

大白私刊

书签 · 第 009 期 2026-11-22

大模型分家以后，真正在算的是哪几笔账

从腾讯混元 Hy4 开放权重与 WorkBuddy 千万日活，到《词元之外》第 011 期的信任飞轮、TokenHub 原厂直供 DeepSeek 免收服务端检索费、1984 年川菜馆的算账争议，以及叠衣服时玩的配袜子游戏与睡前用手藏的小金鱼。

2026-11-22

本期目录

今日速览

Anthropic IPO 交割、混元 Hy4 开放权重、智算标准与三则跨界一看

长专题

《词元之外》#011（01:05:03 起谈隐私与信任，01:19:32 起谈阿里、腾讯的模型竞争与 WorkBuddy）：大/小模型怎么分工，Agent 又靠什么攒出第一份信任，本期附有节目与逐字稿入口

追踪

此前关注的项目本期暂无新消息可更新

小说书房

《1984：从川菜馆开始》里的川南烟火气，与做过餐饮的读者对账本漏算人工、房租的质疑

生活亲子

十分钟干净袜子寻宝赛，和不用翻书也能玩的《小金鱼逃走了》

个人创作方法

从一张柠檬水取餐码开始，看《浮屏》怎样守住一个入口

在线阅读本期

今日速览

资本与上市 Anthropic 完成 IPO 交割并挂牌纳斯达克。11月20日公司公告，整笔 IPO（含售股股东）以每股83美元发行11.27亿股、募资毛额约935亿美元，其中公司出售1,103,351,354股、扣费前取得约916亿美元；同期独立私募约260亿美元，A类股19日起以代码 ANT 交易。

开源权重 腾讯混元 Hy4 与 Hy4-flash 开放模型权重。旗舰 Hy4 语言主干 770B 参数、每 token 激活 49B，支持 1M 上下文；两者均按腾讯 Hy 许可托管模型权重。

云端上架 阿里云新加坡站点上线 Macaron 新版本。日志显示 mindlab/macaron-v1.1-venti-pro 11月20日上线国际文本生成，支持1M上下文与读图；qwen4-max 已于16日上架。

接口计费 腾讯云 TokenHub 免除部分 DeepSeek 联网费。北京时间11月18日12:00起，带「原厂直供」标识的 DeepSeek 模型通过 Anthropic API 或 Responses API 的 web_search 触发服务端 Search 与 Fetch 不再计费，模型 Token 费照收；公告承诺起算后产生的相关费用将于11月30日23:59:59前统一退还，无需单独申请。

算力底座 2026 世界计算大会发布智算规模与互连标准。截至9月全国智算达 2592 EFLOPS、建成52个万卡以上智算设施；大会发布3项 CLink 互连标准，刘文强预测2030年推理将占我国智能算力需求的85%。

硬件终端 华为星芒 OSTA 01 Pro 官网公布参考售价。官网展示参考起价 5,699 元，配备 16:9.5 阔感屏、典型值 7000 mAh 电池与 HarmonyOS 7 小艺体验。

校园开发 鸿蒙高校创新赛现多端协同照护设计。少数派11月21日刊发上周末在南京大学举行的颁奖典礼现场采访：山大团队用眼镜、手机与90点触觉手柄帮视障者摸清路况；北工大团队以柔性压阻阵列辅助养老卧床护理。

不妨一看

「邂逅·天津」摄影展今天迎来最后一天。

天津美术馆一层1号展厅自11月6日展出至11月22日，由天津市文旅局主办，分河海、历史、发展、人文、文化五板块呈现城市地景。

查看展讯

天津各大湿地迎候鸟迁徙高峰，多处刷新纪录。



秋日湿地里，白鹳在芦苇间停歇。

《天津日报》11月13日报道湿地保护区监测情况：七里海单日最高现东方白鹳1600只、时隔20年再现3只丹顶鹤；北大港一次性记录27只大红鹳，花脸鸭超7万只。

阅读监测报道

《GTA VI》官方音乐页开放原声曲目与游戏内电台试听。

游戏已于11月19日登陆 PS5 与 Xbox Series X|S；官方音乐页上则可试听34首原创曲目的原声专辑、Justice 的配乐单曲以及多档游戏内电台选段。

前往试听页面

长专题

什么任务才值得用旗舰：从混元 Hy4 到 Personal Agent 的真实取舍

过去大家总盼着一个模型包打天下，现在大厂自己把路线拆开了：重型旗舰留给长程难题，轻量模型顶上高频成本账，而真正卡住个人智能体的，反而成了用户愿不愿意多交出一点信任。



同一只纸鸟，可以有繁简不同的折法。

旗舰 Hy4 先把能力边界摊在台面上

腾讯这次在 GitHub 发布了说明与代码仓库，并在 Hugging Face、ModelScope 与 GitCode 按 Tencent Hy License 同步开放了 Hy4 指令版、FP8 量化版与预训练基座权重。语言主干总参数 770B、每 token 激活 49B，支持 1,048,576 tokens 上下文与内置 3.3B DFly2 投机解码器。但官方仓库在已知限制里并没有藏拙，明确写出在研究级数学、长时程编程等困难内部基准上，Hy4 与顶级闭源模型仍有差距；团队成员 No-Cherry2850 也在 Reddit AMA 里承认，Hy4 为追求能力上限存在思维链偏长的老毛病。

WorkIndex 中，Hy4-flash 怎么把平均单任务成本压到 Hy4 的二十分之一

日常干活不需要每次都请旗舰出马。Hy4-flash 约 220B 总参数（含 64B Engram，结构表拆分主体 155.5B、Engram 63.5B、ViT 0.43B 与 DFly2 0.85B），每 token 只激活约 12B，改用 1:3 稀疏与滑动窗口混合注意力、LatentMoE，并用强化学习刻意压短思维链。按腾讯在 WorkIndex 评测的口径，Hy4-flash 拿到 42.1 分（高于 Hy4 preview 的 39.7、低于 Hy4 的 51.4），平均每任务成本却只有 Hy4 的 1/20；在 85 名 WorkBuddy 内部盲测用户里，它以 6.7 分钟中位耗时和 0.26 元人民币的单任务成本，换来接近 Claude Haiku 5.5 的 2.79/4 满意度。官方因此自己划了线：长时程软件工程用 Hy4，做幻灯片、改文档和日常搜索交给 Hy4-flash。至于社区追问的更小尺寸 Hy4-mini 与后续更大模型，团队成员在 AMA 里分别作了回应：DiWang01 确认 Hy4-mini 已在路线图上，aiden_agent 则表示计划训练更大的模型，但两者的规格与时间表均尚未公布。

WorkBuddy 千万日活说明：模型必须长在入口里

模型拆得再细，也得有地方用得上。腾讯 11 月 18 日在港交所发布的三季报给出了一个硬数字：按大语言模型词元消耗计，WorkBuddy 已成国内使用量最高的 AI 产品，每日超 1,000 万名用户合计消耗 100 万亿以上词元；财报同时提到，受 WorkBuddy 的带动以及腾讯增加 MaaS、GPU 租赁算力投入两方面因素影响，云服务收入增速有所提升。《词元之外》第 011 期在 11:14 先谈到日常编程及普通软件评测趋于饱和，又在 01:19:32 聊到 Qwen4-Max 与 Hy4-flash 的卡位时提到，模型如果不在性能与成本的帕累托前沿上，只靠发新版本很难留住人；WorkBuddy 这类高频办公使用入口，恰好成了 Hy4-flash 跑通真实任务的现成试验场。

Personal Agent 的真门槛：第一轮帮上忙，云端账算得清

相比办公入口，面向个人的 Agent 难在哪里？《词元之外》第 011 期在 01:05:03 切入隐私与信任讨论，橘子在 01:07:29 点得很直白：Personal Agent 的本质就是信任，你只有先看到它真能办成事，才会放心地交出更多上下文，进而得到更贴合的帮助；如果头一两回显得笨拙，用户就不会再给任何信息。节目在 01:02:09 分享了一个小例子：家里买电钢琴时，把三款候选型号交给名为 Cola 的 Agent 做调研比较，最终按它的性价比建议直接下了单。在特定云服务的调用成本上，腾讯云 TokenHub 自 11 月 18 日 12 时起对部分联网能力免除单独计费：调用带「原厂直供」标识的 DeepSeek 模型，并通过 Anthropic API 或 Responses API 的 web_search 工具触发服务端 Search 与 Fetch 时，不再单独收取检索与抓取费用，仅按模型 Token 用量计费。

《Next Token丨词元之外》#011

2026-11-21(11月20日北京象外空间录制) · 歸藏、杨攀、橘子、向阳乔木

从让 Cola 对比三款候选电钢琴的功能需求与性价比、家人最终按推荐下单的真实经历切入，拆解 Personal Agent 如何通过先帮上忙积累上下文信任；节目还独立聊到 Neon 的剪辑试用(46:08)与 Subcast 介绍(01:17:26)，以及阿里 Qwen4-Max 与腾讯混元 Hy4 flash 的模型竞争，和 WorkBuddy 怎样为腾讯混元提供试用场景。

01:02:09(电钢琴调研案例)、01:05:03(隐私与信任飞轮)、01:07:29(信任正反馈机制)、01:19:32(Qwen4-Max 与 Hy4-flash 讨论)

阅读逐字稿

追踪

此前关注的项目本期暂无新消息可更新



小说书房

《1984：从川菜馆开始》

轻语江湖 连载



川南小饭馆的手擀面，从一张撒了面粉的案板开始。

故事一开场就把主角周砚扔进了 1984 年川南小镇的烂摊子里。作为穿越回来的现代美食博主，他刚接盘一家因原主经营不善而濒临破产的川菜馆——此前原主只顾着闷头做菜、人情一塌糊涂而被厂食堂赶出来，又盲目堆出三四十道菜的菜单，把日子过得鸡飞狗跳。周砚手里可用的牌很实在：一是现代厨艺见识与绑定的美食系统，二是阴差阳错救人攒下的一点好名声，以及家里杀牛匠老汉、麻利老妈、嘴馋又能带货的小妹、曾号称卤菜教主的奶奶和一众勤快叔伯；但摆在眼前的约束同样硬邦邦——八十年代初小镇物资紧巴、食客兜里钱不多，饭馆也没多少本钱可烧。他只能先大刀阔斧删掉冗余菜单，从成本最低也最考验基本功的手

擀面起手，靠一道道扎实的吃食与顺理人情，把小饭馆从倒闭边缘一步步盘活。

从这里进去

书海鱼人的三位书友(一人推荐、两人试读)把这本的爽感拆得很清楚。首先是两位书友共同点赞的做菜实在感:「推书少女西米荞」肯定蒸炸炒卤的火候、调味写得细致又有可行性，没有脱离年代像别的作品那样拿白面粉洗大肠、只留蛋黄扔蛋白般浪费食材;「只有一点点小愿望」则赞赏菜品从手擀面到跷脚牛肉、卤菜循序渐进，配上章节插图很下饭。其次是川南一家子的烟火日常，「只有一点点小愿望」尤其偏爱老汉钓鱼老空军要靠儿子买鱼撑面子、老妈收到新鞋出门显摆、小妹啃猪蹄就能当销冠的群像细节，再夹杂带喜感的四川方言，让日常过得热气腾腾。「弗虑山」还给了一个很有参考价值的坐标对比:认为它在年代人情与剧情铺陈上有《生活系游戏》的影子且更精妙，但单论做菜本身的“道道”不如《妙厨》，作者下厨更像美食短视频博主“知其然不知其所以然”，少了一层直击食客的勾人深味。

阅读门槛

最容易劝退读者的门槛集中在经营常识与年代物价的经不起细抠。在2026 年 2 月这组书评里，「弗虑山」指出两个明显裂口:一是 1984 年牛肉本属凭票供应的稀罕物，书中却借“全家屠夫”的设定让主角天天端得出牛肉;二是按当时全国职工月均约 80 元(年均 961 元)的收入水平，一个川南小城很难撑起每天上百号人下馆子且营收持续暴涨，「弗虑山」表示他个人的阅读兴趣就停在这里了。「推书少女西米荞」以在饭馆打过一年工的经验吐槽更直接——主角算账只算食材成本，人工、水电、房租和利润预备金全不见了，而且备货量偏少、单价又贵，现实里客人连扑几次空早就不再回头;她另回忆读过作者写当年开店破产关门的《上架感

言》。此外，2 月书评里「只有一点点小愿望」当时读到的感受是部分情节偏巧合，书友彼时推测女主应是那位落水女大学生，两人以带着年代感的书信往来；本书适合冲着家庭群像和下饭菜谱去看的读者试读前几十章，若对餐饮经营算账格外较真则需提前降一档预期。

生活亲子

今晚玩什么

干衣堆里的袜子寻宝赛

约 10 分钟　一篮待叠的干净衣物(里面混着几双成对的袜子)、一个孩子够得着的低矮抽屉或收纳篮，可选配一个计时器

趁着收拾刚洗完的干净衣物时，先别急着把袜子挑出来，直接把所有袜子混在衣堆里，或者平铺在床面与地垫上，和孩子约定玩一次「找另一半」的寻宝赛；如果家里有不少颜色接近、只差尺码或微小花纹的袜子，低龄宝宝刚开局时可以先悄悄撤掉几双最容易混的。

1 挖宝

大人一边叠大件衣服，一边请孩子蹲在干净衣堆或衣篮旁，用手从里面往外翻找出所有藏着的袜子，每摸到一只就大声报出它的颜色或上面印着什么图案。

2 配对



大人手里举起一只单袜，请孩子在摊开的袜子堆里帮它找回「另一半」。若是两个孩子一起玩，还可以定一个 5 分钟小计时器，铃响时数数各自配成了几双，配得最多的就赢。

3 比长短

碰到颜色、图案一模一样但尺码不同的兄妹袜或亲子袜时，让孩子把两只袜底对齐放在一起比长短，挑出真正一般大的一对。

4 收进家

每配好一双就由大人团成一对，请孩子每次把团好的一双袜子亲手放进够得着的低抽屉或收纳篮里归位。

观察孩子在这一轮里最起劲的是哪一段：是更喜欢「从衣堆里往外掏」、对照花色配对，还是「自己推进最下层抽屉」，顺带练了颜色辨认、大小比对和帮忙干活的成就感；若发现孩子遇到同色微差袜子开始烦躁，下次就减少混放数量、只留花纹反差大的几双。

今晚睡前绘本

《小金鱼逃走了》

五味太郎 · 3岁亲子共读

新书种草 · 家里没书也能玩

一条红色小金鱼从鱼缸里溜出来后，接连躲进和自己颜色、形状相似的房间角落：红色圆点窗帘、花丛、糖果罐、水果盘……孩子每翻一页，都要从这些图案里把小鱼找出来；中间它还在镜子上变成三条一模一样的小金鱼考验眼力，最后游进满是同伴的大池塘里，让孩子从许多条金鱼中认出最初逃走的那一条。

买还是借 这本低幼找图绘本全靠五味太郎鲜亮的色块，把小金鱼藏在窗帘花纹、花丛等相似图案里让孩子反复指认，很适合亲子一页页共读。如果家里有正爱玩找东西和指认游戏的小朋友，可根据阅读需求考虑买或借——豆瓣上既有「LV姑娘」分享自家 18 个月的宝宝就能独自翻看（不过她提到，结尾「镜面三条鱼」和「大池塘找原鱼」对当时这孩子来说还分辨不出），也有另一位家长记下快两岁的孩子第一天看了六遍；同时也有家长提到自家娃很快能找到鱼，但好像没觉得找到就很开心；如果拿不准自家孩子会不会喜欢，也可以先去图书馆借回试试再决定收不收。

今晚玩法 关灯前父子俩先并排躺进被窝，爸爸把声音放轻开场：“嘘，刚才床头鱼缸里‘扑通’响了一声，有条红彤彤的小金鱼逃跑啦！它最狡猾了，专门挑跟自己长得一样红扑扑、圆滚滚的地方藏起来。你猜它这会儿躲进咱们被窝哪儿了？”等孩子学小金鱼“咕噜咕噜”吐泡泡或者伸手指向被角、爸爸袖口时，爸爸就把一只手掌并拢成鱼尾巴，顺着孩子指的方向从被面上“游”过去，贴着被窝褶皱藏住半个手背，接着爸爸把那只“鱼手”悄悄藏进两个被角中的一边问：“哎呀，小金鱼这次躲到被角底下啦！快用小手指点一点，把藏着爸爸那只‘鱼手’的一边找出来！”

准备入睡 抓到最后那条小金鱼后，爸爸带着孩子把双手合起来做成小鱼游动的样子，慢慢往枕头边游去，小声说它找到好多朋友回池塘睡觉啦，请孩子吐完最后一声轻轻的“咕噜”泡泡，把手放回被子里闭眼睡下。

个人创作方法

只接一个动作：从《浮屏》看小工具的克制生长



很多独立开发的小工具最后越做越重，往往是因为一开始就想替用户管全套生活；独立开发者 Nezuko 做的《浮屏》，切口却小得具体——夏天夜跑后买一杯柠檬水，等餐时总要一遍遍切回小程序看取餐码。他没有按电影票、火车票、取件码拆出好几套表单，而是只定下一条铁律：不管截图来自哪个 app，全都用同一个动作交给浮屏，识别、分类和锁屏实时活动排版全在后面完成。

真正花心思的不是多接几个场景，而是想清楚每一种卡片「在锁屏上最该突出哪几行字」。更有趣的是，原本用完就该划掉的临时卡片，在被不断

点成「完成」之后，反过来催生了最初没规划过的「回顾」日历——那些过去会烂在相册里的电子票根，顺势长成了一本不用动笔的生活流水账。

同时作者没有回避代价：因为靠云端模型解析图片，截图会上传给模型服务商处理，速度也受制于网络与推理延迟，只能尽量压到 10 秒内；换来的取舍是识别效果不必等客户端发版就能随时微调。把一个高频动作先收窄到顺手，坦白边界，再顺着自己每天打开的习惯长出新功能，是个人小工具很扎实的生长法子。

出处与索引

Anthropic IPO 交割公告

仅采信 Anthropic 2026年11月20日官网公告：A类股11月19日在纳斯达克挂牌（代码 ANT），IPO 发售与同期N类股私募总额分开核对，金额均以美元计。

腾讯混元 Hy4 官方仓库与模型卡

核自腾讯混元官方仓库与 Hugging Face 模型卡（huggingface.co/tencent/Hy4）：语言主干 770B/49B 激活、1M 上下文，按 Tencent Hy License 开放指令版、FP8 版与基座权重；已纳入已知局限说明。

腾讯混元 Hy4-flash 模型卡与 Reddit AMA

参数取自腾讯模型卡正文（约220B总参数含64B Engram、约12B激活）及组件表（155.5B+63.5B+0.43B+0.85B）；每任务成本与评分属腾讯 WorkIndex 及 85 位 WorkBuddy 内部盲测口径。另结合 Reddit AMA（reddit.com/r/LocalLLaMA/comments/1y15w4a/）署名账号说明：No-Cherry2850 确认团队已用 RL 压短 Hy4-flash 思维链，DiWang01 与 aiden_agent 分别提及 Hy4-mini 已入路线图及后续更大模型计划，后两者均未公布具体规格与时间表。

阿里云 Model Studio 新加坡上架日志

核自阿里云官方新加坡地区变更日志：mindlab/macaron-v1.1-venti-pro 2026年11月20日上线国际文本生成范围，qwen4-max 条目标注11月16日；仅代表该地区上架记录，不外推为全球同日发布。

腾讯云 TokenHub DeepSeek 服务端搜索免费公告

核自腾讯云公告：2026年11月18日12:00起限带「原厂直供」标识的 DeepSeek 模型经 Anthropic API 或 Responses API 的 web_search 触发的 Search/Fetch 免单，仍收 Token 费；北京时间2026年11月30日 23:59:59前统一退还为公告承诺。

2026 世界计算大会开幕与智算趋势报道

核自新华网11月20日现场报道及湖南省政府门户11月21日转载湖南日报报道（hunan.gov.cn/hnszf/hnyw/sy/hnyw1/202611/t20261121_34107799.html）：区分截至2026年9月全国智算2592 EFLOPS与刘文强对2030年10000 EFLOPS的预测口径；CLink 3项标准仅按现场发布方向记录。

华为星芒 OSTA O1 Pro 官网产品页

规格与价格仅采信华为官网产品页：¥5,699 为参考起价（实际以电商页为准），电池典型值 7000 mAh、额定值 6815 mAh，16:9.5 阔感屏。

少数派丨2026 鸿蒙高校创新赛总决赛现场采访

核自少数派编辑部2026年11月21日发稿：方案细节为记者在南京大学总决赛现场对参赛团队的访问记录，设备规格属团队介绍口径。

天津美术馆「邂逅·天津」摄影展公告

核自天津美术馆官方展讯页（2026年11月6日发布）：展期至11月22日（一层1号展厅），未核及当日现场客流或预约余量。

天津政务网丨秋冬候鸟迁徙监测消息

核自天津政务网2026年11月13日转载《天津日报》报道：保留单日最高、一次性记录与累计观测的不同统计口径。

Rockstar Games《GTA VI》官方音乐页与发售公告

核自 Rockstar 官方音乐页及 Take-Two 2026年11月19日发售公告（taketwointeractivesoftwareinc.gcs-web.com/node/32696/pdf）：仅确认 PS5 与 Xbox Series X|S 发售及34首原创原声、Justice 配乐与电台试听信息。

腾讯控股 2026 年第三季度业绩公布

核自港交所2026年11月18日腾讯三季报原文：披露 WorkBuddy 每日用户超1,000万、合计每日消耗超100万亿词元及云服务收入增长。

《Next Token | 词元之外》第 011 期节目与逐字稿

核自节目官网、官方 RSS（2026年11月20日16:00 GMT / 北京时间11月21日00:00发布；11月20日北京象外空间录制）及中英文逐字稿（/weekly/011/transcript 与 /en/weekly/011/transcript）。仅采信有明确说话人署名的案例与观点，跳过标记待核发言，且不以口播替官方数字背书。

书海鱼人《1984：从破产川菜馆开始》原评汇编及作品信息

原评摘自书海鱼人2026年2月5日存档（涵盖「只有一点点小愿望」「弗虑山」「推书少女西米荞」三位读者的历史阅读反馈）；作品名、作者与连载状态另经小说网作品主栏（xs.cn/bookquery/qbqbmiiqpefo，显示2026年11月21日23:45更新）核对，起点首发页与最新章详情未独立读回。

Born Learning 手册与家长洗护配对实践

活动框架核自 United Way Born Learning 手册（幼儿拿放衣物、学龄前颜色分类与袜子配对）；操作灵感分别参照 Reddit 两则公开讨论：r/toddlers/comments/qyzsry（发帖人显示为 [deleted]）中叠衣时请两岁半孩子在干净衣堆找袜、配对并放回低抽屉的做法，以及 r/UKParenting/comments/1h094mo 中 oldfilmjanitor 提及不同尺码的袜子易混、Wavesmith 分享三岁孩子爱玩配袜与两人五分钟限时配对赛的经验。

福音馆书店《きんぎょが にげた》与豆瓣、国会图书馆书目

绘本主创与找图设计核自福音馆官网及日本国立国会图书馆书目（1982年8月日文版，23页、22厘米，ndlsearch.ndl.go.jp/books/R100000038-I4717296）；适龄与买借判断参考豆瓣页（book.douban.com/subject/10427241/）署名短评（LV姑娘、蜗牛）及书评（锦瑞《逃走了》，2016-09-09），未把页面聚合评分或库存 UI 冲突信息作依据。

少数派丨《App+1丨浮屏》独立开发者自述

核自独立开发者 Nezuko 于2026年11月20日发表的创作自述，仅采信其披露的产品起点、入口取舍、回顾页演进与云端识别隐私说明；实际应用商店审核与上架状态未另核。