

虚拟人仿真调研 · 方法论与校验白皮书

—— 我们怎么保证结论可用，以及它不能用来做什么

白皮书 v1.1 | 修订日期 2026-10-12 | 对应调研平台版本 V1.6.0 | 文中所有数字均为本平台实测值

两个版本号分别代表什么（先说清，避免误以为本文档过期）：

「白皮书 v1.1」是本文档自己的修订号 —— 方法、边界、校验结果有变化时才升；

「V1.6.0」是调研平台的版本号 —— 每次平台自己的功能上线时升。

⚠ 虚拟城市引擎的升级不记在平台版本号里（那是另一套版本），见《虚拟城市_版本说明》。

两者节奏不同：本文档多数平台版本发布时并不需要改。所以看到"平台已经是 V1.x、白皮书还是 v1.1"是正常的；

判断本文档是否过期，请以"对应调研平台版本"这一行为准（我们在每次实质修订时会同步这一行）。

v1.1 相对 v1.0 的变化：新增第四节《与两类同类方案的区别》（与「固定参数的合成用户」「用真人访谈训练的数字孪生」逐项对照，

并一并写出我们不比别家强的地方）；第三节补充三条边界（不建模社会互动、抓不到意外发现、情感浓度不足）；

并修正了原文档的章节编号缺陷（此前有两节同编号"五、"）。

虚拟城市引擎的一次升级（2026-10-12，不属于平台版本号）：城市补全了全学段教育，

因此平台能读到的市民信息多了就读学段与年级、以及升学关口（引擎在运行中真实记录的原文）。

连同一条必须一并阅读的边界：

在读研究生（硕士、博士）暂不标注 —— 城市记录的是已取得的学位，没有"是否在读"这一项，

因此无法可靠区分「在读」与「已停止求学」；我们没有用估算补这个标签（那会变成编造人口结构）。

目前可可靠标注的是 18 – 21 岁的本科在读。另：城市里若出现高校点位，**一律是模型生成的虚拟院校、

不使用任何真实高校名称**，点位只表示落在该行政区内，不代表真实校园位置。

详见《虚拟城市_版本说明_V1.5.2.md》。

⚠ 本文档的正文数字（第 6 节）仍对应 V1.5.0 的校验结果，尚未重跑；V1.6.0 的实测数字见版本更新说明。

一、这份白皮书回答三个问题

机构客户在评估仿真调研时，问的其实是三件事。本文按这三件事组织，并且每条结论都给出可自行核对的位置。

客户的问题	我们的回答	在哪核对
数据从哪来、像不像真人？	两个样本来源：平台性格库（按你的比例精确构造）与虚拟城市市民（可回查的真人档案）	报告「人群构成」与「样本示例」
结论能不能复现？	同一份参数任何时候重跑，逐题逐项完全一致	报告里的「数据校验」页，当场重跑给你看
哪些事它做不了？	它不是真人回答，不能替代确证性研究；城市里没有的人变不出来；开放题文字是模型生成	本文第三节与第八节

二、它是什么

- 面向机构客户的人群分布模拟器：给定目标人群结构与一份问卷，输出答案分布与交叉分析。
- 两个样本来源，按需选择：
 - ① 平台性格库：按你设定的比例**精确构造**人群（实测各维偏差 **0.0** 个百分点），代价是受访者不是城市里可回查的真人。
 - ② 虚拟城市市民：从正在运行的虚拟城市调集**真实市民**（每位带唯一市民 ID，可核对档案与当日日程），代价是人群比例受城市实际人口约束。
- **三层作答模型**：① 性格主轴（稳定的人格结构）→ ② 特性细分（思维倾向、观察习惯、决策方式、表达习惯、自我约束、受他人影响）→ ③ 人口属性修正。同一份参数，同一个人永远给同一个答案。

说明：作答由确定性函数完成，不依赖任何一次性的模型生成 —— 这是下面「可复现性」的前提。

三、它不是什么（先把边界划清）

我们宁可把话说窄，也不让客户拿去用错。以下每条都会被写进导出文件的口径说明里：

- **不是真人回答**。它是稳定性格结构对问卷的推演分布，不代表任何具体个人的真实意见。
- **不是个体预言**。可以看「这类人会怎么答」，不宜看「张三会怎么答」。
- **不是确证性研究**
。适合做前期假设筛选、概念与文案预测试、方案对比；重大结论仍应以真实访谈与专家研判为准。
- **开放题的文字是模型生成的**，不是任何真实个人的原话 —— 报告中会明确标注，请勿作为真实用户引语引用。
- **城市模式不能突破城市人口上限**：城市成年人里 18 - 24 岁只占 7.5% 时，你把目标设成 30% 也做不到（详见第六节）。
- **不建模社会互动**
：受访者之间不互相影响。涉及从众、群体压力、家庭共用设备、社交传播类题目，本平台**不适合**。
- **抓不到「意外发现」**
：受访者会沿最常见、最合理的路径回答。以"发现意料之外的问题"为目标的探索性研究（如可用性缺陷发现）**不在能力范围内**。
- **情感浓度不足**：情绪、犹豫、前后矛盾这类"真人的毛边"，本平台不模拟，也不声称能模拟。

四、与两类同类方案的区别（为什么「像不像真人」不是唯一的问题）

客户与投资人最常拿两类东西来对比我们。先把三者摆清楚，再说我们强在哪、弱在哪。

	① 固定参数的合成用户	② 用真人访谈训练的「数字孪生」	本平台
受访者从哪来	按提示词 一次性生成 一个画像	用 真人深访数据 训练的个体模型	持续运行的虚拟城市里真实存在的市民
有没有经历	没有经历，只有设定	有真人访谈片段	城市里生活数千天 ：岗位、区县、收入、家庭、健康与一生轨迹
能否回查个体	不能	不能（是聚合模型）	市民 ID 可回查档案与当日日程
同一个人跨次是否一致	每次现编， 可能不同	取决于实现	确定性绑定 ：同一市民 ID 永远同一套性格
同参数跑两次	不保证	强调"改一个条件重推"	逐格一致 （见下）
规模上限	无限（但每个个体没有来源）	受真人访谈量限制	数十万 （城市成年人口）

4.1 我们确实强的地方（有实测数字）

同参数跑两次（各 40 人）：

--	--

检验项	结果
抽到的是 同一批市民	40/40 = 100%
答案 逐格一致	40/40 = 100%
性格文本完全一致	40/40 = 100%
人口维度完全一致	40/40 = 100%
放大到 200 人 再测一次	200/200 = 100% （排除小样本巧合）

个体差异同样是实测的：40 人的 6 道题里 **6/6 道都出现了不止一个选项**，

且 40 个人的**完整答案组合去重后是 40 种**（各人各不相同）。

这意味着别家结构性做不到的两件事：

- ① **同一批人可以跨调研逐人配对**（换一次运行还是同一批市民）——这是面板研究才有的能力；
- ② **每个受访者都可以回查来源**（他是谁、住在哪、做什么工作）。

4.2 我们不比别家强的地方（必须一并读）

维度	事实
与真人的准确度	本平台目前没有公布的对真人效度数字 (我们有的是"可复现"与"人群构成精确"，不是"与真人比对的结果")
答案的分散度	实测：与"按参数生成画像"相比，开放题开头重复率 8% vs 10% —— "受访者有真实生活状态"这件事，并没有让答案更分散
个体知识边界	人设带真实处境后回答明显更具体，但仍会出现"其实不知道却照样答"

所以对外口径应当是：我们卖的不是"更像真人"，而是

"一个有来源、可回查、可复现、可逐人配对的人群"。

这是与"一批生成出来的画像"**不同的东西**，不是同一个东西的好坏之分。

4.3 与「数字孪生」路线的唯一关键差异

数字孪生的结构性天花板：要造一个孪生，就得对**那个真人**做深度访谈 ⇒ **越精确，越不可扩展**。

我们的路：市民是构造出来的，不需要对每个市民做深访 ⇒ 可做到数十万；

代价是**他们不对应现实中的任何具体人**（我们从不声称对应）。

我们声称的是：他们对应**"城市里这类处境的人"**，而且每个都能查到"这个人是谁、过着什么日子"。

两者**不是替代关系**：要"代表某个具体真人的态度" ⇒ 数字孪生更合适（但贵且不可扩展）；

要"看一类人群会怎么答、并且能重复验证、能逐人配对" ⇒ 本平台更合适。

五、可复现性：与「每次重新生成」最本质的区别

很多同类工具每次运行都会重新生成回答，于是「这次跑出来是这样，下次可能变个样」。我们不这样：

- 同一份人群参数 + 同一份问卷，**任何时候重跑，逐题逐选项完全一致**（实测：300 人 × 2 题 = 7 个「题目 × 选项」格子，逐格一致）。
- **这一点不需要你相信我们**
：报告里每个结果都有「数据校验」入口，它会用这份记录保存的**同一份参数当场重跑一遍**，把比对结果摆给你看。

- 该页同时给出分布体检、样本量稳定性与人群构成偏差 —— 都是当场算出来的，且只读、不写任何记录。

六、校验方法与实测结果

以下数字来自本平台 2026-10-10 的实测（脚本与测试用例均在交付物中，可复跑）。

6.1 样本构成：性格库精确，城市受供给约束

维度（客户设的目标）	性格库模式实际	城市模式实际（未对齐时）	城市能供给的上限
职业 · IT/互联网 目标 20%	20%（偏差 0.0）	1%	约 1.2%
年龄 · 18 – 24 岁 目标 30%	30%（偏差 0.0）	6%	约 7.5%
学历 · 高中及以下 目标 30%	30%（偏差 0.0）	65%	约 65%
收入 · 3000 元以下 目标 20%	20%（偏差 0.0）	5%	约 5.1%

结论：**性格库模式可以把每一项比例做到精确**；城市模式的偏差主要来自城市实际人口，而不是算法。为此平台会在你**填写目标时就标注每个档位的「城市上限」**，超限会变红提示。

6.2 后分层对齐：把可供给的部分尽量拉近

城市模式下，平台会另抽一批市民作为候选池，把**可供给**的档位换到接近你的目标（不是加权，因此不会改变报告口径）。实测：

样本量 / 维度	对齐前最大偏差	对齐后最大偏差	耗时的变化
400 人 · 职业/年龄/学历/收入	47.2 个百分点	22.3 个百分点	一次调研约多 2 秒
1000 人 · 收入维	- 26.9	- 11.7	—
1000 人 · 年龄维	- 28.1	- 18.9	—
1000 人 · 职业维	- 22.8	- 18.3	—

残留偏差即城市供给上限所致 —— 报告里同时给「城市上限%」列，并在超限时写明「不是抽样算法的问题」。

6.3 样本量稳定性：人数够不够，当场算给你看

- 把人数各减/增 10% 重跑，比较各选项占比的**最大漂移**：实测 300 人样本为 **4.22 个百分点**（判定「尚可」）。
- 漂移小 ⇒ 结论对人数不敏感、当前人数够用；漂移大 ⇒ 建议加大人数再跑。若人数未生效，页面会**如实标注「检查无效」**，不给假数字。

6.4 分布体检：主动告诉你「这题没用」

每道题都会体检，并主动标出可疑信号（宁可以说结论弱，也不粉饰）：

- 头部占比 ≥95% ⇒ 这批人在此题上几乎没有分歧；
- 各选项极差 ≤3 个百分点 ⇒ **区分度很低**（换人群参数再看是否变化）；
- 有选项为 0% ⇒ 可能人群里没人处于该状态，也可能选项设置不当。

6.5 性能

场景	实测
性格库模式 300 人（含多选与开放题）	约 134 毫秒
城市模式 + 配额抽样 + 后分层 400 人	约 3.3 秒（含另抽候选池）
城市人口构成（供给上限提示）	首次取数约 2 秒，一小时内缓存、其后 0.0 毫秒

七、方法与口径

- **抽样框架**：在世、成年（18 岁以上）、不在押的城市市民；平台对每次调研强制校验人数与等级上限。
- **人群控制**：分层配额抽样 + 后分层对齐（可供档位）；明确区分「你用城市真实分布」与「按我的比例从城市里抽」两种模式。
- **多选题口径**：分母为受访人数，各选项占比合计会大于 100%（报告与导出均标注，并给「平均选几项」）。
- **开放题口径**：由大模型按深度样本逐人生成（默认数十人，可在面板设定）；报告给回答原文、关键词频次（简易 n-gram，非语义分析）与分组差异，并标注来源。
- **导出一致性**：每份导出都带「口径与边界」说明；同一份记录可反复导出，无需重跑。

八、能力边界清单（使用前请读）

边界	含义与建议
城市里没有的人变不出来	城市模式受真实人口约束；需要精确比例请改用性格库样本
只改选项文字不影响仿真版作答	实测：引擎按选项位置与维度作答，不读文字措辞。 要测文案差异请用 LLM 模式 ，或把文案做成不同选项顺序
深度样本不等于全量	LLM 深度解读是佐证，不替换上面的分布结论；报告里两套口径并排
PDF 问卷不支持导入	PDF 取字常不干净；请另存为 Word/Excel，或直接粘贴文本
关键词不是语义分析	只用于看趋势；请结合回答原文阅读
敏感话题	平台只做分布模拟；使用场景与责任边界见合同附件《仿真数据使用条款》

九、怎么用它（实务建议）

- **预测试**：预算有限时先跑仿真，筛掉明显走不通的方案，再把真人预算花在剩下的少数方案上。
- **人群细分**：用小众人群（城市里很难约到访谈的群体）先看分布，找出值得深挖的方向。
- **两版对比**：同一批虚拟人在两版问卷/两个方案上的差异做了**显著性检验**（两比例 z 检验），并注明「差异不是抽样噪声」。
- **不要**用它替代重大决策前的真人验证 —— 这是我们在合同与文档里反复写明的一条。

十、可验证性清单（客户可自行核对）

你要验证的	怎么做
结论可复现	报告 → 「数据校验」→ 看①可复现性（当场重跑比对）
比例是否达标	报告 → 人群构成表（目标/实际/偏差 + 城市上限% 列）
样本量够不够	报告 → 「数据校验」→ ③样本量稳定性（人数上下各 10% 重跑，看漂移）
题目有没有区分度	报告 → 「数据校验」→ ②分布体检（头部占比/极差/0% 选项）
受访者是不是真人	城市模式报告 → 样本示例里点「回查」，可在城市页面看其档案与当日日程
导出内容是否完整	下载 Excel/PDF → 含逐题分布、交叉分析、样本明细与口径说明

本白皮书中的全部数字均由本平台实测得出，实测日期见各节标注；如后续有变化，我们会更新此文档并同步"对应平台版本"一行。

文档用途：客户评估与内部参考。对外引用时请保留出处与版本号 ——

白皮书 v1.1 | 修订日期 2026-10-11 | 对应调研平台版本 V1.6.0。