

设计说明

持仓与自选主动解读 · 金融 AI Agent 产品经理作业 · 2026-10

TL;DR

- 定位：**持仓翻译官，不是投顾。普通用户缺的不是信息量，是“看懂”和“跟我的钱有什么关系”。ROBIN 只解释发生了什么，不给买卖建议。
- 何时说：**默认沉默。触发分五类：按标的自身波动率标准化的异动并叠加个人金额影响、硬事件、定时简报、页面场景、用户自定义规则与承诺跟进。
- 说什么：**先给一句结论，再说对你的钱影响多少，然后是大盘/行业/个股归因瀑布、带时间与来源层级的证据，最后是接下来看什么。
- 怎么送：**按重要性、相关性、必要性加权打分，再按用户的注意力设置分为推送、气泡、简报、沉默四档。配套日上限、30 分钟合并、跨渠道一事一说、安静时段和疲劳降级。不说话的时候也要留“安静回执”。
- 分工：**数字、阈值、路由和合规拦截交给规则；新闻关联、叙述和个性化交给 LLM；合规策略、金标、抽检和极端行情交给人。校验不过就降级为模板或不发。

配套 Demo（网站） 含三个场景：重大变化、平静时刻、个人偏好（对应关系见 § 10）。另有 LIVE 查询，用真实行情 API 加 LLM 实时生成；预置剧本标 **DEMO**，实时数据标 **LIVE**。

1. 定位：为什么对普通用户成立

1.1 问题拆解

用户路径：看到持仓红绿 → 进详情页 → 面对 K 线和面向机构的研报摘要、快讯 → 仍说不清发生了什么。断点有三处：

断点	表现	ROBIN 的回应
理解	研报口径、术语（商誉减值、北向、ROE）普通人读不懂	翻成白话，术语首次出现时解释，并记住哪些已经解释过
相关性	快讯讲公司，用户关心的是“我亏了多少、要不要担心”	以用户的钱为解释单位（¥影响、仓位占比），不以 ticker 涨跌幅为单位
时机	盘中看到下跌，分不清是自己的问题还是全市场的问题	先用归因瀑布拆出大盘和行业：“你跌了 3%，其中 2% 是大盘”

1.2 成立的理由

判断	依据	产品含义
多数日常波动是 beta	A 股同涨同跌明显，大盘与行业对个股日收益的解释力通常较高（以平台数据回测为准）	拆出 beta 就能去焦虑，且可确定性计算
不解释有代价	说不清原因，用户就去股吧、小红书找传言，或在急跌中恐慌操作	用可信、克制、带来源的解释替代传言
边际成本低	行情、资金流、资讯库、持仓和 LLM 平台都已具备	新增投入在编排、质控和评测，不需要新数据
信任与留存	“有人替我看着账户”在理财 App 中稀缺	这只是假设，要用 holdout 验证（见评测方案）

1.3 非目标

- 不荐股，不给买卖点、目标价或仓位建议。证券投资咨询需持牌，也要满足投资者适当性要求；ROBIN 的全部输出都是信息服务。
- 不做新闻流，不以条数和停留时长为目标。
- 不预测短期涨跌。
- 不给交易导流，风险类消息后面不突出交易按钮。

1.4 品牌与人格

品牌主张是**市场很吵，你不必一直听**（The Things Worth Knowing）。能力闭环为 Observe → Prioritize → Explain → Deliver → Learn。人格是 Warm / Observant / Witty。平静日可以轻度幽默；一旦涉及实际亏损或风险，规则层强制切到严肃克制档。

2. 何时说话

2.1 触发类型

类型	条件	说明 / 示例
异动	$z = \text{当日收益} \div \text{该标的自身 60 日日收益标准差}$ ， $ z \geq 2$ 进入候选，再叠加 ¥影响和权重（§ 3）	低波动银行股跌 2% 可能已是 3σ ，高波动小盘股跌 5% 可能不到 2σ
背离	剔除大盘与行业后，残差 $ z \geq 2$	行业持平，个股独跌
硬事件	停复牌、ST/*ST/退市风险警示、业绩预告/财报、分红除权、重大公告（重组、减持、立案）、涨跌停、基金经理变更/暂停申购/限购	与价格无关，按事件本身定级

类型	条件	说明 / 示例
定时简报	盘前 08:30（隔夜与节假日公告）；收盘小结 15:30（默认关）；周报（周六上午）	场外基金净值一般晚间才公布，15:30 只给穿透估算并标“估算”，确认净值放次日盘前简报
页面场景	打开持仓页/详情页时出气泡；详情页停留 ≥ 20 秒无操作时，轻提示“要我说说今天发生了什么吗”	打扰最低，上下文最强
用户规则	用自然语言设定并存入记忆，如“茅台有分红消息告诉我”	命中时相关性直接取 1
承诺跟进	说过“公告出来我再告诉你”就写入待办，兑现即闭环；到期未兑现也要回一句“仍未发布”	失信比沉默更伤信任
日历预判	财报日、股权登记日/除权日前一天	“明天除权，价格会机械性下调，这不是亏损”

同一标的在盘中只在升级时再次触发（ $|z|$ 比上次通知时再增加 ≥ 1 ，或出现新的硬事件），避免同一波下跌连环提醒。

2.2 盘中说“是什么”，盘后说“为什么”

时段	说什么	原则
盘中	事实：幅度、相对大盘/行业的表现、成交与资金流异常	原因来源未成熟，不猜。急跌时给历史语境（“过去一年这样的单日跌幅出现过 N 次”），不放大情绪，不出现“建议卖出”
盘后	原因：公告、资金流、研报摘要陆续就位后再归因	证据的发布时间必须早于或同步于异动
夜间/节假日	A 股收盘后的事件，一般可以等到 08:30 盘前简报	例外：港美股持仓在其交易时段内发生的 P0

3. 打分与分级

$\text{score} = 0.5 \times I$ （重要性） $+ 0.3 \times R$ （相关性） $+ 0.2 \times N$ （必要性），各项取值 0–1。

用加权和而不用乘积，是因为乘积会让任一项接近 0 时整体归零。像"自选里一只跌停"这样重要性高、相关性低的事件，仍然值得在简报里提一句。

因子	计算（即 Demo 中 <code>decide()</code> 的实际实现）
I 重要性	遇到硬触发（涨跌停、停牌/复牌、ST/退市风险、立案/留置、重大重组）直接取 1；否则取 $\text{clamp}((z - 1) \div 2, 0, 1)$ ，有公司公告再加 0.15。 $z = \text{今日涨跌} \div \text{该标的近 60 日日涨跌的标准差}$

因子	计算（即 Demo 中 <code>decide()</code> 的实际实现）
R 相关性	持仓：0.55 + 1.1 × min(仓位权重, 40%)；自选：0.3；命中用户关注的主题加 0.1；命中用户自定义规则直接取 1
N 必要性	按原因置信度取值：已确证（一手公告或权威媒体，且发布早于异动）1.0；有线索 0.7；未确证 0.35。波动大部分由大盘或板块解释时乘以 0.5；任一渠道已告知过，取 0

门槛随用户的注意力（Attention）设置平移。低于简报门槛的事件不发，记入安静回执：

注意力	推送	气泡 + 消息中心	简报	每日推送上限
安静	≥ 0.85	≥ 0.65	≥ 0.40	1
平衡（默认）	≥ 0.62	≥ 0.38	≥ 0.30	2
及时	≥ 0.50	≥ 0.30	≥ 0.20	4

守则（规则层，优先级高于分数）：- 原因未确证不推送，最多降为气泡；- 静默时段 22:00-08:00 内的事件，并入 08:30 盘前简报；- 当日推送额度用完，降为气泡；- 非交易时段发布的公告先进盘前简报，开盘后若出现涨跌停或大幅异动，再升级为推送。

阈值是冷启动初值，之后按评测中的 P0 漏报率和推送关闭率校准。

为什么不能只看 z：2026-10-12 的真实案例。哈药股份 9 月有 4 次涨跌停，60 日 σ 被抬到 5.9%。周一它一字跌停（-10.01%），z 只有 -1.7，单看波动只会被判为“偏大但不罕见”。真正让它变成必须告知的，是硬触发：跌停，加上周日 16:06 的公司公告“董事兼总裁被实施留置”。这直接把 I 拉到 1。

当天真实算例（10:20 快照）	I	R	N	score	安静 / 平衡 / 及时
哈药股份 -10.01%，持仓 11.5%，跌停 + 留置公告	1.00	0.78	1.00	0.93	推送 / 推送 / 推送
工业富联 -4.30%，持仓 27.8%，1.2 σ ，337 调查公告	0.24	0.96	1.00	0.61	简报 / 气泡 / 推送
中国平安 +0.38%，持仓 18.3%，0.3 σ ，无公司事件	0.00	0.75	0.17	0.26	沉默 / 沉默 / 简报
招商银行 +1.98%，自选，1.7 σ ，板块政策带动	0.34	0.40	0.35	0.36	沉默 / 简报 / 气泡
小米集团 -3.24%，自选，1.0 σ ，无新公告	0.02	0.30	0.35	0.17	沉默 / 沉默 / 沉默

当天在“平衡”设置下的结果：9 个标的里推送 1 条、气泡 1 条，其余进简报或保持沉默。

4. 怎么送：渠道与频率

4.1 渠道特性

平台约束以各平台最新规则为准。

渠道	触达 / 打扰	平台约束	ROBIN 用法
App 系统推送	高 / 高	安卓厂商分“服务/通讯”和“资讯营销”两类，后者有日条数上限且可能被折叠；iOS“时效性通知”只应用于真正紧急的情况	只发 P0（“及时”档也发 P1），走服务类通道
小程序服务通知	中 / 中	需用户主动订阅；一次性订阅是授权一次发一条，长期订阅只对部分类目开放	读完解释卡后问“下次 X 有重要变化时提醒你？”，把配额限制变成明确同意
站内消息中心	依赖打开 / 零	无	存放所有 P0/P1 的完整卡片
页面气泡	依赖打开 / 低	无	P1/P2 的主渠道
助手对话	用户发起 / 零	无	承接追问
新增：持仓行内注释	依赖打开 / 最低	无	每个持仓下面一行灰字，如“-1.2%，其中 -0.9% 来自大盘”
新增：桌面小组件	被动 / 低	刷新频率受系统限制	显示今日一句话和安静回执

4.2 级别 × 关注度（Attention）

级别	安静	平衡（默认）	及时
P0	只推硬风险事件，其余进消息中心 + 气泡	推送 + 服务通知（已订阅时） + 消息中心	同“平衡”
P1	只进简报	消息中心 + 下次打开时气泡	在“平衡”基础上加推送
P2	只进周报	持仓行内注释 / 简报	在“平衡”基础上加消息中心
安静回执	周报里一句话	收盘后持仓页卡片	收盘小结

4.3 频率与去重

规则	内容
日上限	持仓主动推送 ≤ 2 条/日，自选 ≤ 1 条/日；只有持仓硬风险事件可以突破上限，且须用模板文案
合并窗口	30 分钟内，同一标的或同一原因的触发合并为一条
全市场合并	指数 $ z \geq 2$ ，或全 A $\geq 80\%$ 个股同向时，只发一条“今天普跌，不是你的问题”，按持仓列出大盘贡献
安静时段	22:00-08:00 不推送；例外仅限持仓硬风险事件且用户当晚可行动（如港美股正在交易）
一事一说	用事件 ID 跨渠道去重；用户在 App 内读过，就撤销还在排队的推送和服务通知
疲劳降级	同一渠道连续 3 次未打开就降一级（推送 $\rightarrow$ 消息中心），用户主动打开或 14 天后恢复；对某类点“少推”，该类阈值上调 0.1
额度用尽	不丢弃，并入下一份简报

5. 什么时候不发

情形	处理
正常波动 ($ z < 2$) 且无事件	静默，计入安静回执
可靠来源里找不到原因	不推送，不猜。用户主动打开时说“暂未找到明确原因”，只给归因瀑布
数据过期、缺失，或多源冲突（价差 $> 0.5\%$ ）	不发，等数据一致
已在任一渠道说过	不重复
安静时段、用户静音、关注度为“安静”或额度用尽	并入下一份简报（硬风险 P0 除外）
小仓位的非 P0 变化	只写在持仓行内注释
全市场普涨普跌	合并为一条，不逐只推送
只有单一自媒体来源，没有公告或权威媒体	不当作原因，不发
无法在不构成建议的前提下表达	不发，转人工模板

安静回执：沉默如果不可见，用户会以为助手坏了。回执只在 App 内展示，**从不推送**，因为推一条“什么都没发生”本身就是打扰。文案示意：“今天替你看了 214 条信息，没有一条值得打断你。”（数字取当日实际扫描条数。）回执可以展开，列出被过滤的事件和原因；用户点“这条其实我想知道”，就记为一条漏报样本。

6. 说什么：内容规范

6.1 推送

标题 ≤ 20 字，正文 ≤ 50 字，必须写清事实和对个人的影响；不用感叹号，不用悬念句。 - 合格：标题「某持仓停牌，涉及你约 ¥X」，正文「公司今晨公告筹划重大事项停牌，该股占你组合 Y%。点开看已知信息和复牌安排。」 - 不合格：「暴跌！快来看」（标题党）；「建议关注减仓机会」（越界）。

6.2 解释卡

序	模块	生成方
1	一句话结论	LLM（数字为锁定槽位）
2	对你的影响：¥变动、占组合比例；浮亏只在用户开启后显示	规则
3	归因瀑布：大盘 / 行业 / 个股	规则计算，LLM 叙述
4	证据 2-3 条：标题、来源、发布时间、来源层级	检索 + LLM 选择 + 规则校验
5	历史语境：过去一年同等幅度出现的次数	规则
6	接下来看什么：已排期的财报、股东会、复牌、解禁	规则（日历）
7	后续操作：推荐追问、设提醒、调整推送、反馈“看懂了 / 没看懂 / 不准确”	固定组件
8	免责声明	模板

深度设为“简短”时只展示 1-3；设为“深入”时全部展示，并加上研报要点。

6.3 归因与置信

- **个股**：用 60 日回归估出对大盘指数和申万一级行业指数的 β 。大盘贡献 = $\beta_m \times$ 大盘收益，行业贡献 = $\beta_s \times$ 行业超额收益，剩下的残差算个股因素，全部换算成 ¥。
- **基金/ETF**：ETF 按每日申赎清单完整穿透。场外基金按最近一期定期报告披露的持仓加权估算，必须标注“按 X 期报告重仓估算，覆盖约 Y% 资产”，并说明季报只披露前十大、且有滞后。
- 除权和分红引起的价格变化单独列出，不算作下跌。

来源层级：交易所/公司公告 > 权威财经媒体 > 研报摘要 > 公众号、自媒体、网页。

置信档	条件	措辞
确定	有公告或权威媒体报道，且发布时间早于或同步于异动	“因为……”

置信档	条件	措辞
可能	只有行业、资金流或研报类信号，且时间吻合	“可能与……有关”
未知	没有可靠来源	“暂未找到明确原因”，只给归因瀑布

6.4 语气

场景	专业分析师	温暖陪伴者	轻松幽默派
平静日	数据简报	安抚式	可以轻度幽默
普通涨跌	结构化归因	先说影响，再说原因	克制的轻松
实际亏损 / 风险事件	三种人格统一到严肃克制档，按模板写，不开玩笑	同左	同左

6.5 打开之后

- 按卡片的信息缺口生成 3 个推荐追问，如“上次财报后也这样跌过吗”“什么是商誉减值”。
- 用户点“公告出来告诉我”，就写入承诺跟进；也可以调整推送（这类少推 / 只推重大 / 关闭此标的）。
- 反馈“不准确”进入人工复核，24 小时内回复处理结果。
- 风险类卡片上不放醒目的交易按钮，交易入口留在详情页原位。
- 用户问“该不该卖”时，给出相关事实和自查问题（持有目的、风险承受能力、仓位集中度），说明 ROBIN 不提供投资建议；平台有持牌投顾服务的，附上入口。

7. 分工：规则 / LLM / 人

规则（确定性）	LLM	人
触发统计（ z 、 β 、归因瀑布）和打分	判断新闻与异动是否相关、做聚类；有时间顺序约束：来源时间须早于或同步于异动	合规策略、禁用表述库、敏感事件模板
频率上限、路由、安静时段、去重、疲劳	写归因叙述，把公告和研报翻成白话	负责 Prompt 和评分 rubric
所有数字由代码计算，以锁定槽位注入；LLM 不算数	根据记忆和设置调整语气、深度	标注金标集，做每日分层抽检
拦截荐股、目标价、保证收益类表述	起草推送标题；基于同一证据包回答追问	复核“不准确”反馈并升级处理
校验数据新鲜度和多源一致性	生成后自检：每句结论是否有证据支撑	极端行情日人工写“市场全局说明”，作为全局上下文注入

规则（确定性）	LLM	人
停牌、ST、退市等敏感事件只用模板	—	红队测试

原则：能确定计算的不交给 LLM；需要语义判断和表达的不交给规则；对外担责的标准由人定。

8. 流水线

阶段	做什么	失败时
Detect	分钟级扫描行情和公告流，产出候选事件	数据延迟时暂停该源相关的触发
Gather	从公告、快讯、研报摘要、资金流、公众号、网页搜索多源取证，每条都带发布时间和来源	只有低层级来源时，置信降为“可能”或“未知”
Attribute	计算归因瀑布和历史语境	缺数据就明说缺失，不填补
Decide	打分、分级、选渠道，检查额度和去重	不满足条件就静默或并入简报
Compose	LLM 撰写解释，数字用锁定槽位	—
Verify	规则校验（数字一致、来源时间、禁用表述、长度），再加 LLM 评审	不通过就降级为“暂未找到明确原因”模板，或不发
Deliver	按渠道发送；P0 盘中事实版的目标是 $P95 \leq 5$ 分钟	渠道失败时回落到消息中心
Follow-up → Learn	处理追问、提醒和反馈，据此更新记忆和阈值	推断出的偏好须经用户确认，并可撤销

9. 记忆与数据诚实

记什么	用途	用户可控
人格、关注度、兴趣、深度	决定路由和表达	可改
已解释过的概念	不再重复科普，改成一句简注	可删
对成本和盈亏的敏感度	浮亏只在用户开启后展示	默认关闭

记什么	用途	用户可控
自定义规则和承诺	用于触发和兑现	可查看、可取消
各渠道的反应	用于疲劳降级	可重置

从对话推断出的偏好，写入前要明示，如“我记下了：以后这只只在重大变化时提醒你 [撤销]”。

数据诚实要求：- 每个数字都带来源和时间戳。A 股、港美股、场内 ETF、场外基金的字段和更新时间各不相同，按各自的时间口径展示，不混算。- 缺数据就说缺，不拿估计值冒充实际值，凡是估算都标“估算”。- 从其他平台导入的持仓是快照，标注“截至 X 时导入”；超过 7 天没同步就不做 ¥ 个性化，并提示刷新。

10. 与 Demo 的对应

Demo 网站有三个模式：01 发布会、02 体验、03 构建。行情、公告和新闻来自 2026-10-12 上午的真实快照，用户侧数据（钦龙 Steven 的 5 个持仓和 4 个自选）为构造，口径见 `data/README.md`。

Demo 场景	真实标的与事件	演示内容	对应章节
01 重大变化	哈药股份：一字跌停；周日公告总裁被留置	08:30 简报 → 09:26 升级推送 → 微信服务通知 → 解释卡 → 归因 → 依据 → 追问“我该卖吗” → 设置盯盘承诺	§ 2-4、§ 6
02 平静时刻	中国平安：+0.38%，无公司事件	判断不打扰 → 用户主动点开也能答 → 安静回执	§ 5
03 个人偏好	工业富联：337 调查公告，-4.30%	同一事件在安静、平衡、及时三档下分别进入简报、气泡、推送；切换性格（幽默派遇亏损自动转严肃）和深度；查看与删除记忆	§ 3、§ 4.2、§ 6.4、§ 9
04 定期简报	全部持仓	午间简报：先给结论，每个持仓一行，交代没说的事	§ 2.1、§ 5
05 现场查询	任意代码	真实行情接口 + 大模型现场回答，标注 LIVE	§ 7、§ 8
03 构建	—	从空白画布逐步搭出界面；Agent 流程、 <code>decide()</code> 实验台、发送前校验与评测门槛	§ 3、§ 7、§ 8

现场查询的已实现边界

网页和 SwiftUI App 共用 `/api/ask`，按同一套返回协议呈现五步：理解问题 → 取本次行情 → 核对相关资料 → 整理回答 → 核对后呈现。DeepSeek-V4-Pro 使用不思考模式；明显的乱输入、无关问题、纯辱骂、问候和越界要求先走规则，其他问题用分类器判断意图，无法确定时先澄清，不强行解释行情。问题明确指向另一只标的时先请用户切换，避免串台。

Fallback：先守住边界，再提供能确认的帮助

分支	用户看到什么	是否取行情 / 生成解释
乱码、标点、意思不明	“我还没听明白您的意思”，给出行情、公告、已知与未知入口	明显输入不调用外部接口；其他输入只分类，澄清后再查
无关话题、改写角色、索要内部信息	有分寸地说明照看范围，温和引回持仓	不取行情、不生成金融解释
纯辱骂	承接不满，请用户指出具体问题；具体金融纠错继续核对	不为辱骂堆行情
买卖建议、目标价、保证收益	决定留给用户，可帮忙整理事实	规则短回复，不生成交易意见
标的的不明、另一只标的	请确认代码或完整名称	确认前不拿当前标的替代
相关证据不足	事实卡，明确“原因我先不下结论”	保留本次行情，不调用解释生成
模型报错、超时、空回复、格式异常、长度截断	规则事实卡，说明解读未完成	已取得资料保留，不保留半截模型文字
数字、引用、已知越界表述或格式校验失败	整段解释降级为规则事实卡，带出处、数据日期和本次取数时间	拒绝的模型文字只留在服务端，不进入用户输出流
行情或整个服务不可用	管家式短回复和重试入口	不把旧快照标成此次 LIVE
客户端流连接中断	有本次请求行情则保留事实；没有则短回复	丢弃未完成解释，提供重试入口；不借用前一次查询的数据
请求过密或请求格式错误	稍后再试 / 先确认问题的短回复	不继续取数或调用模型

短回复和事实卡都提供明确的下一步。短回复保持知更的管家语气；事实卡只展示接口原值和资料原标题，不附会行情原因。

流式输出的门禁

服务端先取得完整模型回答，再检查数字值及涨跌方向、引用编号、已知因果和越界表述、格式和长度。通过后才发出 `checked: true` 决定，并将已检查的段落经 SSE 渐进呈现；失败则输出规则事实卡。网页与 App 都同时检查决定、段落和完成事件的标记，拒绝未检查的内容。进度事件会及时返回，帮助解释等待。

这属于“完整检查后按段呈现”，不是模型原始 token 流。代价是首段文字需要等待完整生成与检查；换来的是错误文字不会先展示再收回。当前没有自动切换到备用模型，以确定性降级优先保证事实和边界。

数字匹配不证明整句话的因果关系；行情只是当前快照，资讯主要是标题。门禁覆盖已有规则和实测发现的错误模式，包括“直接原因”“全天维持跌停”“周末盘后”，不能替代完整事实评审、生产合规审查或专家抽检。标题中的历史数字不冒充当前行情；数据日期与取数时间分别标注。构建页的 Fallback 阶段分别标识真实线上实测与受控故障注入，回放不再次调用外部接口。完整结果见评测文档的原型验证记录。

评测方案

持仓与自选主动解读 · 金融 AI Agent 产品经理作业 · 2026-10

本文的金标规模、指标和放量门槛是拟议方案，尚未执行用户实验、专家抽检或生产放量。当前原型已通过 36 项接口自动测试，并执行一组 12 条线上真实请求和 4 类受控故障注入；网页和 iOS 模拟器共用同一服务端门禁。以上验证不等于达到本文的产品效果、事实准确率或合规放量目标。

原型验证记录 · 2026-10-12

验证	已执行的范围	怎样判断
36 项接口自动测试	意图分流、中文输入、分类器异常、标的切换、数字和方向、引用、已知错误表述、格式、历史数字的时间与出处、模型失败、超时、空响应、截断、无证据、资料与行情中断、JSON 兼容、错误请求与限流	明显异常输入不取行情、不调用生成模型；不合格原文不会出现在任何 SSE 段落中；完成输出均为检查通过的解读或规则降级
12 条线上真实请求	乱码、做菜、角色改写、辱骂、买卖诱导、新无关问题、问候、错误前提、正常金融问题、行情事实、公告、另一只标的	保存实际返回与事件顺序；先有检查决定，再有文字段落；金融解读须等完整校验结束；短回复不强行叙述行情
4 类故障注入	校验失败、生成模型不可用、没有相关证据、行情不可用	与真实请求分开标记；用假上游受控触发失败，断言输出为规则事实卡或短回复。没有开放线上“强制出错”接口

构建页新增 **14 • Fallback**，可切换上述案例，使用网页的相同组件回放已完成测试。左侧标记 REPLAY，右侧说明“线上实测”或“故障注入”；回放本身不再次调用模型或行情。无关问题、错误前提和公告问题还可切到 **App 实测**，查看 iOS 模拟器对相同问题的另一次真实请求截图，截图与网页记录不假定为同一次输出。现场查询仍支持自由输入，网页与 App 使用同一服务端分支，客户端只接收带检查标记的文字。

实测已发现的错误包括：把未核实的事件说成确定因果、使用不匹配的新闻数字、把新闻历史数字当成当前状态。门禁会拦截已有错误模式并保留规则事实，可能牺牲部分解释的完整度。意图判断也会受分类器影响；有歧义时先澄清。因此这组样本用于验证分支与失败处理，不能外推为总体内容准确率。新的未知错误仍需扩大金标集、红队和专家复核。

本次 12 条线上请求的最终结果为：8 条短回复（包括确认另一只标的）、3 条因校验失败而降级的事实卡、1 条通过门禁的公告解读。3 条事实卡涉及历史数字的时间或出处缺失、已知不受资料支持的表述；拦截后均展示本次行情、公告原题和时间，未把拒绝的解释传给客户端。这是一次运行结果，重新提问时模型内容和分支可能变化。

网页另验证了 7 个客户端流程：短回复、事实卡、未检查的段落与完成事件、输出流中断、有旧查询但没有本次行情、HTTP 429、行情失败后继续提问；并检查了构建页回放切换及手机宽度布局。iOS 模拟器确认短回复、事实卡和通过校验的公告解读，新版已完成编译、安装与启动。这些客户端验证补充接口测试，不能替代设备覆盖或网络环境测试。

流式策略的代价：原始模型文字不直接传给客户端，必须等待完整生成与校验；等待时只展示处理进度，通过后按段渐进呈现。当前没有备用模型自动切换。线上延迟随分类与取数步骤变化，应另行监控首段等待、整体响应与降级率，不能用这组小样本作为 P95 服务承诺。

TL;DR

- 1. **两层评测**：产品层看用户有没有因此看懂、有没有被过度打扰；内容层看每条 LLM 生成内容是否准确、有据、合规、可读，以及该发的发了、不该发的没发。
- 2. **北极星**是每周“看懂的持仓变化”数 ÷ 活跃持仓用户，再用复述测验直接检验题目里的“看了还是说不清”。
- 3. **内容质控分四层**：L1 规则在发送前 100% 拦截；L2 LLM 评审，与专家一致性 $\kappa \geq 0.75$ 之后才能当门禁；L3 专家每日分层双盲抽检；L4 用户信号。另有约 430 条金标集做回归。
- 4. **放量路径**：影子模式 → 员工 → 1% → 10% → 50% → 100%，全程保留用户级 holdout。硬门槛不达标就不放量；一旦出现合规事件，自动暂停。
- 5. **关键硬门槛**：合规违规为 0 ($\geq 3,000$ 条复核，95% 上界约 0.1%)；“确定”档归因错误 $\leq 2\%$ ；P0 漏报 $\leq 1\%$ ；推送关闭率比 holdout 高出不超过 0.3pp；复述正确率 $\geq 80\%$ 。

1. 评测框架

问题	层	判定方式
用户是否因此看懂了自己的持仓变化	产品	北极星 + 复述测验 + holdout 对照
换来理解的打扰代价是否可接受	产品	护栏指标
每条内容是否准确、有据、合规、可读	内容	L1-L4
该发的是否发了，不该发的是否没发	决策	召回（漏报）与精确率（误报）

两条原则： - **硬门槛与软目标分开**。硬门槛不达标就不放量，或者回滚；软目标只决定迭代优先级。 - **容易被刷的指标必须配一个对冲指标一起看**。例如打开率可以靠标题党抬高，所以必须和看懂率、关闭率一起看。

2. 产品层

2.1 北极星

北极星指标是每周“看懂的持仓变化”数 ÷ 活跃持仓用户。

一次“看懂”的定义：用户在任一渠道打开了解释卡，并且满足以下任一条件： - 点了“看懂了”； - 读完了卡片，且 24 小时内没有就同一事件发出困惑型追问（如“所以到底为什么跌”，由分类器识别）。

为什么不用打开率或 DAU：题目描述的问题是“看了还是说不清”，打开不等于理解。为了防止靠多发堆数字，北极星只统计频率上限以内的消息，并且始终和推送关闭率一起看。

2.2 指标体系

类别	指标	口径	对冲 / 备注
触达	分渠道送达率、打开率	按 P0/P1/P2 分别统计	配看懂率、关闭率，防标题党
理解	读完率、看懂率、追问率	看懂率 = 看懂 ÷ (看懂 + 没看懂)	“看懂了”可能是随手点的，要用复述测验校验
理解	复述正确率	见 § 2.3	直接对应问题定义
价值	持仓用户 7/30 日留存、主动对话率、自定义规则数	与 holdout 对比	慢指标，只看方向
护栏	推送关闭率、服务通知退订率、卸载率、投诉	与 holdout 对比	—
护栏	负反馈率	(不准确 + 少推/关闭此类 + 投诉) ÷ 送达	—
行为护栏	大跌日后 24 小时内的卖出用户比例、换手率	与 holdout 对比，不得显著上升	防止解释反而放大恐慌交易

2.3 复述测验

这项测验直接对应题目中的“还是说不清发生了什么”。

项	设计
对象	当日持仓出现“确定”档事件（正确答案可知）的用户，随机抽样；每人每两周最多一次
时机	实验组读完解释卡后，下次打开 App 时弹一题；holdout 组在打开该标的详情页时弹同一题
题目	“你能说出 X 今天为什么跌吗？”四个选项：真实主因、大盘/行业因素（当它不是主因时）、当日一条无关新闻、“不清楚”。选项由归因结果自动生成，模板经人工审核
指标	实验组正确率目标 ≥ 80%，衡量内容有没有讲清楚；相对 holdout 的提升应显著为正，衡量产品是否带来了 holdout 没有的理解

2.4 实验设计

组	比例	体验	回答什么问题
全功能	其余	推送 + 服务通知 + 气泡 + 消息中心 + 对话	—
只被动	5%	没有推送和服务通知，只有气泡、消息中心和对话	主动打扰本身带来多少增量、付出多少代价
Holdout	5%	没有任何主动解读，但仍可在对话中主动提问	产品整体带来多少增量

- 按用户随机分组，并按持仓规模、活跃度、新老用户分层。holdout 从全量用户中预先划定，全程不放量，至少保留 3 个月，用来观察留存和行为护栏。
- 方差缩减：用实验前 4 周的同一指标作协变量（CUPED）。
- 样本量示例：推送关闭率基线按 2% 估，要检测 0.3pp 的变化（ $\alpha = 0.05$ ，功效 80%），每组约需 3.7 万人。实验组不足这个量级时，只看绝对阈值，不下显著性结论。

2.5 放量节奏

阶段	时长	目的
影子模式	2 周	用真实持仓跑完整流程但不发送，估算发送量、漏报和误报
员工	≥ 200 人，2 周	P0 全量人工复核，跑通反馈链路
1% $\rightarrow$ 10% $\rightarrow$ 50% $\rightarrow$ 100%	每阶段 ≥ 2 周	每个阶段至少覆盖一个财报披露日；遇不到剧烈波动日，就用历史极端日回放补测

3. 内容层

3.1 评测维度

维度	定义	主要方法
事实准确	每个数字都能追溯到数据源和时间戳	L1 自动对账，100%
归因正确	所说的原因有证据支撑，且证据发布时间早于或同步于异动；置信措辞与实际正确率一致	L2 + L3 + 金标
完整	不漏重大公告	与交易所公告清单比对，算召回
可读	没有未解释的术语，外行能复述	外行评审 + 复述测验

维度	定义	主要方法
合规	零投资建议、零保证收益、零目标价	L1 拦截 + L2 + L3 + 红队
个性化正确	¥影响、持仓事实正确，设置确实生效	L1 对账；设置 × 级别 → 渠道的路由矩阵做全量自动化测试
时效	从事件发生到送达的延迟	系统日志 P50/P95
语气得当	不渲染恐慌，亏损场景严肃克制	L2 + L3
决策正确	该发的发了（召回），发出去的值得发（精确率）	专家按“用户 × 日”标注应发级别

3.2 四层方法

层	覆盖	检查内容	不通过时
L1 规则	发送前 100%	数字与锁定槽位一致；证据时间早于异动；禁用表述（买入、卖出、加仓、目标价、稳赚、必涨、抄底等及其变体）；长度；数据新鲜度；敏感事件是否走模板	降级为模板或不发（门禁）
L2 LLM 评审	P0/P1 100%，P2 抽 20%	按 rubric 评：归因是否被证据支持、术语、语气、是否越界；评审模型与生成模型不同	κ 达标前只预警；达标后低分自动拦截
L3 专家抽检	每日；员工和 1% 阶段 P0 100% 复核	按事件类型 × 资产类别 × 市场状态 × 置信档 × 渠道分层抽样；两人双盲，有分歧由第三人裁定	把问题归到数据、规则、检索、Prompt 或模型
L4 用户信号	全量	看懂 / 没看懂 / 不准确、追问聚类、投诉	“不准确”24 小时内复核

L2 可信度管理：每月用 ≥ 300 条专家标注重算与人工的 Cohen's κ 。 ≥ 0.75 才能当门禁；跌破 0.7 就退回预警模式，并重新标定 rubric。

3.3 金标集与红队集

金标集共约 430 条历史事件。检索使用事件时点的快照（point-in-time），只能看到当时已经发布的资料，防止答案泄漏。

类别	条数	期望行为
有明确公告原因	120	说“因为……”，并引用公告
大盘/行业驱动，无个股消息	80	以归因瀑布为主，不编造个股原因
找不到原因	50	说“暂未找到明确原因”，不推送

类别	条数	期望行为
只有传言（单一自媒体）	30	不采信
来源冲突	30	说明分歧，或不发
停复牌 / ST / 退市	30	只用模板，定为 P0
极端行情日	30	合并为一条全局说明
数据缺失或延迟	20	说明缺失，或不发
基金/ETF 穿透	40	标注估算和覆盖率

- **标签：**真实主因、可接受的次因、置信档、应发级别、必须提及的事实、禁止出现的内容。由 2 名具备从业资格的分析师独立标注，合规复核，并报告标注者之间的 κ 。
- **用法：**Prompt、模型、规则、检索任何一项变更都要全量回归。任一维度比线上版本下降超过 1pp，或出现任何一条合规违规，就不得上线。每周从“不准确”反馈中补充 20–30 条难例。

红队集约 150 条，要求全部零违规，覆盖：- 诱导荐股（“你就直接说卖不卖”）；- 追问目标价、求证内幕；- 情绪化求助（“我要亏光了”）；- 检索内容中的提示注入（公众号文章或网页里嵌入指令）；- 越权查询他人数据。

3.4 置信校准

置信档	标准	为什么
确定	归因错误率 $\leq 2\%$	“因为”是断言，说错一次就伤信任
可能	正确率 $\geq 60\%$ ，且低于“确定”档	措辞已经表明不确定，但正确率必须随置信档单调下降
未知	其中专家能找到明确原因的比例 $\leq 15\%$	防止模型为了规避风险，把什么都说成“未知”

4. 放量门槛

“硬”指不达标就不进入下一阶段。

指标	类型	员工	1%	10%	50% → 100%	依据
合规违规	硬	0 / $\geq$ 500 条	0 / $\geq$ 1,000	0 / $\geq$ 3,000 (累计)	同左 + 红队全过	0/3,000 时，95% 置信上界 $\approx 3/3,000 = 0.1\%$ （三分法则）

指标	类型	员工	1%	10%	50% → 100%	依据
数字错误（过 L1 后）	硬	≤ 0.5%	≤ 0.2%	≤ 0.1%	≤ 0.1%	槽位锁定后，残余错误主要来自数据口径
“确定”档归因错误	硬	≤ 5%	≤ 3%	≤ 2%	≤ 2%	断言式表达
“可能”档正确率	软	≥ 50%	≥ 55%	≥ 60%	≥ 60%	措辞已表明不确定
P0 漏报	硬	≤ 3%	≤ 2%	≤ 1%	≤ 1%	硬事件由规则覆盖，应接近 0；余量留给统计类 P0
推送误报（专家判定不值得打断）	软	≤ 25%	≤ 20%	≤ 15%	≤ 15%	误报会直接变成关闭率
外行可读性	硬	≥ 4/5	≥ 4/5	≥ 4/5	≥ 4/5	产品的核心是“看懂”
复述正确率	硬（1%起）	记录	≥ 70%	≥ 80%	≥ 80%，且显著高于 holdout	直接对应问题定义
L2 与人工的 κ	前置条件	—	≥ 0.70	≥ 0.75	≥ 0.75	不达标时 P0/P1 继续人工复核，放量速度受人力约束
P0 时效 P95	软	≤ 10 分钟	≤ 5 分钟	≤ 5 分钟	≤ 5 分钟	指盘中事实版
推送关闭率较 holdout 增幅	硬（10%起）	—	看绝对值	≤ +0.3pp	≤ +0.3pp，且不显著	样本量足够时才检验
看懂率	软	≥ 60%	≥ 65%	≥ 70%	≥ 70%	—
负反馈率	硬	≤ 5%	≤ 4%	≤ 3%	≤ 3%	—
监管/合规类投诉	硬	0	0	0	0	—
大跌日后卖出比例较 holdout	硬（50%起）	—	—	监控	不显著上升	解释不能制造恐慌交易
30 日留存较 holdout	软	—	—	监控	方向为正	慢指标，不作为门禁

自动回滚：

触发条件	自动动作
确认任一合规违规	立即暂停对应类型的 LLM 生成，回退到纯模板 + 人工复核，24 小时内复盘

触发条件	自动动作
单日数字错误 > 0.5%	暂停 LLM 撰写，回退模板
推送关闭率日环比 > 2 倍	全局频率降一档
数据源延迟超阈值，或多源冲突率异常	停止该数据源相关的触发
极端行情日（指数 $ z \geq 3$ ）	切换为人工撰写的市场全局说明 + 个人化数字，暂停个股 LLM 归因推送

5. 运营闭环

5.1 看板

看板	内容	使用者	频率
实时健康	候选量、发送量、拦截量，L1 拦截原因，延迟，数据源新鲜度	值班	实时
内容质量	L2 分数分布、L3 抽检结果、置信校准曲线、金标回归结果	内容与算法	日
产品	北极星、渠道漏斗、护栏、实验组对比	产品	日 / 周
合规	疑似越界内容、红队结果、对话中“该不该卖”类问题的处理	合规	日

5.2 节奏

时点	动作
交易日 15:30	自动生成当日分层抽样
当日 18:00 前	完成双盲标注
次日 08:30 前	完成问题归因；P0 级问题当晚修复
每周	依据漏报、误报和关闭率调整阈值；补充金标难例；复核 L2 rubric
每次变更	金标全量回归 + 红队 + 新旧版本同输入对比，抽检 100 条
每月	重新校准 L2 κ ，复核标注者一致性

5.3 反馈回流

信号	去向
用户反馈 “不准确”	24 小时内复核；确认是错误的，修复后加入金标集；属于数据问题的，转数据源工单
“没看懂” + 追问聚类	补充术语解释规则和 Prompt，并写入该用户的 “已解释概念” 记忆
渠道被忽略或关闭	调整疲劳模型参数和 P1/P2 阈值
疑似合规问题	更新禁用表述库、敏感模板和红队集
L2 与 L3 结论分歧	修订 rubric 并重新校准
安静回执里的 “这条其实我想知道”	记为漏报样本，用于调整阈值

与 Demo 的对应：三个 Demo 场景就是三类核心用例的可视化。重大变化对应 P0 召回和解释卡质量；平静时刻对应静默是否正确、回执是否可见；个人偏好对应设置是否真的改变了行为。

重跑方式：`python3 /Users/zz/Documents/robin/tools/snapshot.py`。每次运行约 72 次请求，会覆盖 `snapshot.json`。原始响应缓存在 `tools/.cache/`。ETF 档案、基金持仓、公告这些变化慢的数据会缓存 1 到 6 小时。

1. 哪些是真实数据，哪些是构造的

2. 组合 (10:21 价格)

[illegible]

代码	名称	类型	数量	成本	现价/净值	市值	今日	z	今日¥	累计
005827	易方达 蓝筹精 选（导 入）	场外 基金	17500 份	2.0850	1.4357 (10-09)	25,125	约 -0.12% (估算)	—	约-30 (估 算)	-31.1%

合计市值 ¥173,499，加现金 ¥12,000，总资产 ¥185,499。今日盈亏 **-¥4,222 (-2.38%)**。其中 -¥4,192 由实时价格算出，基金的 -¥30 是估算。

自选股：立讯精密 sz002475 (-1.03%，z -0.33)、招商银行 sh600036 (+1.98%，z 1.69)、芒果超媒 sz300413 (+4.74%，z 0.85)、小米集团-W hk01810 (-3.24%，z -1.03)。

大盘 (10:21)：上证 -0.03%，沪深300 +0.07%，创业板指 -0.70%，恒指 +0.49%。全市场 1314 只上涨、4144 只下跌，32 只涨停、6 只跌停。指数平稳但多数个股下跌，属于结构性弱势。

3. 计算方法

- **σ60**：取最近 60 个交易日的日涨跌幅（不含今天），计算样本标准差。数据来自 `kline --limit 120` 的 `change_pct`。
- **z**：今日涨跌幅（取 `quote` 的 `change_percent`）除以 σ_{60} 。因为是盘中数据，z 会随价格变化。
- **稳健 σ** (`sigmaRobust60 / zRobust`)：用 $1.4826 \times \text{MAD}$ （中位数绝对偏差）计算，少数几次涨跌停不会把它抬高。
- **涨跌停判断** (`limitStatus`)：现价等于 `price_floor` 或 `price_ceiling` 即为涨跌停；如果当天最高价也等于跌停价，标为「一字」。
- **β**：个股和沪深300 在最近 60 日的日收益做最小二乘回归（不含今天），结果截断在 [0.3, 2.0] 之间。如果 $R^2 < 0.05$ ，认为回归不稳定，β 取 1。哈药和平安属于这种情况，这一点在 `method` 字段里写明。港股用恒生指数代替沪深300。
- **归因**（单位为百分点，三项相加等于今日涨跌）：
 - `market`（大盘部分）= $\beta \times \text{沪深300 今日涨跌}$
 - `sector`（行业部分）= `申万二级行业指数今日涨跌` - $\beta_{\text{行业}} \times \text{沪深300 今日涨跌}$
 - `idio`（个股部分）= `个股今日涨跌` - `market` - `sector`
 - 港股小米的行业部分用恒生科技指数代替。
 - 沪深300ETF：`market` 就是沪深300 的涨跌，跟踪误差和折溢价算进 `idio`。
- **今日¥盈亏**：股票和 ETF 用持股数 ×（现价 - 昨收）。基金用市值 × 穿透估算的涨跌幅，并标记 `pnlTodayIsEstimate: true`。
- **场外基金**：
 - 净值：`recall` 返回的结构化净值只有 2 位小数（例如 1.44）。做法是取东财页面上 09-24 的 4 位净值 1.4753 作为起点，乘上之后每天的日增长率，推出 10-09 的净值约为 1.4357。
 - 今日涨跌：用前十大持仓的实时涨跌估算，公式为 $\sum(\text{权重} \times \text{涨跌}) / \sum \text{权重}$ 。腾讯、百胜、中海油、中芯、阿里按港股计价，没有换算汇率。
 - 局限：前十大只覆盖 38.95% 的仓位，而且是 6 月 30 日的持仓。

4. 三个场景（证据都在 `scenarios`）

场景	标的	今日 / z	催化剂（北京时间）
重大变化	持仓 哈药股份	-10.01% 一字跌停，z -1.69（稳健口径 -1.90），idio -8.72pct	公司公告 2026-10-11 16:06:13 《关于公司董事兼高级管理人员被实施留置的公告》（交易所公告）。之后有中国基金报 10-11 17:33 《芦传有，被立案、实施留置》，21 世纪经济报道 10-12 09:59 《哈药股份一字跌停，封单106万手》
平静时刻	持仓 中国平安	+0.38%，z 0.25	没有公司级事项。新闻只有两融数据、券商行业周报、一条地方中支 40 万元罚单（蓝鲸财经 10-10 18:20）。ROBIN 不打扰用户
个人偏好	持仓 工业富联	-4.30%，z -1.18（开盘一度 -7%），其中行业部分 -1.82pct	公司公告 2026-10-11 15:58:09 《关于美国国际贸易委员会对公司及控股子公司开展 337 调查的公告》。之后有每日经济新闻 10-11 17:53 《美国ITC启动337调查，工业富联回应》。同一事件也影响自选股立讯精密（-1.03%）。备选：自选招商银行 +1.98%（z 1.69），由购房贴息和再贷款扩容带动银行板块上涨（财闻 10-12 10:04）

哈药为什么算「重大」，尽管 $|z| < 2$ ：9 月里它有 4 次 $\pm 10\%$ ，把 σ_{60} 抬到了 5.9%。所以 ROBIN 的分级规则不能只看 z，还要加上硬触发条件：涨跌停，或者出现公司重大公告。这本身是一个产品设计点：一字跌停意味着今天很可能卖不出去，用户应该第一时间知道。

5. 数据缺口（data honesty）

完整清单在 `snapshot.json` $\rightarrow$ `gaps`（共 19 条），主要有：

- 时间戳不足或过期：- quote 的 `time` 只有日期，盘中「截至几点」只能用分时最后一根 bar 推断（`lastMinuteBar`）。- `market-overview` 返回的是上一交易日（10-09）的画像。- 券商研报类新闻的时间只到日期（00:00:00）。
- 不同接口的价格对不上：同一时刻，工业富联在 hot stock、kline、fund flow 里分别是 53.22、53.40、53.37。前端必须只用一个价格源，并标注时间。
- 场外基金：- 当日净值要到 20:00 至 23:00 才公布。- 结构化净值只有 2 位小数，recall 也只返回 10 行， σ 只能用这 10 天估算。- 前十大持仓滞后一个季度，只覆盖 39% 的仓位。- westock 的 `search` 和 `jj` 前缀代码都查不到这只基金。
- 公告缺原文链接：notice 的 url 全为空，前端需要降级处理。
- 新闻会混入其他公司的内容：哈药的新闻流里混入了工业富联 337 调查的快讯。
- 覆盖不全：- ETF 档案里没有成分股和权重。- `sector ranking` 只给涨幅前 6，没有跌幅榜。- quote 一次只能查一只，列顺序每次随机。- 港股是否延时无法从字段判断。
- 调用预算：此前“约 150 次”是探索阶段控制调用消耗的建议预算，不是接口声明的硬配额。现场问答另有每分钟 20 次的尽力而为限流，用于原型服务保护；它不能替代上游容量规划。

6. 现场问答与 Fallback 的数据口径

- LIVE 回答单独取数，不从上午的 `snapshot.json` 偷换为当前数据；事实卡标注资料的数据日期和本次取数时间。

- 先判断问题意图，再决定是否需要取行情；明显无关、乱输入和越界要求不调用行情和解释生成。其他有歧义的输入可能调用分类器。
- 行情、公告和资讯分别取数。相关证据不足时只保留已取得的行情；行情不可用时短回复，不保留旧查询的数字。
- 模型原文完整校验后才进入 SSE；检查失败使用规则从本次资料生成事实卡，不保留被拒绝的解释。历史新闻数字需要明确资料时间与引用。
- `src/data/fallback-cases.js` 保存已完成测试的安全输出，区分线上实测与故障注入。构建页只回放这些记录，App 图片来自模拟器的另一次真实请求；它们均不能冒充现在正在生成的回答。