AI 技术每日分析：Google 推出 Gemini Decisions API：一次处理最多 1000 个独立问题

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 11 月 12 日

摘要

11 月 11 日的 AI 消息，一端是输入输出清晰的判断服务，另一端是科研模型与长任务工具。Google 上线 Gemini Decisions API，Plurai 开放智能体仿真与评估平台；DeepMind 发布 AlphaGenome 2，阿里云和百度分别公开 Qwen4-Max-Preview 与 ERNIE 5.5 Flash Preview。这些消息的开放范围并不相同：已经上线的 API、限时预览和计划开放的权重，需要分开看。

Contents

1. Google 推出 Gemini Decisions API：一次计算完成上千项判断 2
2. DeepMind 发布 AlphaGenome 2：把 DNA 变异效应细化到细胞类型 2
3. 初创 Plurai 获 1300 万美元融资并开放智能体仿真平台 3
4. 阿里云开放 Qwen4-Max-Preview：多模态旗舰限时预览 3
5. 百度发布 ERNIE 5.5 Flash Preview：2800 亿总参数、140 亿激活 4
工程与产业延伸 4
趋势观察 5
参考资料 5

1. Google 推出 Gemini Decisions API：一次计算完成上千项判断

Google 在 11 月 11 日发布 Gemini Decisions API，底层是 Gemini Decision 1 模型，明确服务工具选择、模型路由、内容评分以及策略检查等有边界的判断任务。官方介绍，开发者可以把最高 256K 词元的多模态输入与最多 1000 个独立问题交给一个接口，支持文字、图像、音频、视频及 PDF。Google 给出的端到端时延为 150—250 毫秒，输入定价每百万词元 0.10 美元，无输出词元收费。上述性能和报价来自官方产品说明，并不代表各种负载下均能达到同一延迟。

传统生成式模型对每一个分类问题都生成一段文字，会产生额外解码开销，也容易使下游系统不得不解析不稳定的自然语言。专用决策接口输出概率和置信度，工程团队可以将低风险自动处理、高风险人工审核、模糊结果重新判断区分开来。架构分工可以更直接：通用模型负责长文本，高频、可枚举的判断交给专用接口。实际应用还要检验概率校准、业务样本漂移、阈值选择及不同行业数据的准确率。

2. DeepMind 发布 AlphaGenome 2：把 DNA 变异效应细化到细胞类型

Google DeepMind 于 11 月 11 日宣布 AlphaGenome 2，面向遗传变异与细胞状态关系研究，训练材料涉及超过 1 亿份单细胞谱系信息，覆盖 68 种组织。新模型以目标组织或细胞状态作为条件，预测特定 DNA 序列的调控作用。官方在 TenK10K 单细胞数据上的对比称，识别候选因果变异的召回结果从约 16% 提高至 30%，方向性基因效应的相应指标由 23% 增至 43%，均是所披露基准场景的结果。模型当天通过 API 向非商业研究开放，商业研究处于私人预览阶段。

这项研究将预测问题缩小到具体组织或细胞状态，而非只看一段序列的总体作用。它能够为研究人员提供后续湿实验筛选线索，但候选结果还须经过独立生物实验验证。DeepMind 明确该工具仅供研究，尚未获得临床诊断用途验证或批准。

3. 初创 Plurai 获 1300 万美元融资并开放智能体仿真平台

美国初创公司 Plurai 于 11 月 11 日宣布获得 1300 万美元种子融资，由 Team8 领投，并宣布智能体仿真平台全面开放。其技术路线不是单靠模型增大规模，而是在客户知识库、工具和策略基础上构造代理运行环境，在隔离沙箱中安排大量正常及对抗性用户场景，提前暴露越权、工具失败、错误答案和敏感信息泄漏等故障，再利用生成的数据改进较小的任务专用模型。

公司在公告中声称，部分设计合作客户的开发周期从数月缩至数小时、成本下降 97%、故障下降 82%，并公布特定项目的成本性能对比；这些均是企业自报成果，未提供足以推广到所有场景的独立样本。对工程团队来说，可复现的仿真场景比一组漂亮的测试分数更有用。企业智能体面向复杂权限和动态业务环境，只用固定测试问答很难预测真实长尾问题；可复现的数字场景有望成为上线验收的重要基础。

4. 阿里云开放 Qwen4-Max-Preview：多模态旗舰限时预览

阿里云模型服务文档 11 月 11 日更新，Qwen4-Max-Preview 进入开发者限时预览。模型支持文本、图像、视频输入以及工具调用、联网搜索、结构化输出，默认启用推理，可设置思考强度；官方标明最长 100 万词元上下文。预览于北京时间 11 月 11 日 10 时开始，持续至 11 月 25 日 24 时，期间只对 Token Plan 开放，不提供按量付费，结束后这一预览版本

将下线。

这一预览只持续两周。用户要留存版本标识和评测结果，不宜把它直接接入不可中断的生产流程。接口兼容性降低迁移成本，但视频内容理解、长上下文稳定性和结构化输出仍需根据业务数据检验。

5. 百度发布 ERNIE 5.5 Flash Preview：2800 亿总参数、140 亿激活

百度文心团队 11 月 11 日推出 ERNIE 5.5 Flash Preview，官方给出 2800 亿总参数、140 亿激活参数以及 100 万词元上下文。模型 API 已上线千帆；此前公开的基础模型经过后训练形成此次预览版，权重计划在 11 月 25 日前以 Apache 2.0 方式开放，目前不得称其已经完成开源。官方展示了通过 FreeCAD 和 Python 脚本生成可编辑机械结构、在 Blender 内制作三维白模以及多文档研究报告等任务。

演示里更值得看的是脚本、文件与结果是否可交付，而非截图是否好看。厂商同样说明预览模型是中间检查点，仍可能存在后训练未完成造成的能力缺陷；对于 CAD 或财务报告等用途，模型产生的参数、运动关系与计算仍需要专业审查。

工程与产业延伸

这五条消息对应不同的交付状态。API 已经上线，可被开发者实际请求；限时预览意味着接口、版本和价格可能改变；“未来开源”只是厂商承诺；研究专用服务则有授权范围和验证边界。采购方在比较模型能力时，应把这些状态写入同一张表，并安排下线、迁移和审计预案。开发环境的可复现程度也十分关键，尤其是当模型要操作代码库、外部搜索、企业数据或科研计算软件时，输入资料和最终产物都应留下版本信息。

趋势观察

这些发布各有明确边界：专用判断接口重视时延与概率校准，通用模型承担更复杂的长任务，科学模型聚焦可验证的研究问题，仿真平台把失败发现前移到上线之前。企业采购和开发应把可复现评测、权限审计、版本管理与真实费用放在同一个验收表中。

参考资料

1. Google DeepMind ｜ Google 推出 Gemini Decisions API：一次计算完成上千项判断｜ 2026-11-11 ｜用于核验主体、当日动作、参数及事实边界
   https://blog.google/innovation-and-ai/technology/developers-tools/gemini-decisions-api/

2. Google DeepMind ｜ DeepMind 发布 AlphaGenome 2：把 DNA 变异效应细化到细胞类型｜ 2026-11-11 ｜用于核验主体、当日动作、参数及事实边界
   https://deepmind.google/blog/alphagenome-2-deciphering-how-cells-read-dna-to-better-understand-disease/

3. Plurai ｜初创 Plurai 获 1300 万美元融资并开放智能体仿真平台｜ 2026-11-11 ｜用于核验主体、当日动作、参数及事实边界
   https://www.prnewswire.com/news-releases/plurai-raises-13m-to-build-the-simulation-infrastructure-for-trustworthy-self-improving-ai-agents-302936422.html

4. 阿里云｜阿里云开放 Qwen4-Max-Preview：多模态旗舰限时预览｜ 2026-11-11 ｜用于核验主体、当日动作、参数及事实边界
   https://help.aliyun.com/zh/model-studio/qwen4-max-preview

5. 百度文心团队｜百度发布 ERNIE 5.5 Flash Preview：2800 亿总参数、140 亿激活｜ 2026-11-11 ｜用于核验主体、当日动作、参数及事实边界

   https://ernie.baidu.com/zh/release/ernie-5-5-flash-preview/

机构介绍与关注

新质生产力工作委员会

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

关注高促会新质生产力工委会公众号

关注工业智能算网平台

https://gyznsw.cn