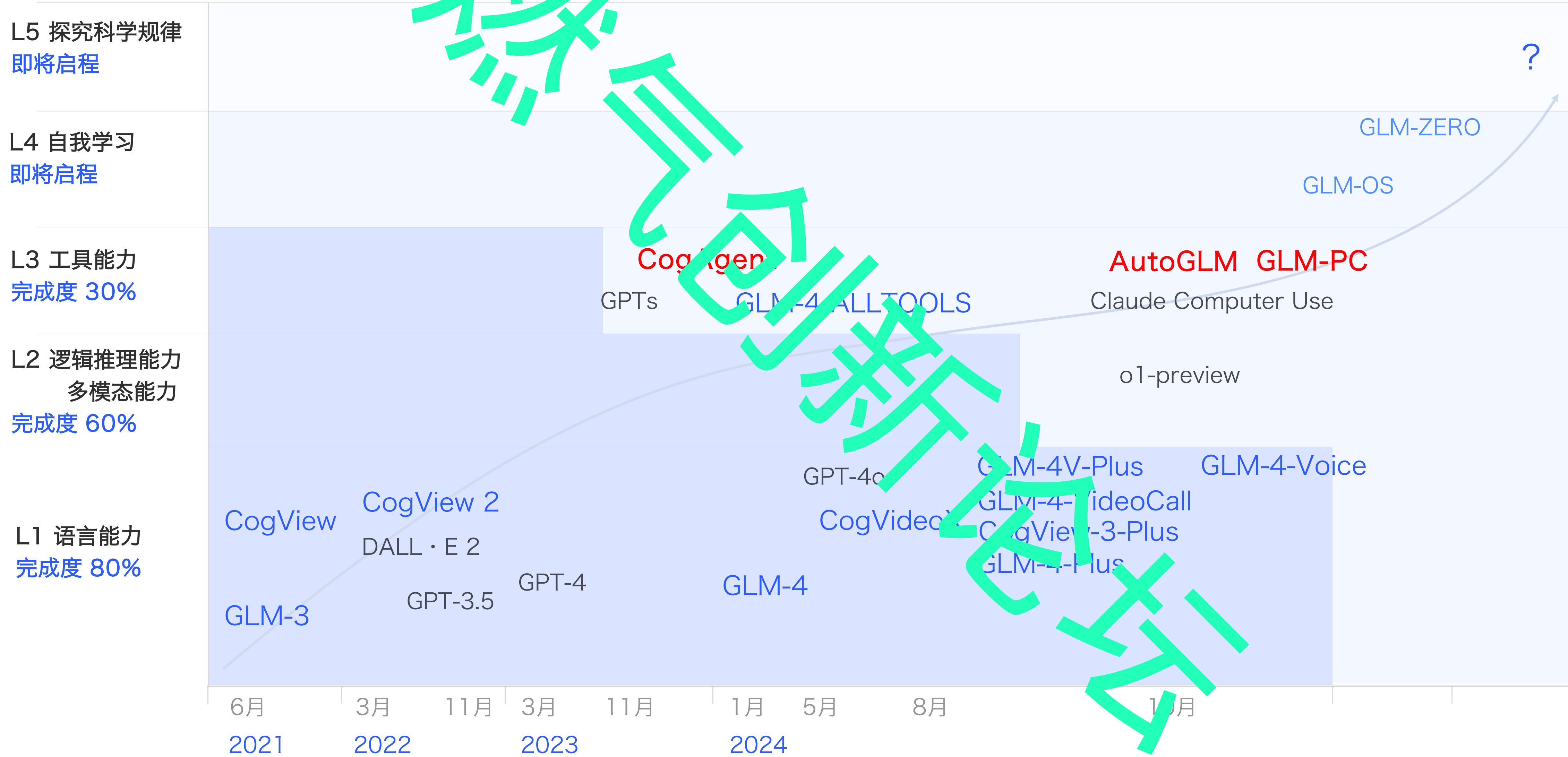


Open AI

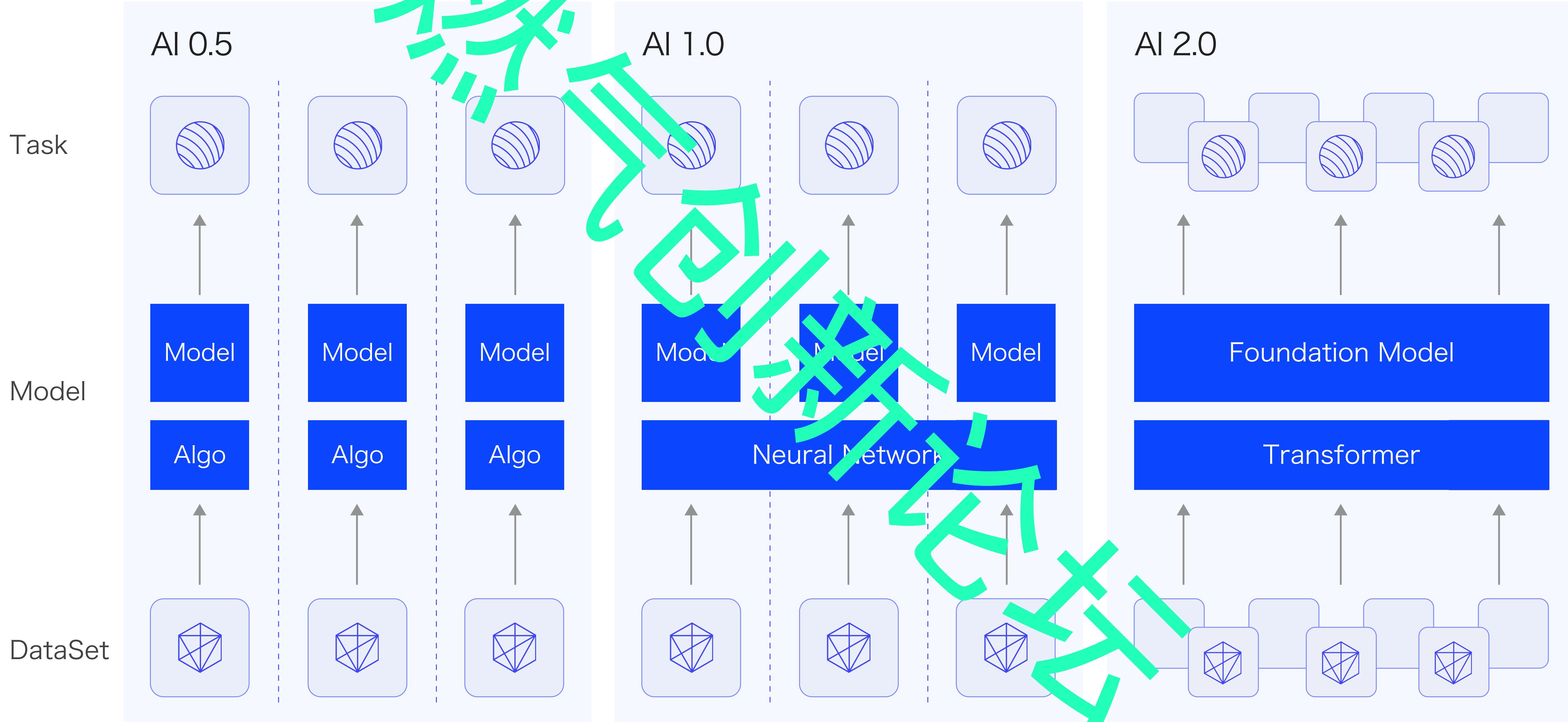
我们的思考



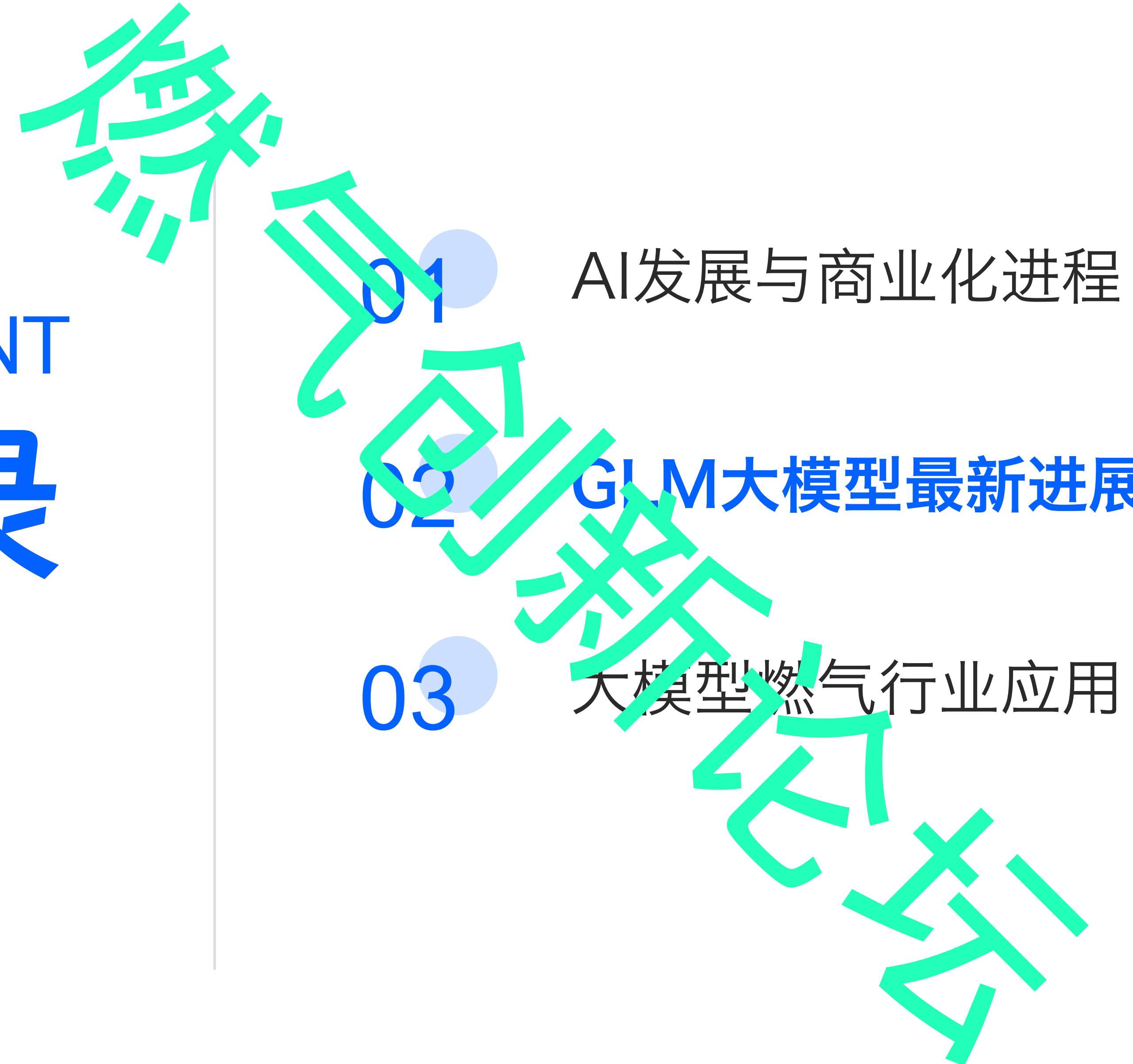
智谱的 AGI 之路：专注大模型研究，持续高速迭代



大模型的技术新范式



CONTENT
目 录



模型矩阵全面覆盖文本、代码、多模态等，横跨云侧和端侧



语言大模型

从端侧小模型到经济型模型，再到千亿参数旗舰大模型，提供丰富的语言模型矩阵，满足用户在成本、应用场景等多维度的需求；最新旗舰模型GLM-4-Plus 性能媲美国际顶尖模型

GLM-Edge-1.5B-Chat
端侧模型

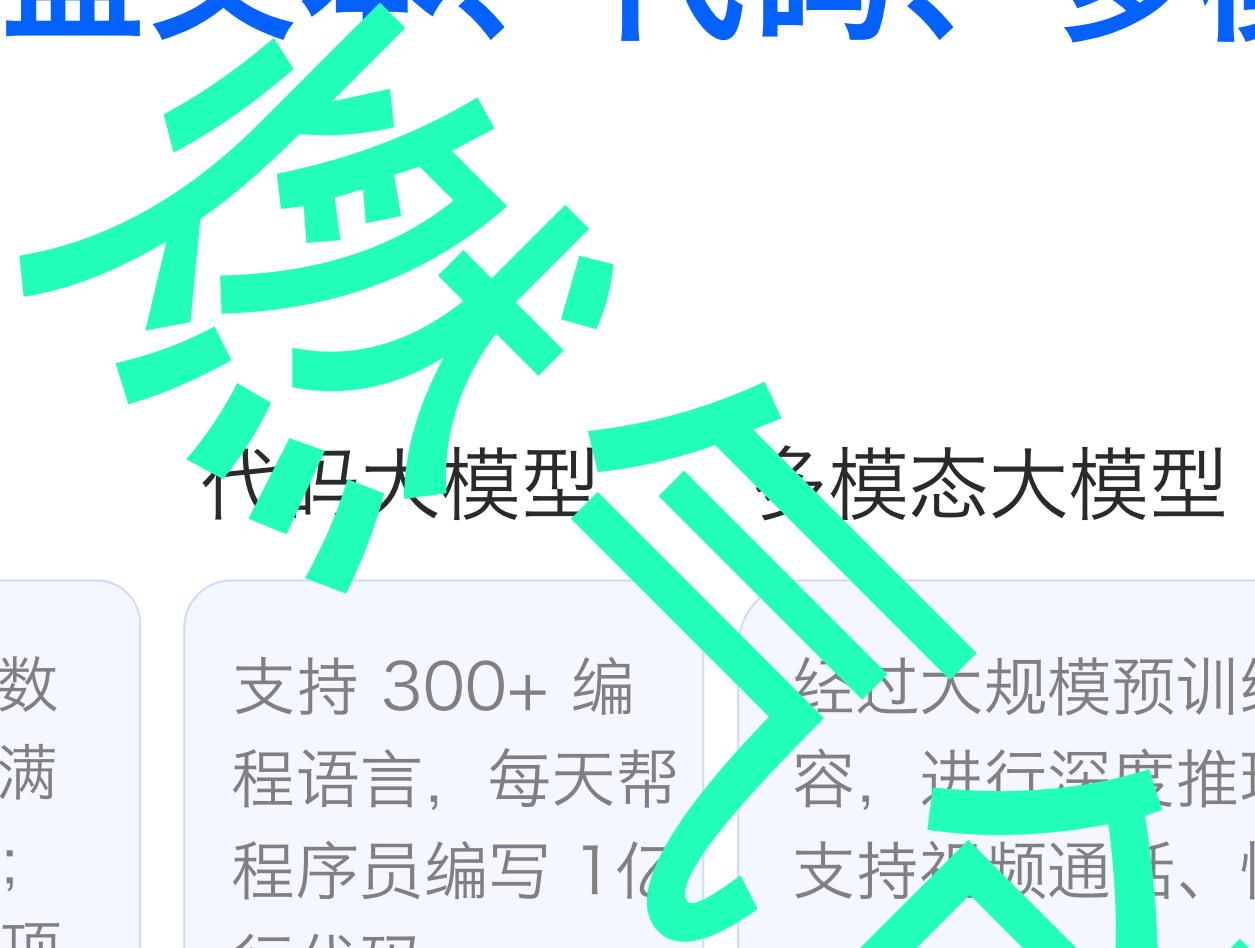
GLM-Edge-4B-Chat
端侧模型

GLM-4-9B
最新开源

GLM-4-Flash
免费调用

GLM-4-Air-0111
经济

GLM-4-Plus
最新旗舰



支持 300+ 编程语言，每天帮程序员编写 1亿行代码

经过大规模预训练，能够准确理解图片、视频等多模态内容，进行深度推理；生成高质量的图片、视频和音频素材；并支持视频通话、情感语音等跨模态实时互动

CodeGeeX4
最新升级

CodeGeeX4-ALL-9B
最新开源

GLM-Edge-V-2B
端侧多模态理解

GLM-Edge-V-5B
端侧多模态理解

GLM-4V-Flash
免费云侧多模态理解

GLM-4V-Plus
云侧多模态理解

CogVideoX-3-Flash
免费图像生成

CogVideoX-4
图像生成

CogVideoX-Flash
免费视频生成

CogVideoX-2
视频生成

CogSound
音效生成

GLM-Realtime
实时视频

GLM-Zero Preview
推理模型

GLM-4-Video
情感语音

具有立体化人格，支持情感共情、超长记忆、基于人设的角色扮演

CharGLM-4
角色扮演

Emohaa
心理陪伴

像人类一样感知环境、规划动作、并在手机和电脑上执行复杂任务

CogAgent
GUI Agent

AutoGLM
手机/网页操作助手

GLM-PC
电脑操作助手

多模态生成模型

文生视频

输入：

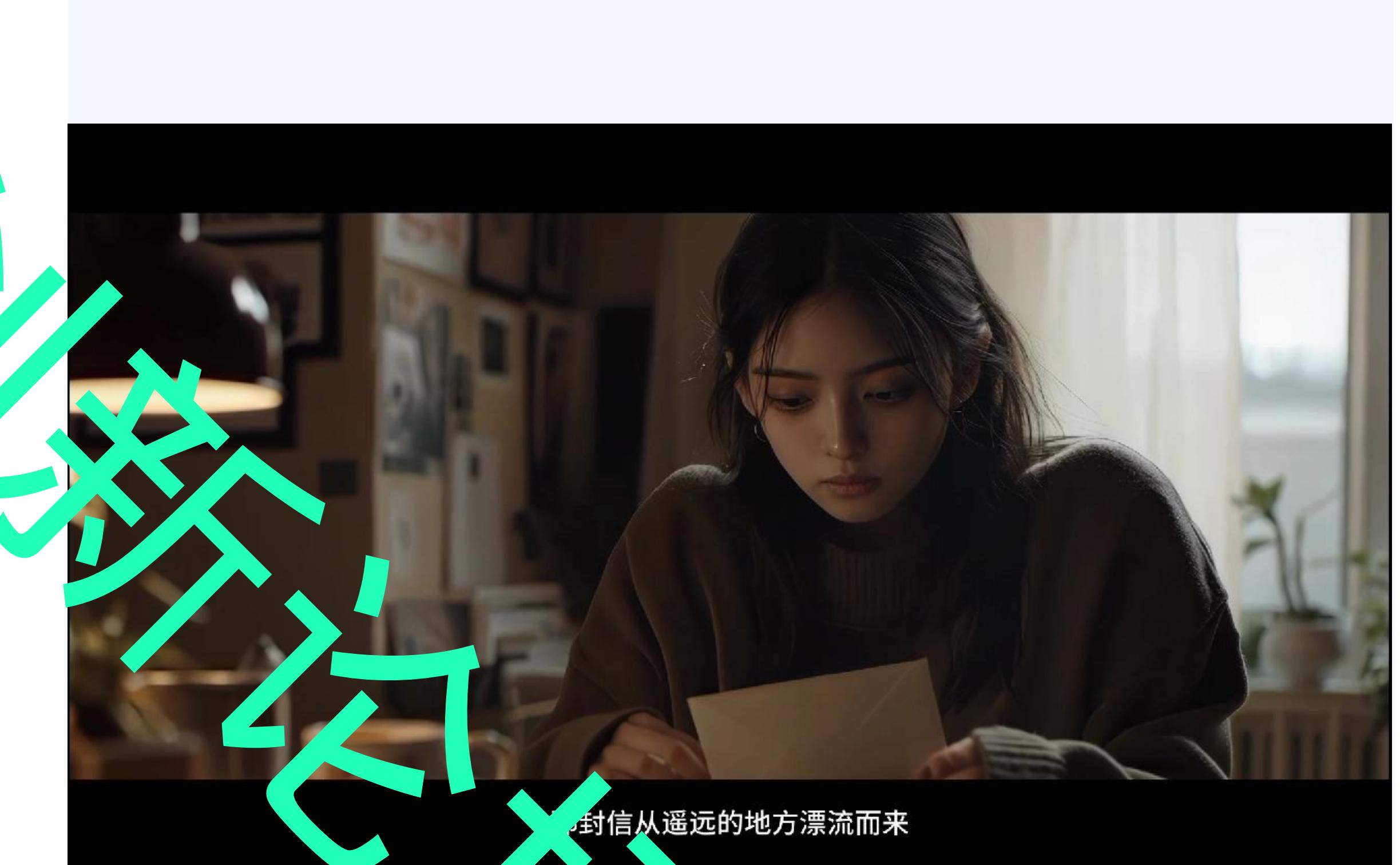
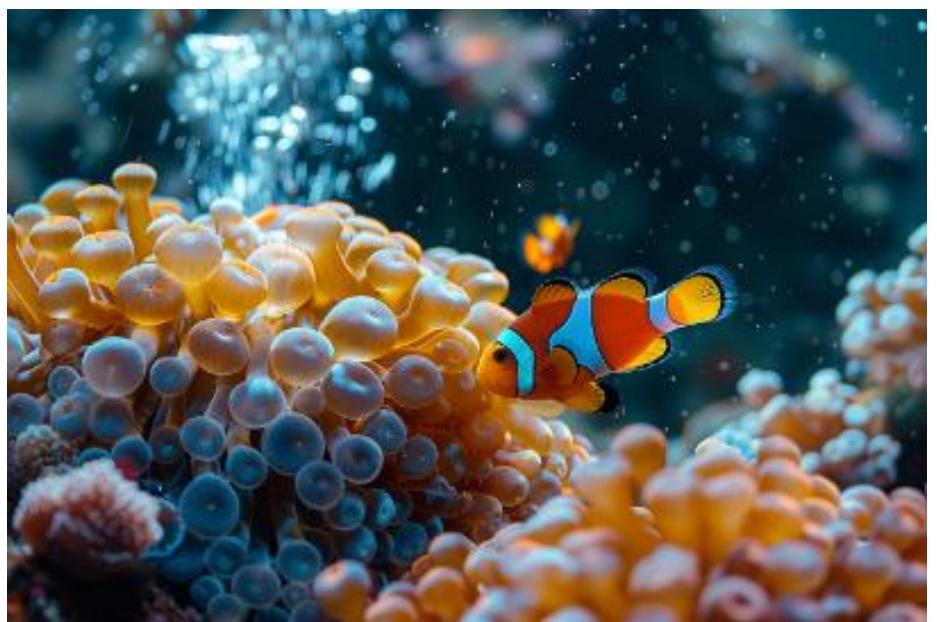
动画场景，展示一个粉色的毛绒绒的小怪物划船，3d风格，需要注重画面的细节，小怪物的神情充满喜悦，表现出顽皮天真的样子。温暖的颜色和氛围的灯光。



视频生成模型应用

图生视频

输入：小丑鱼围绕着珊瑚游荡，水里面的气泡向上浮动，水下摇晃的镜头



一封来自远方的信漂流而来

GLM-4V-Plus：国内首个通用视频理解模型 API，再次升级



具备视频总结 + 视频修改 + 推理能力 + 多轮对话 + 时间问答能力，能够理解并分析复杂的视频内容，同时具备时间感知能力



全新升级，性能表现显著提升的同时计算成本更低



端侧模型GLM-Edge为AI原生设备而生



端侧模型小巧而强大

端侧语言模型

端侧语言模型	最高解码速度 65 tokens/s	最快首响时间 0.07s	端侧语言模型	最高解码速度 24 tokens/s	最快首响时间 0.17s
GLM-Edge-1.5B-Chat			GLM-Edge-4B-Chat		

端侧多模态理解

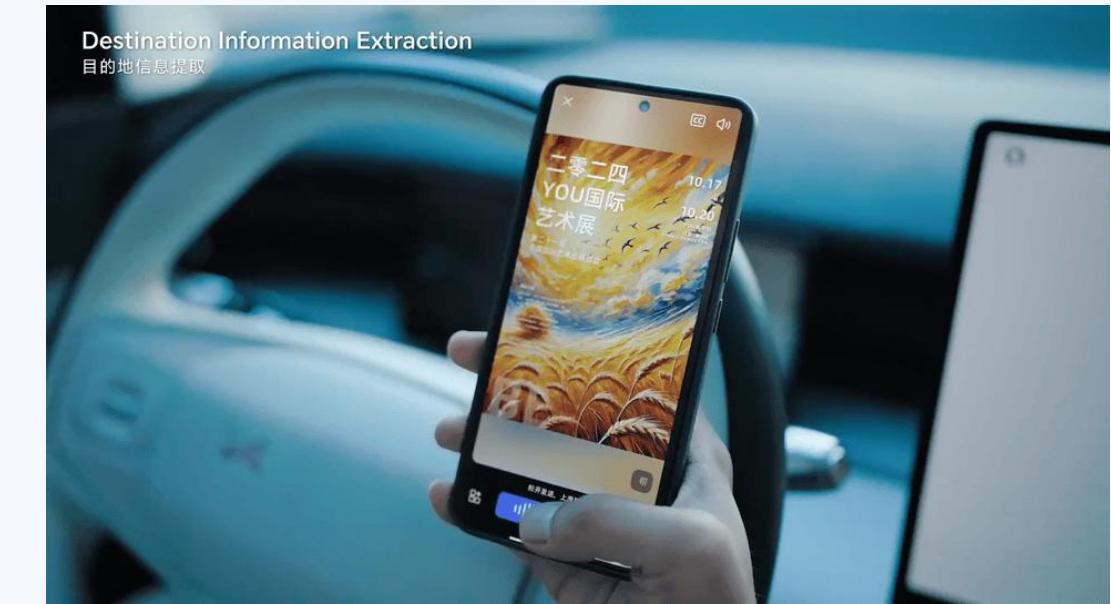
端侧多模态理解	最高解码速度 70 tokens/s	最快首响时间 1.50s	端侧多模态理解	最高解码速度 24 tokens/s	最快首响时间 2.00s
GLM-Edge-V-2B			GLM-Edge-V-5B		

端云同源的协同架构

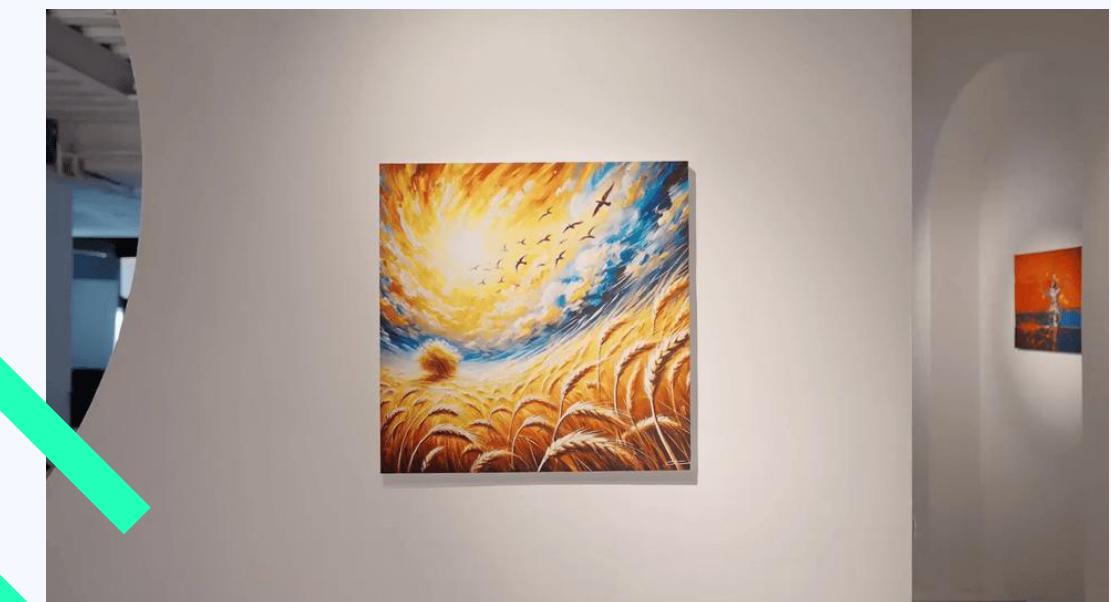


智谱 x 高通骁龙8至尊版移动平台

自动获取目的地信息并导航



画作鉴赏与推荐



GLM-4-Plus-VideoCall: 国内首个视频通话 API



GLM-4-Plus-VideoCall 具备跨文本、音频和视频等多模态内容理解和实时推理能力，支持流畅视频通话，即使用户频繁打断也能迅速反应。适合教育培训、办公、文旅、智能硬件及医疗智能等多类场景

办公场景



教育场景



生活场景



- ✓ 识别并理解企业工作环境，准确判断业务类型
- ✓ 深度解析工作内容，提供专业解答与指导

- ✓ 智能识别与理解试题，解释相关知识点
- ✓ 互动提问，实现多轮全面精准解答

- ✓ 智能识别日常用品，快速提供品牌、产品等信息
- ✓ 针对用户问题，提供清晰的分步操作指引

GLM-Realtime：首个具备清唱能力的端到端实时视频通话模型



全新发布的端到端模型GLM-Realtime实现了低延迟的视频理解与语音交互；该模型融入了清唱功能，并支持长达2分钟的记忆及Function Call功能

实时视频&语音能力

- 让用户体验到近乎实时的智能助手交互
- 在语音交互方面，创新性地实现了清唱功能，让大模型首次具备在对话中的歌唱能力



支持Function Call 功能

- 不仅能够依靠自身的知识和能力，还能灵活调用外部知识和工具



低延迟交互

近乎实时的交流体验



清唱功能

前所未有的趣味性



长记忆

更拟人，更懂你



Function call

更广阔的应用前景

自主智能体：模拟人在手机、浏览器和电脑上执行复杂任务



模型能力

AutoGLM 可以完成人类在电子设备上可以做的任何事，它不受限于简单的任务场景或 API 调用，也不需要用户手动搭建复杂繁琐的工作流。

- 触控执行
解放双手，更加便捷地与设备互动
- 内容理解
高效从大量数据中提取关键信息，提高效率
- 内容生成
创意无限，快速生成高质量内容

CogAgent 专门用于图形用户界面GUI的理解和执行，泛化能力强，可以“像人一样”自主规划和执行各种复杂跨模态任务。

Phone Use

Web Use

旅游攻略

在小红书上检索,阅读并输出旅游攻略

Computer Use

智谱新一代 MaaS 平台，助力企业轻松打造专属模型和应用



一键微调，打造私有模型

+



零代码开发智能体应用

集成了最新的全系列模型矩阵，企业可使用私有数据，在创建微调任务、训练、部署三个步骤后使用企业私有模型



全新智谱大模型微调平台



智谱

Step 1 准备训练数据

按照模板收集并准备数据，导入数据集中以训练模型

Step 2 创建微调任务

可选择 LoRA 和全参数微调方式，支持训练轮数、最大文本长度等参数配置

Step 3 部署微调模型

LoRA 微调模型可直接调用，全参微调模型将部署到私有云服务器

Step 1 模版创建

提供丰富的通用智能体模版，选择业务对应模版，一键复制使用

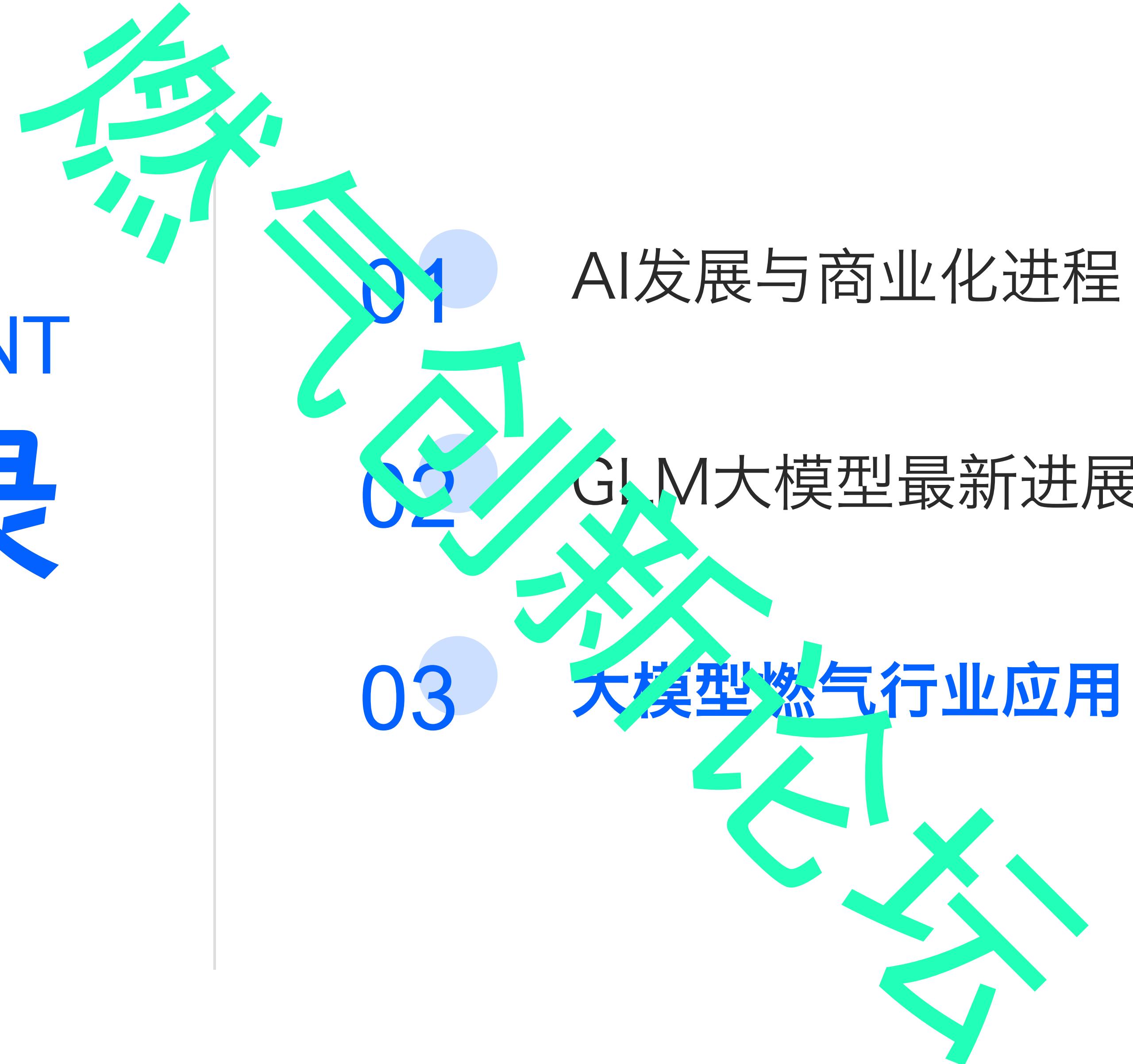
Step 2 智能生成提示词

提示词智能优化，直接使用结构化的提示词或基于优化后的提示词修改

Step 3 可视化配置

页面配置简单可控，一键生成应用引导内容

CONTENT
目 录



燃气行业大模型应用全景图



智能客服：更高效、更满意、更低成本



24小时在线快速响应用户的咨询和投诉。无论是关于燃气费用查询、业务办理流程，还是用气安全知识等问题，智能客服都能准确理解用户意图，并提供及时、准确的回答。

燃气费用查询与账单解析

- 信息提取
- 数据库关联
- 数据解析
- 图表生成

我刚收到燃气账单，显示本月费用比上个月多了50元，能帮我查一下具体用气量和单价吗

信息查询

好的，正在为您查询。根据您的户号XXXX，本月用气量为25立方米，单价为3.2元/立方米，总费用80元。上月用气量20立方米，费用64元。本月气温较低可能导致用气量增加。您可通过‘账单详情’查看每日用气曲线图

AI助手

业务办理流程咨询

- 深层意图理解
- 知识库检索
- 多模态交互

新房需要开通燃气，需要哪些材料？流程多久

燃气报装需提供房产证、身份证及开户申请表。您可通过‘在线报装’上传材料，工作人员将在24小时内预约上门勘测。开通后3个工作日内完成安装。点击链接直接跳转办理页面

AI助手

用气安全紧急咨询

- 图像理解与判别
- 工单自动派单

天然气突然变红，有焦糊味，是不是漏气了？

检测到火焰异常，可能存在燃烧不充分或燃气泄漏危险，请立即关闭阀门、开窗通风，切勿触碰电器。已为您联系最近的维修人员，15分钟内上门。请保持电话畅通

工单生成，推送维修人员

AI助手

投诉处理与情绪安抚

- 情绪感知
- 动态决策支持
- 工单调取
- 调度系统联动

你们维修师傅迟到两小时！耽误我工作，必须给个说法！

工单调取

非常抱歉给您带来不便！已查询到您的工单号XXXX，原定14:00上门，因突发抢修任务延误。我们已优先安排另一位师傅预计16:10到达，并补偿您50元燃气券。您看这样可以吗？

行吧，尽快解决

感谢您的理解！补偿券已发放至账户，可随时查看。后续将优化调度系统，避免类似问题

AI助手

经营管理：燃气知识管理



基于大模型的燃气知识管理解决方案，帮助企业有效沉淀企业知识资产，并更好地用知识赋能一线员工，转化为生产力。

场景挑战

- 燃气企业大量高价值的经验和知识分散在各类文档或系统中，无法及时反馈生产环节，企业知识资产缺乏有效沉淀，知识存而不用
- 燃气业人才流动率提升，“老带新”的知识传承和培训方式成本高、周期长，且企业经验知识随人员流失而流失，无法形成资产
- 传统知识库建设运营成本高，用户使用体验差，不利于知识价值释放

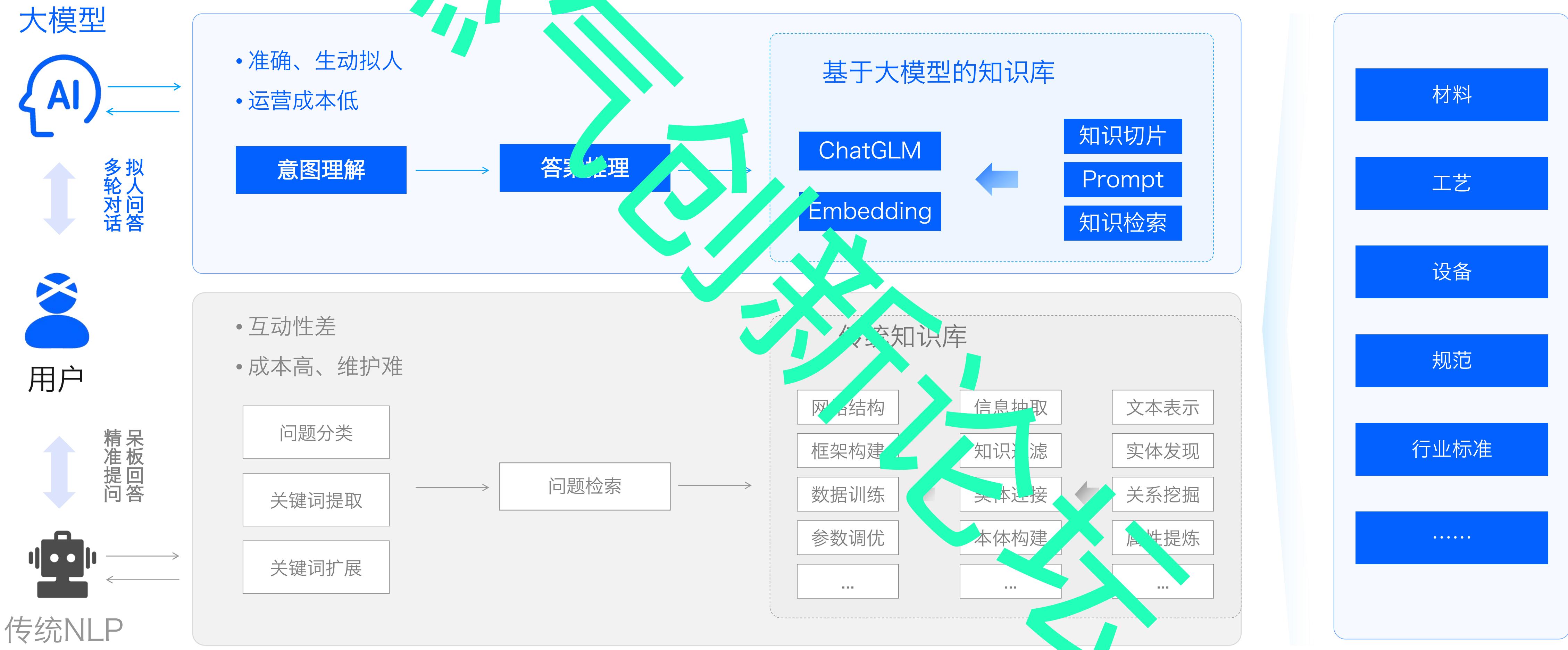


业务价值

- 有效沉淀企业知识资产，并转化为生产力
- 加速新员工培训周期，提升员工各环节工作效率
- 释放资深经验员工的精力，投入到更高价值的工作中

知识管理示例：知识问答

对比基于传统NLP的解决方案，基于大模型的知识问答在用户体验和运营成本上都具备显著优势。



设备运维助手



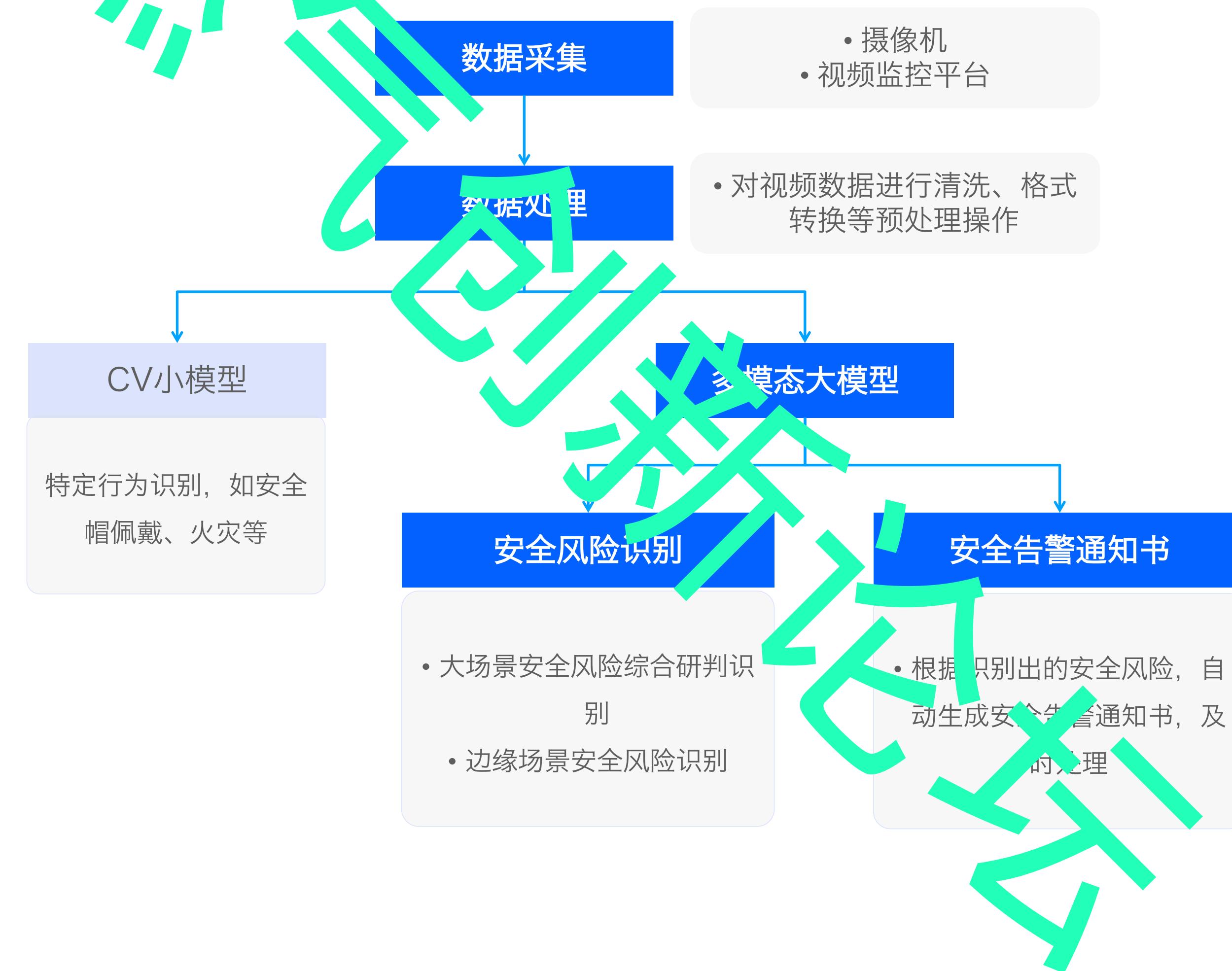
- 设备维修SOP生成：基于设备知识库，结合大模型的语义理解和生成能力，维修工程师能够通过自然语言交互的方式获取设备故障维修SOP建议。
- 设备预测性养护报告生成：大模型可以根据设备运行数据自动生成预测性养护报告，及时监测设备运行状态、输出保养建议。



多模态大模型可以和传统CV小模型结合，对好地理等安全监控画面进行综合研判，识别小模型不易发现的安全隐患和违规操作行为，并自动生成告警通知。

场景挑战

- 传统安全管理依赖人工巡检和监控，效率低下，难以及时发现和处理违规行为和安全隐患
- 企业安全管理常常面临人手不足的问题，导致无法对所有关键区域和时段进行充分监管
- 传统 CV 小模型泛化能力弱，训练成本高，难以识别 Corner Case，易出现漏报



业务价值

- 降低安全风险：及时发现安全隐患，及时通知安全管理人员进行处置，从而降低安全风险
- 降低监管成本：大模型自动识别和告警有效减少对人力资源的依赖，且模型泛化能力强，降低监管成本

大模型时代，企业如何构建自己的竞争力？

智谱



让机器像人一样思考



thanks

联系我们：400-6883-991

北京市海淀区中关村东路1号院9号楼10层

No.1 ZHONGGUANCUN EAST ROAD, HAIDIAN DISTRICT, BEIJING

版权所有© 北京智谱华章科技有限公司 2024