

上一期结尾留了个问题

数百万司机在线找货



货被抢走

几分钟

模型推理

几十秒

同一根时间标尺 · 长度就是时间

这个问题怎么解？

他们给的解法就一句话

让 AI 当编译器，~~别当司机~~

✗ 不铺知识点

✓ 跟着这一个真实系统走

难题 ①

难题 ②

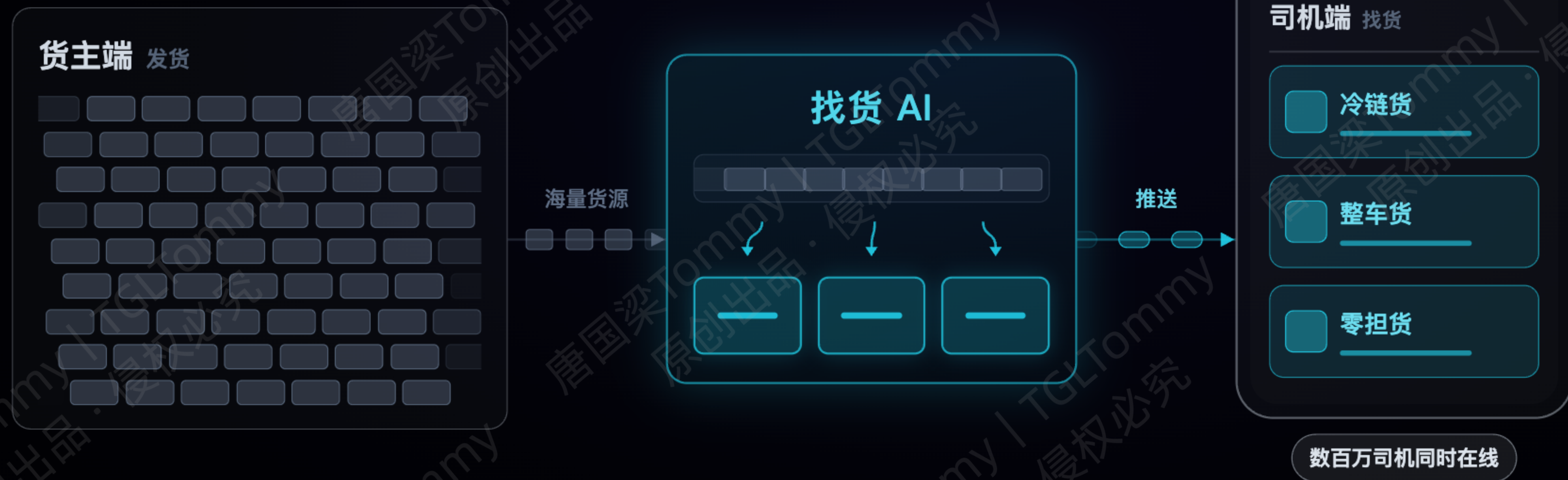
难题 ③

一个一个解开

答案不只属于货运这一行

运满满

数字货运平台



在海量货源里，盯住适合他的那些货

三个难题

听起来就是个标准的 AI 推荐场景

难题 ①

时间

难题 ②

钱

难题 ③

险

这块板，后面一格一格填



几分钟就被抢走

难题 ①
时间

难题 ②
钱

难题 ③
险

等模型把这一轮推理跑完

货早没了

货源存活



几分钟

模型推理



几十秒

二十秒

六十秒

说不准是哪一个



吞吐量



加机器 → 变多 ↑

单次延迟



加机器 → 不变 =

难题 ②

钱

找货不是一锤子买卖

难题 ①

时间

难题 ②

钱

难题 ③

险



几小时 / 人

长时间持续的任务

×



数百万司机

同时在线

×



模型循环

持续运转 · 每人一个

是乘出来的，不是省出来的

一天烧掉的调用费



这个业务做不下去

难题 ③ · 险

找货 AI 能做的动作，风险天差地别

难题 ①

时间

难题 ②

钱

难题 ③

险

动作 ①

推送一条货源消息

顶多是打扰

同一个开关，三种代价

难题 ③ · 险

找货 AI 能做的动作，风险天差地别

难题 ①
时间

难题 ②
钱

难题 ③
险

推送

顶多是打扰

外呼

骚扰投诉的风险

抢单

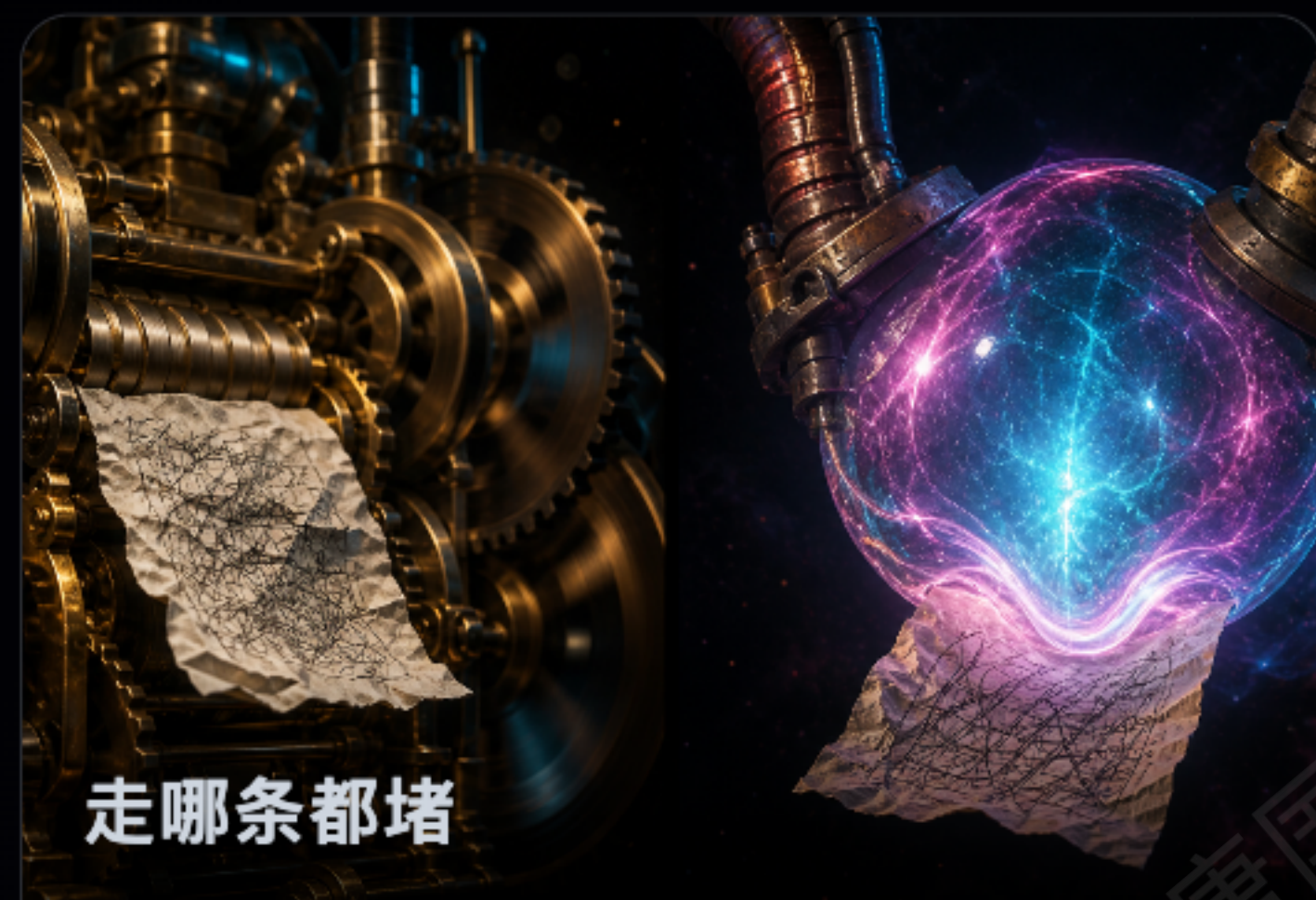
真金白银 · 订金 + 信任

保守：什么都不敢做

一套逻辑通吃三种动作

激进：早晚出事

权限分级 → 第 3 讲



走哪条都堵

司机说的

「我想跑附近的冷链货，
价格别太低」

- ⊗ 规则引擎读不进去
- ⊗ 学不会个人偏好

聪明

大模型

听得懂
又慢又贵

?

既快 · 又聪明

快的不聪明
聪明的不快

规则引擎

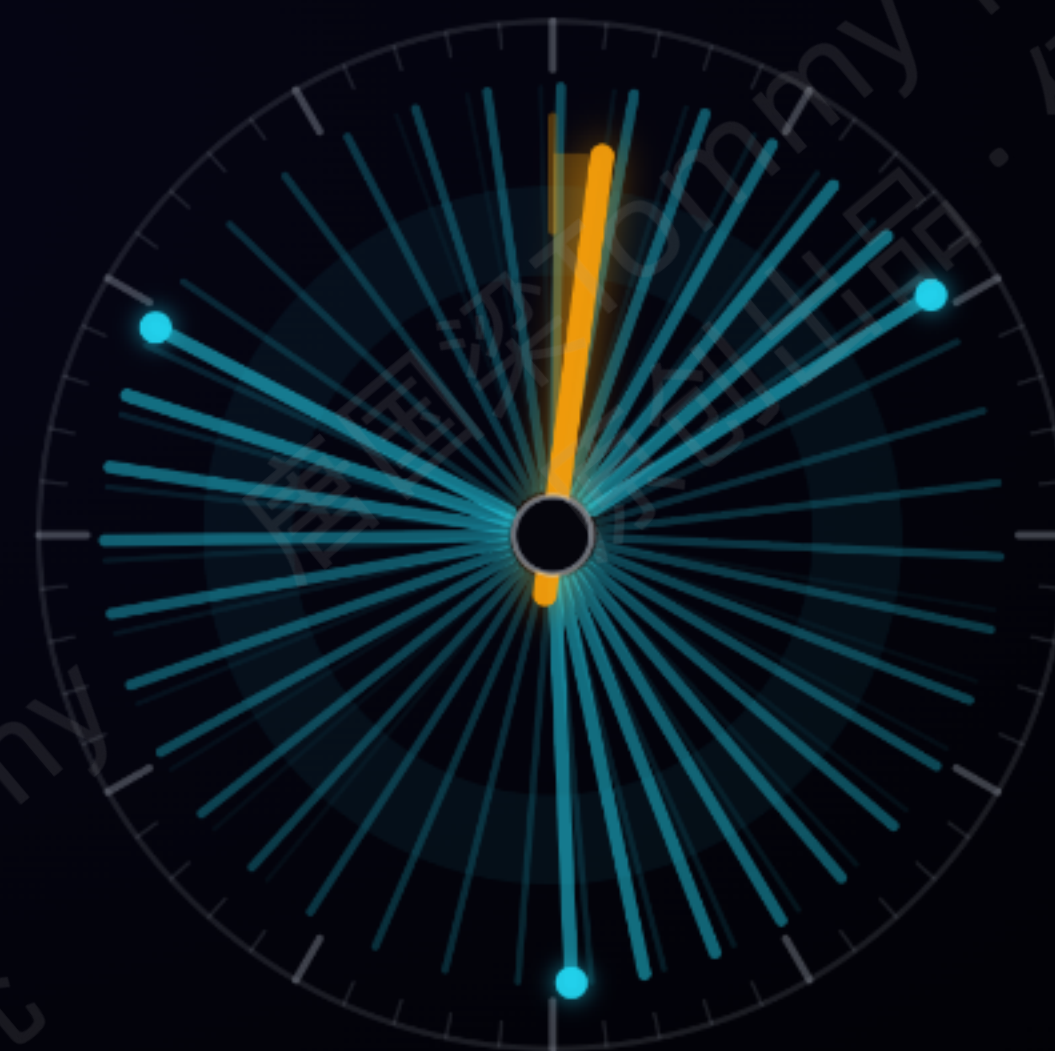
够快够便宜
听不懂

快

先把问题重新定义一遍



两种变化的速度



司机的意图，变得很慢

同一根时间轴 · 同一段时长

意图

分钟级 / 小时级

换一条线路

到货卸完了

行情涨了跌了

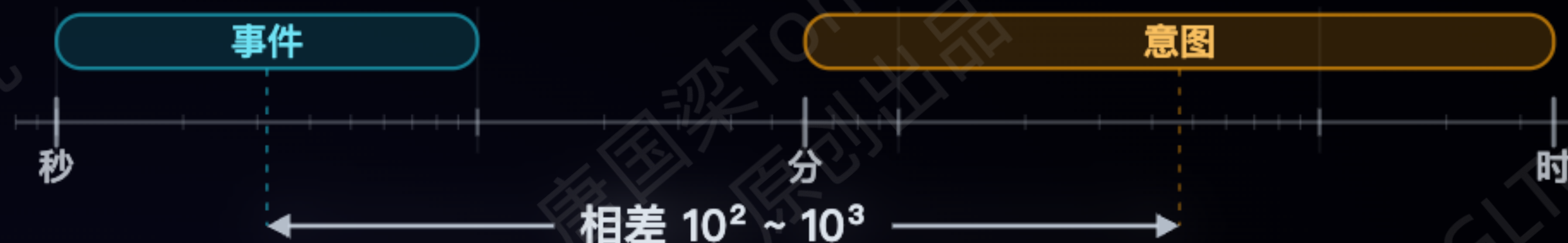
事件

秒级

新货上架 · 货被抢走 · 货主改价

时间尺度

对数轴 · 一格 $\times 10$



慢循环

分钟级 / 小时级

换一条线路

到货卸完了

行情涨了跌了

快循环

秒级

新货上架 · 货被抢走 · 货主改价

慢循环

分钟级 / 小时级

快循环

秒级

慢的那边



大模型

慢循环

干四件事



推断任务

回家

赚钱

顺路捎一单

①



提取条件

自然语言



结构化约束

②



学习偏好

不拉危险品

只走高速

③



借鉴经验

跑得好的老司机的路子

④

盯盘计划

结构化

四件事做完 → 编译成一份结构化的盯盘计划

慢循环

分钟级 / 小时级

快循环

秒级

盯盘计划 ---

货源流

秒级不断涌入

规则引擎

①

匹配

②

评分

③

决定推不推

快循环

推



不推



难题 ①
时间

难题 ②
钱

难题 ③
险

「让 Agent 做判断，工程保证执行的确定性」

—— 运满满

▶ 只按计划执行，不做决策

铁律

司机的对话

历史

模型

盯盘计划

① 召回条件 看哪些货

■ 出发地
从哪出发

■ 目的地
去哪

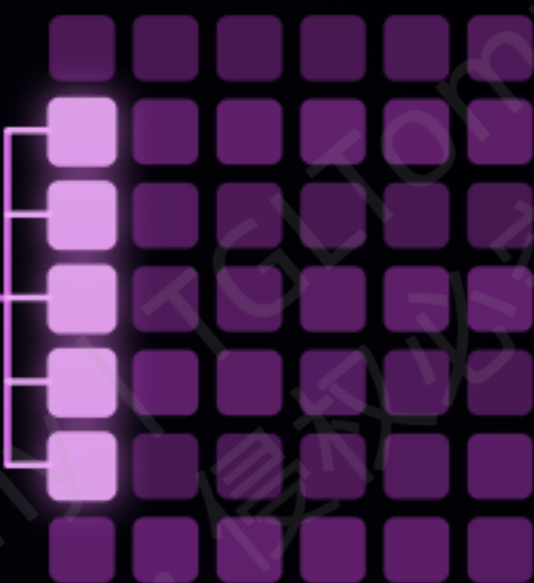
■ 货型
什么货型

■ 价格
价格下限

■ 车型
什么车能拉

② 阻断规则 什么货坚决不看

③ 执行参数 怎么推



共 40 多个维度

盯盘计划

① 召回条件 看哪些货

出发地 · 目的地 · 货型 · 价格 · 车型

40 多个维度 模型已填好

② 阻断规则 什么货坚决不看

■ 黑名单货主

■ 超重超限

■ 推荐过没反应

▶ 每条独立判定 · 命中即拦 · 不需要模型参与

③ 执行参数 怎么推



盯盘计划

① 召回条件 看哪些货

出发地 · 目的地 · 货型 · 价格 · 车型

40 多个维度

模型已填好

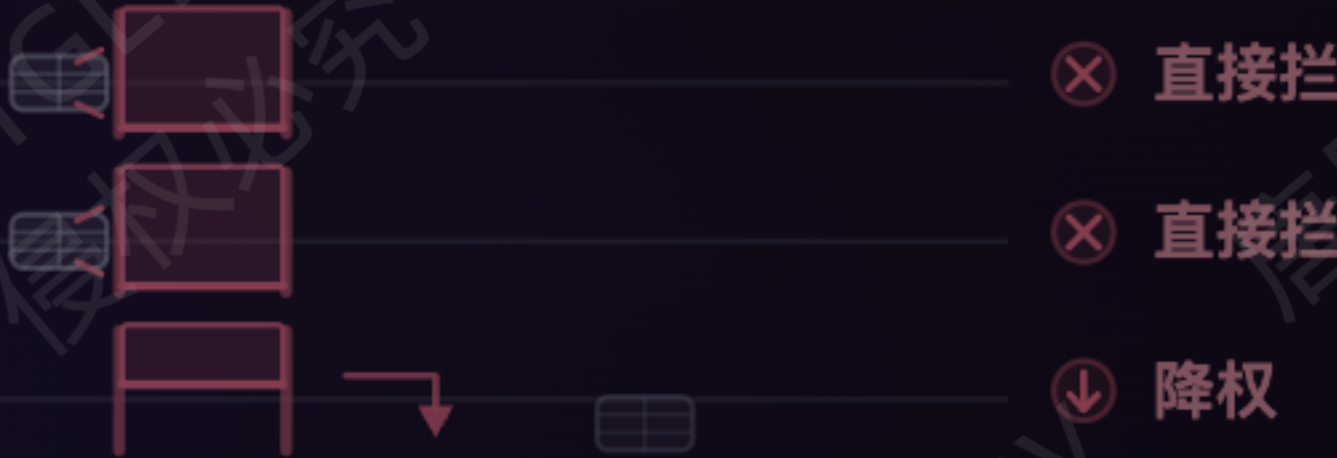
② 阻断规则 什么货坚决不看

■ 黑名单货主

■ 超重超限

■ 推荐过没反应

▶ 每条独立判定 · 命中即拦 · 不需要模型参与



③ 执行参数 怎么推

单次条数

最多 3 条



推送间隔

≥ 5 分钟



日配额

一天最多 20 条



司机说的

「我想跑附近的冷链货，价格别太低」

↓ 编译

			DSL
1	origin_radius_km	: number ≤ 150	
2	cargo_type	: enum ['冷链']	
3	price_floor	: number ≥ 2.4 元/公里	
4	⋮		

字段名

类型

取值范围

Schema

DSL

计划本身长什么样

Schema

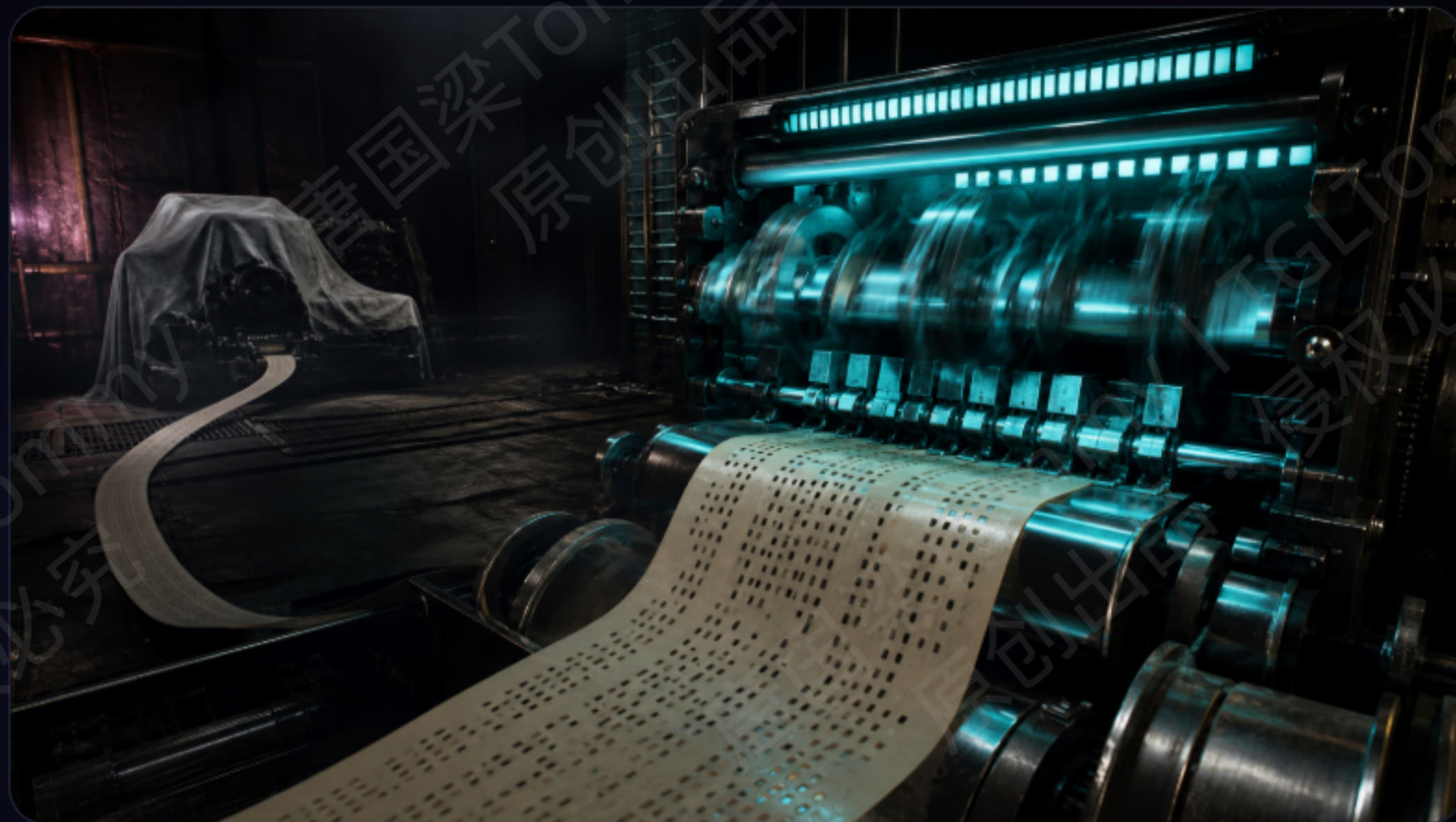
计划必须长成什么样

把一句模糊的话，变成机器能校验、能执行的东西

这个过程，其实就是编译

| 编译器 · 把代码翻译成机器指令

| CPU · 拿着指令执行



源代码



编译器



机器指令

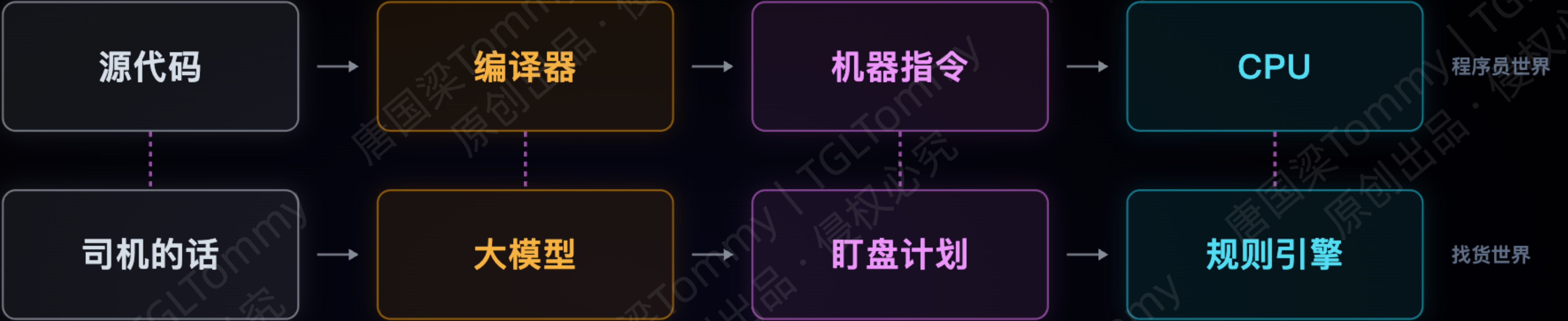


CPU

程序员世界

翻译完就退场 · 运行时并不在场

这里一模一样



同一套结构 · 换一个世界

让 AI 当编译器

→ 只负责翻译一次意图

别当司机

→ 别让它亲自处理每一单货

盯盘计划

v3.2

① 召回条件 看哪些货

价格下限

② 阻断规则 什么货坚决不看

③ 执行参数 怎么推



盯盘计划

v3.2

① 召回条件 看哪些货

价格下限

② 阻断规则 什么货坚决不看

③ 执行参数 怎么推



| 三道保险

第一道

Schema 校验

- 字段类型
- 取值范围
- 必填项

机器自动检查

↓
不合法的计划根本发不出去



三道保险

计划

① Schema 校验

机器自动检查

✓ 字段类型

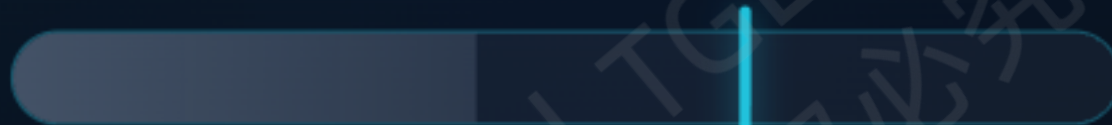
✓ 取值范围

✓ 必填项

不合法的计划根本发不出去

② 置信度门槛

没听懂司机意图



把握不足

?

司机

反问一句确认，不硬编

③ 版本管理

每份带版本号

v1

v2

v3

生效

一键回滚到上一版

发布

第三道保险

顺手多出来的一件事

可审计



能指住的那一行



指不住的一团雾

司机投诉

「为什么半夜给我推这单货？」

有结构

查 ↓

盯盘计划

v3.2

① 召回条件 看哪些货

② 阻断规则 什么货坚决不看

③ 执行参数 怎么推

→ **push_window: 00:00-06:00**

max_per_day: 20

min_interval: 5min

✓ 答案一目了然

没结构

翻 ↓

大模型当时的对话记录

?? ? 猜它为什么那么想

结构化的计划 = 执行的依据 + 事后追查的证据

难题 ②

钱

一次编译 · 一万次执行

难题 ①
时间

难题 ②
钱

难题 ③
险

慢循环
分钟级 / 小时级

同一份计划 · 一直有效

快循环
秒级

新货上架 · 货被抢走 · 货主改价

模型编译

1 次

规则引擎执行

10,663 次

模型调用费

¥ 0

什么时候叫醒模型？

事件驱动

四类



前三类：外部世界变了 | 第四类：AI 自我诊断

「长程任务 $\neq$ 持续运行」

运满满原话
事件还分轻重缓急

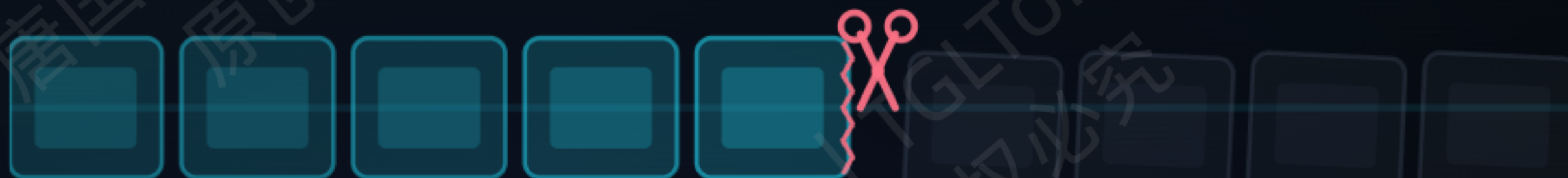
① 抢占

🗨️ 司机开口说话



最高优先级

正在跑的活动



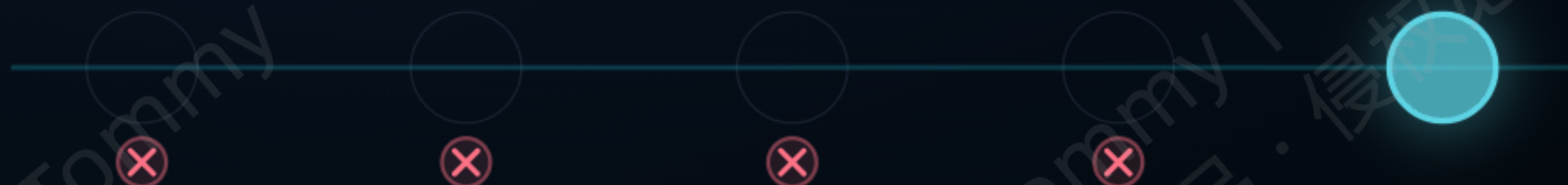
立刻中断，先响应人

② 去重

🔄 周期性的巡检信号

积压了一堆

信号队列



只认最新那条

前面的全作废
同样的活不干第二遍



0h

5h

找货任务 · 几小时

模型

意图变化的那几个瞬间

引擎

其余时间全在值班

还差最后一步才闭上

推送结果 · 持续回流



执行归执行，学习归学习 —— 各按各自的节奏转

快

① 时间

省

② 钱

还不误事

③ 险



这套设计，
只有货运才用得上？

| 不止运满满一家

荣耀

手机助手

华泰证券

交易

大数据运维

运维

腾讯金科

合规

运满满

货运

互不相识 · 各自的生产压力 · 同一个设计

荣耀
手机助手

华泰证券
交易

大数据运维
运维

腾讯金科
合规

运满满
货运

互不相识 · 各自的生产压力 · 同一个设计

荣耀 手机助手

输入 用户口头交代的自动化任务

产物 受约束的触发规则 + 动作规则

执行 注册前确认一遍 · 触发时再校验

▶ 没确认 · 没校验，就不触发



同 构

华泰证券 交易

输入 真金白银的交易场景 · 交易意图

产物 结构化提案 —— 不是操作

执行 独立校验 + 签名，执行器才认

▶ 没签名，执行器不认

安全篇 → 第 3 讲

两个行业 · 同一张表

荣耀
手机助手

华泰证券
交易

大数据运维
运维

腾讯金科
合规

运满满
货运

互不相识 · 各自的生产压力 · 同一个设计

大数据运维 运维

输入 运维意图

产物 符合 Schema 的动作意图

执行 另一条经过校验的通道

▶ 模型碰不到服务器



同 构

腾讯金科 合规

输入 监管法规

产物 决策表

执行 决策表执行审核

▶ 模型只负责理解和取证

模型

服务器

执行



理解 · 取证

加上运满满，一共 **五** 家

荣耀

手机助手

华泰证券

交易

大数据运维

运维

腾讯金科

合规

运满满

货运

场景八竿子打不着 · 做法却撞在了一起

不是哪篇论文



工程里最硬的证据 = 独立收敛

内核就一句话

生成计划阶段

模型自由度

100%



0

100

四件事

推断

提取

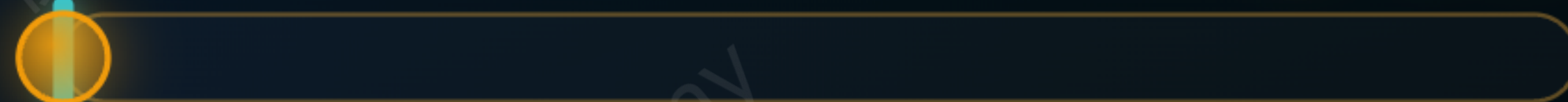
学习

借鉴

执行阶段

模型自由度

0%



0

100

铁律

只按计划执行，不做决策

Agent =

Model

⊕

Harness

上限

下限

本期

上一期的等式



聪明劲儿只许进 计划



不许直接进 动作

最后两件事

不是万能的

第一 · 硬前提

决策逻辑 必须能结构化

方模具吃不下毛边的东西



决策逻辑必须能结构化

硬前提

司机说「价格别太低」

`price_floor : number ≥ 2.4`

✓ 可翻译

司机说「帮我找点顺眼的货」

???

✗ 翻不了

↓ 硬翻的后果

一份看起来合法的计划

✓ Schema 校验 · 通过

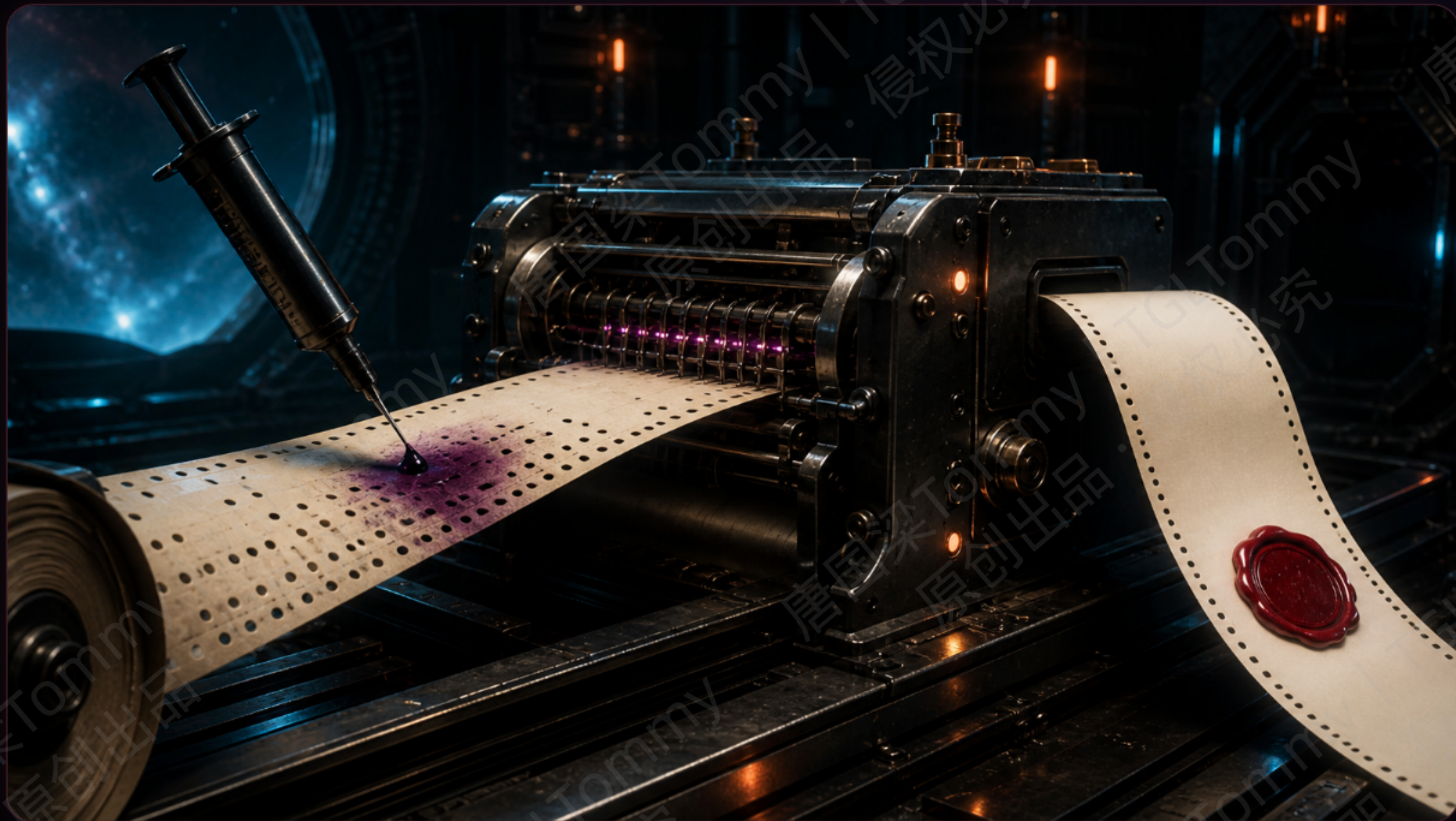
其实完全不符合本意

✗ 结果全错

正确做法：反问澄清 不是硬编

一个还没解决的风险点

② 更值得记住



| 计划是模型从输入编译出来的

| 输入本身被人动了手脚

⊗ 机器毫无察觉，照常盖章

✖ Prompt 注入



普通聊天

一次性

本期结构

忠实执行



全行业尚无经过验证的防御方案

三个难题

解决了两个

难题 ①

时间

靠：快慢分离

难题 ②

钱

靠：事件驱动

难题 ③

险

靠：授权体系

剩下的风险

只挡住了一小半

推送
?

外呼
?

抢单
?

凭什么给 AI 开放不同等级的动作权限?

~~一份计划~~ × 管不住

一整套授权体系 →

第 3 讲

第 3 讲

为什么 Prompt 里写「请不要删库」**根本没用**
怎么让 AI 就算想犯错，也做不到

《Agentic AI 工程五讲》· 第 2 讲完 → 第 3 讲

版权声明

唐国梁 Tommy 原创出品 | 熵智未来（深圳）科技有限公司版权所有。保留所有权利。

本课程及配套资料，包括但不限于课程视频、PDF 文档、课件、代码、图表、示例项目、知识库文档及其他学习材料，仅供购买本课程的用户本人学习使用。未经版权所有者书面许可，任何个人或机构不得以任何形式对本课程及配套资料进行录制、复制、传播、上传、公开分享、分发、转售、改编、翻译、汇编，或用于商业培训、企业内训、商业化产品、知识库建设、数据集构建、人工智能模型训练、微调及其他商业用途。如发现侵犯本课程及配套资料相关权益的行为，权利人有权依法要求停止侵害、删除下架、赔偿损失，并追究相应责任。

本课程及配套资料中涉及的第三方开源软件、模型、代码、数据集、商标、图片及其他素材，仅在相关许可证允许范围内进行教学展示，其相关权利归原权利人所有。

侵权举报 / 授权合作: legal@tgltommyai.com

微信公众号: 唐国梁TGLTommy | 官方网站: TGLTommy.com (国内访问需科学上网)