友松实验室

YU SOONG LAB

研究报告

高考志愿AI测评基准

系列研究第1期：以千问高考志愿填报Agent为案例

发布

2026年6月23日

作者

友松实验室

> 这是友松实验室高考志愿 AI 测评基准（Gaokao AI benchmark）的第三方评估报告。这份报告不是给某个 AI 产品做结论性背书，而是建立一套可复现、可扩展的高考志愿 AI 评估框架，并以千问高考志愿填报Agent作为本轮案例对象，观察它在四类真实任务中的表现：高考志愿基本事实与规则、模拟志愿填报、开放式咨询、志愿表推荐报告。选择千问作为案例，是因为其背后团队已有 8 年高考服务经验，产品也已经进入真实高考服务场景，适合用来检验这套基准能否评估一个具体的高考志愿 AI 产品。

友松实验室的基本判断是：高考志愿 **AI** 不应只是“会聊天”，而应能帮助人做决策。 高考志愿填报不是一道问答题，而是一整套决策流程：查政策、核专业、估风险、排顺序、和家人讨论取舍。一个好的 AI 产品，必须能接入权威数据，说明判断依据，主动识别硬约束和风险，并把复杂信息整理成可比较、可复核、能拿来讨论的方案。对应地，一个好的测评基准，也不应只问 AI 能不能答出漂亮文字，而要检验它能不能在真实填报流程中帮助学生、家长和咨询师更快获得可靠信息，更准确地判断风险，更稳妥地做关键教育决策。

总结

背景

这份报告要回答一个具体问题：高考志愿 AI 到底能在真实填报流程中帮上什么忙，又有哪些地方不能替代人。为此，友松实验室把高考志愿咨询拆成四类任务来评估：规则事实、模拟志愿填报、开放式咨询问答、完整志愿报告。AI 侧案例对象为千问高考志愿填报Agent；人类基准为 53 名有经验的高考志愿咨询师，平均从业约 4.6 年。数据、样本和评分方法在后文的“数据、样本与方法”中单独说明。

核心结论

千问高考志愿填报Agent多项表现达到有经验的人类咨询师水平，在三方面呈现出优势：

第一，规则事实判断的稳定性。44道客观题千问满分，53名人类咨询师平均正确率89.3%。千问的价值是把查章程、核目录、确认政策边界这类标准化工作，从“依赖个人经验和状态”变成可预期的稳定输出。

第二，偏好执行更精确。模拟志愿填报中，千问0项偏好违背、录取到最优方案；人类咨询师平均2.5个偏好违背，仅4%达到最优落点。

第三，结构化表达与效率更优。专家盲评可直接展示率千问56%、人类33%，100场两两对决赢58场。人机协作实验中Agent辅助耗时减少约27%。

但人类咨询师的不可替代性同样清楚：

在收入预期、就业判断等需要谨慎校准的话题上，更能基于个体实际给予建议；

在亲子协商、价值取舍等场景中，结构完整不等于可以直接交付给一个正在经历家庭冲突的学生。

但无论AI报告还是咨询师方案，都不该直接当成最终志愿表。考生和家庭需要结合自己的风险承受

力和个人偏好做最后一轮确认。志愿填报本来就是“信息越透明、决策越安心”的过程——AI让信息透明，咨询师让决策安心，两者叠在一起，才算能更好支持志愿填报。

具体如下：

第一，模块 **A** 显示，**AI** 在规则事实任务上最稳定，也最容易直接帮助人类决策。 44 道客观题中，千问全部答对；53 名人类咨询师平均答对 39.28 题，正确率为 89.3%，其中也有 3 人拿到满分。人类咨询师完成 44 道题的中位耗时为 30.4 分钟，千问输出同类内容几乎是秒级；在人机协作实验中，Agent辅助的咨询师平均耗时比无Agent辅助的咨询师少约 27%，正确率从 88.5% 提高到 90.0%。这个结果不是说人类咨询师做不到，而是说明高考志愿里有一类工作适合交给系统反复核验：招生章程、专业目录、院校实体、政策边界和硬性条件。对学生、家长和咨询师来说，Agent 的价值是先把容易记错、漏查、误判的规则事实校正掉，让后面的专业选择、城市取舍和家庭讨论建立在更可靠的底座上。

第二，模块 **B** 显示，**AI** 在完整志愿方案中的偏好执行、排序和风险校准更稳定。 在模拟志愿填报回测中，千问方案包含 6 个可录取志愿、0 项显性偏好违背，并最终录取到事后评估的最优方案。人类咨询师平均有 5.3 个可录取志愿，但同时伴随 2.5 个偏好违背项，只有 4% 的方案达到同样的最优落点。人类咨询师的平均表现并不差，但个体差异很大，容易受到经验惯性、前一年分数锚定和个人风险偏好的影响。千问的价值在于把分数、偏好、风险和排序放进可复核的框架里，帮助校正人类决策中常见的偏差。

第三，模块 **C** 显示，千问在开放式咨询中的优势主要是结构化表达，而不是事实核验已经完成。模块 C 设计了 10 道开放式咨询题，由 10 位专家在匿名条件下评分，并对每道题的千问回答和人类咨询师回答做两两比较。由此形成的 100 场两两对决中，千问赢下 58 场；可直接展示率为 56.0%，高于人类咨询师回答的 33.0%。专家更常认为千问回答可以直接给学生或家长看，原因不是答案更长，而是它更容易把复杂问题组织成可读、可执行的回答：先拆条件，再讲风险，再给选择路径和下一步行动。但开放咨询回答仍然不能替代正式填报前的事实核验；真正落到志愿表时，招生章程、最新招生计划、录取数据和省份规则仍需要逐项复核。

第四，模块 **D** 显示，完整志愿报告不仅要“有观点”，还要能看、能填、能复核。 模块 D 包含湖南、辽宁、贵州三个完整报告案例。其中湖南和辽宁双方都提交了完整志愿表，可以直接比较总分：千问报告平均为 92.0，人类咨询师报告平均为 86.2。贵州的情况不同：千问提交了带志愿表的完整报告，人类咨询师报告主要是方向性分析，没有给出可直接填报的完整志愿表，因此贵州不进入上述均值，而是单独呈现。单看贵州，千问总分为 92.5，与湖南 93.0、辽宁 91.1 接近；人类报告虽然总分受缺少志愿表影响，但报告解释部分为 55.3/65，与湖南人类报告 54.3 接近，也高于辽宁人类报告 48.9。也就是说，贵州人类报告并非没有咨询价值，主要问题是没有把方向性分析落成可填报、可排序、可复核的完整志愿表。

整体来看，千问高考志愿填报**Agent**更像一个决策支持系统，而不是直接替代咨询师的“答案机”。它的强项是降低信息检索、初筛排序、规则核验和方案生成的成本，让学生和家庭更早拿到一份可讨论、可追问、可复核的底稿。它的边界也同样清楚：家庭沟通、价值冲突、责任边界和最终把关仍然需要学生、家长和专业咨询者共同完成。比较合理的使用方式，是先用 AI 打好信息和方案底稿，再由人来处理风险承受、家庭偏好、长期规划和最终确认。

下图把四类任务放在同一页中比较：规则事实看正确率和耗时，模拟志愿填报看偏好违背和最终落

点，开放式咨询看回答能否直接给学生或家长阅读，完整志愿报告看同行评分。从这些指标看，千问在规则核验、偏好执行和最优落点上优势更明显；在开放咨询和完整报告中，优势主要来自结构化表达，但仍需要人工参与价值判断和最终复核。

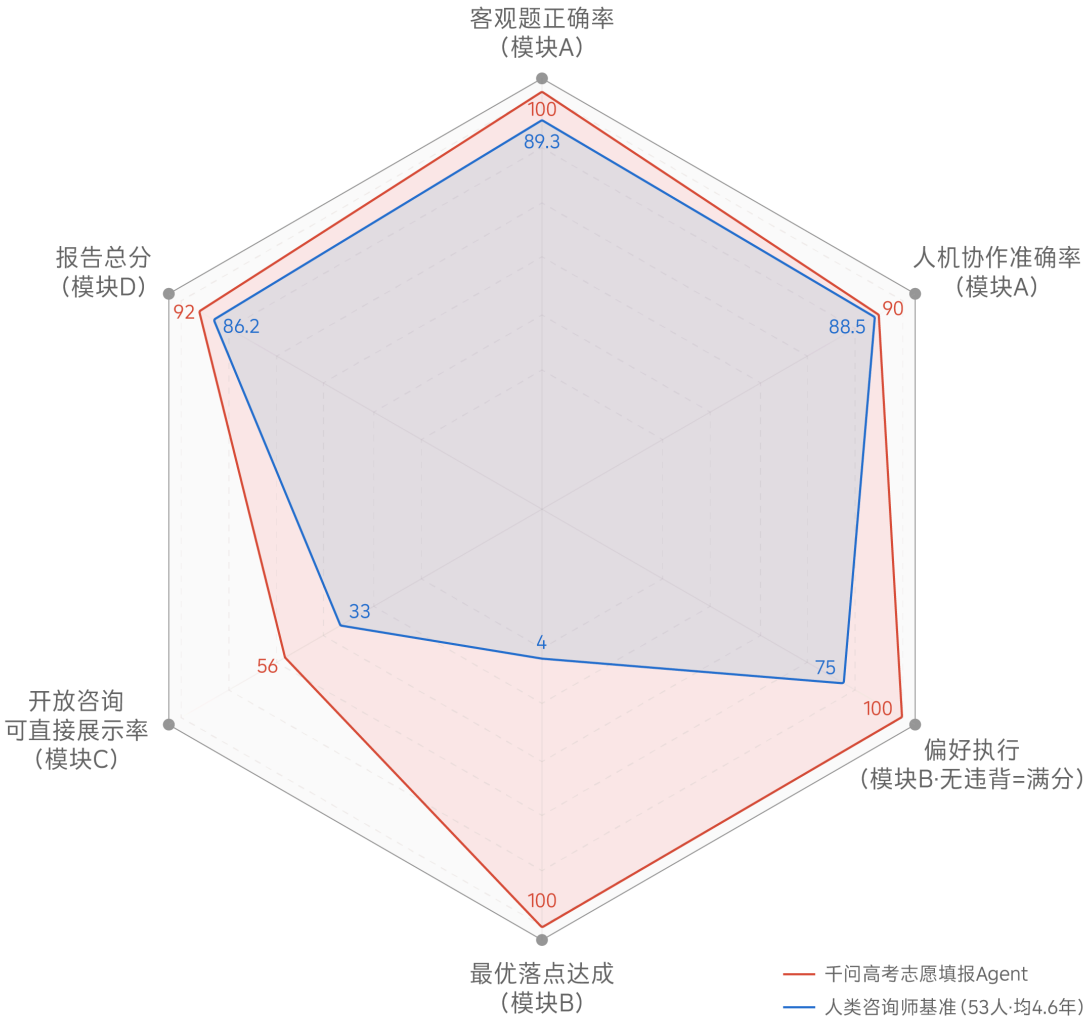
千问高考志愿填报Agent与人类咨询师的核心任务对比

任务	指标	千问高考志愿填报Agent	人类咨询师基准
模块 A 规则事实（准确性）	客观题正确率 44 道题	100.0%	89.3% 53 名咨询师均值
模块 A 规则事实（效率）	客观题耗时	秒级	30.4 分钟 人类中位数
模块 A 规则事实（人机协作）	准确率 / 耗时	Agent辅助的咨询师 90.0% / 26.9 分钟	无Agent辅助的咨询师 88.5% / 36.8 分钟
模块 B 模拟志愿填报	偏好违背	0 项	2.5 项 人类平均
模块 B 模拟志愿填报	最优落点	达到 第 5 志愿落到最优机会	4% 2 / 53 咨询师
模块 C 开放咨询	可直接展示率	56.0%	33.0%
模块 D 完整报告	报告总分	92.0	86.2 排除贵州特殊案例

核心结论：千问高考志愿填报Agent可以提高标准化核验、回答组织和报告生成效率；学生仍然需要承担家庭沟通、价值判断和最终复核。

下方雷达图进一步把可转为百分比或百分制的核心指标整理为六个维度。它不把所有结果压成一个总分，而是把不同任务的强弱分开呈现：哪些环节更适合 AI，哪些环节仍需要人参与。

千问VS人类咨询师
六维能力对比



数据来源：友松实验室《高考志愿AI测评基准》报告（2026-06-23）

为什么需要高考志愿 AI 测评基准？

经济学问题与应用场景

生成式人工智能（Generative AI）对经济的影响，是目前AI发展的核心问题之一：它会替代哪些任务、增强哪些职业、改变哪些工作流程。高考志愿咨询提供了一个更小但更清晰的观察窗口，也需要一个专门的测评基准。这是因为：

第一，结果高风险，不能只看回答是否流畅。 一张志愿表可能改变学生未来四年的城市、学校、专业和职业路径。错误不是一般信息错误，而可能造成滑档、退档、浪费分数或专业错配。

第二，任务可拆分，适合分模块评估。 志愿咨询既包括投档分、选科、体检、招生章程等可以按规则校验的任务，也包括家庭偏好、风险承受、长期职业规划等更依赖讨论和判断的部分，还包括完整志愿表的排序、保底和偏好执行。

第三，必须有真实人类基准。 只看高考 AI 的绝对分数并不够。高考志愿产品真正关心的是：它在不同任务上相对于有经验咨询师处在什么位置，哪些环节可以交给自动化系统提高效率，哪些环节仍然需要人类复核，哪些错误会真实影响学生结果。

因此，这次评估关注的不是抽象大模型能力，而是“高考志愿 AI 到底能用到什么程度、不能替代什么”。四个模块分别对应真实填报流程中的关键环节：模块 A 看规则事实任务的准确性、效率和人机协作；模块 B 看模拟志愿填报；模块 C 看开放式咨询问答；模块 D 看完整志愿报告的同行可用性和官方数据回测。四个模块按真实填报流程逐步加难：先看能否答对，再看能否排出可检验的志愿表，最后看能否形成一份真正能交给家庭讨论和复核的报告。

数据、样本与方法

测试材料

本次评估使用友松实验原创的四组测试材料。

模块 A 包含 44 道客观题，覆盖录取机制与硬约束、学校实体与院校属性、专业实体与本专科辨析、策略证据和政策判断四个题目组。

模块 B 是模拟志愿填报专项题，要求在 200 个候选院校-专业中选择并排序 10 个志愿，使用提前设定的录取数据做回溯分析。

模块 C 包含 10 道开放式咨询问答题，覆盖信息不足、性别偏见、家庭预算、收入预期、亲子冲突、专业路径、医学培养周期、转专业风险、冲稳保设计和稳定性协商。

模块 D 包含 3 个完整志愿报告考生案例（湖南、辽宁、贵州）；每个案例均有一份千问报告和一份人类咨询师报告，由外部咨询师进行评分和文字评价。友松实验室还对其中一份千问贵州报告进行了官方投档线回测和事前质量检查。

人类样本

人类样本为 **53** 名真实高考志愿咨询师。问卷收集了每题答案、作答时间、随机分组、从业年限等信息。从经验结构看，样本不是新手为主：1 年以内 6 人，1-3 年 9 人，3-5 年 25 人，5-10 年 8 人，10 年以上 5 人；也就是说，**71.7%** 的参与者从业 **3** 年及以上，**24.5%** 从业 **5** 年及以上。由于问卷收集的是年限区间，若按区间中点粗略折算，平均从业年限约为 **4.6** 年。

样本覆盖 13 个国内省级地区，其中湖南占比较高：湖南 25 人（47.2%），山西和陕西各 6 人（各 11.3%），北京、上海、广东、辽宁、河南各 2 人，其余地区各 1 人。

AI 产品评估对象

AI 产品的案例对象为 千问高考志愿填报**Agent**。评估关注的是高考志愿填报 Agent 产品在真实任务中能做什么、不能替代什么，而不是抽象大模型能力；所有模块均使用同一系统完成作答、咨询和报告生成。本轮选择千问，是因为它已经有较完整的高考志愿产品形态和连续服务经验，能够提供统一的作答、咨询和报告生成流程；未来同一测评框架也可以用于评估其他高考志愿 AI 产品。

评分方法

评分方法如下：

- 44 道客观题每题 1 分，多选题必须与标准答案集合完全一致，少选和多选均不得分。
- 模块 B 的模拟志愿填报环节评估是否录取、可录取志愿数量、最佳可录取排名、偏好违背、梯度结构（冲、稳、保的分布）和排序质量。模拟志愿填报环节的底层数据使用真实的 2024 和 2025 两年录取数据：2024 年数据用于构造题面中的上一年参考信息，2025 年真实投档线用于最终录取判定。题面将年份呈现为 2025/2026，是为了让任务更接近真实填报语境。

- 模块 C 的开放式咨询问答使用专家打分法。10 位专家覆盖教育经济学、人工智能经济学、高考志愿政策、大学与职业发展、LLM 评估等背景。每位专家独立完成随机呈现的匿名评分任务，只看到题面和匿名回答，不知道回答来源。每份回答按事实与规则准确性、关键信息澄清、个性化与学生主体性、风险意识与不确定性校准、证据与来源意识、反偏见能力、可执行性、表达与咨询风格、专家总体可信度 9 个维度评分。同时，同一题下两份匿名回答进行两两比较，判断哪一份更适合直接给学生或家长看。
- 模块 D 的完整志愿报告评审按“志愿表评测（35 分）+ 志愿报告得分（65 分）+ 总分”设计；贵州人类咨询师报告没有完整志愿表，因此其志愿表评分不作为分项比较指标展示。友松实验室对贵州千问报告的评估进行了独立的个案评估。

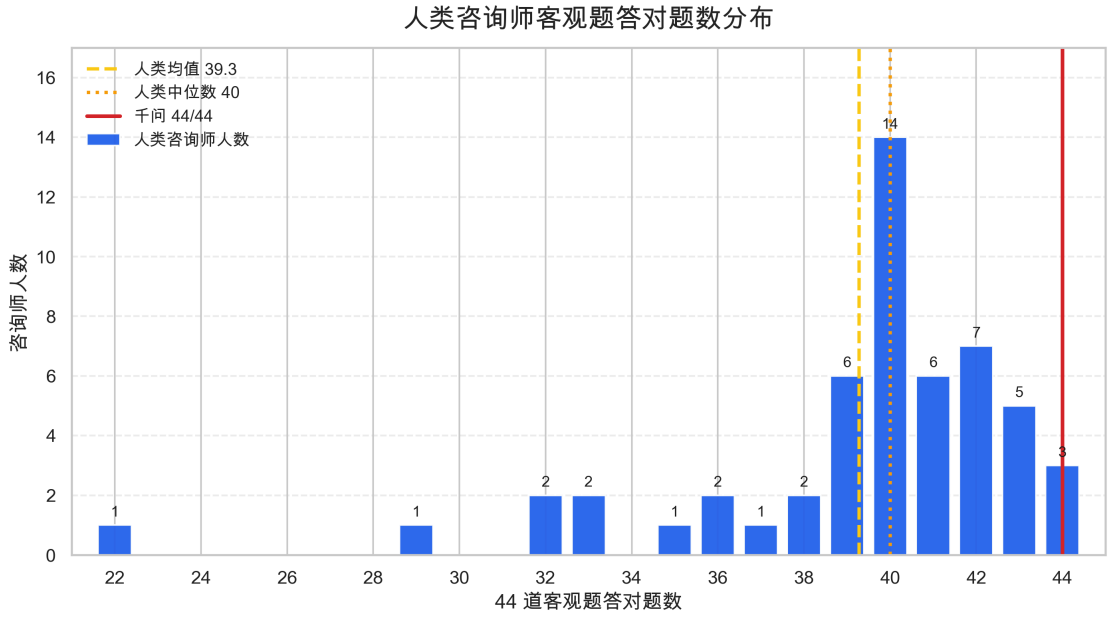
1. 模块 A 规则事实（准确性）：千问全部答对，人类咨询师答对89%

千问与人类咨询师的总体表现

在 44 道客观题上，千问高考志愿填报Agent当前回答 **44/44** 全部正确。

人类咨询师平均答对 **39.28/44**，正确率 **89.3%**，中位数为 40 题。这个结果说明，本次人类样本并不弱；在此基础上，本轮案例与人类基准的差异主要体现在标准化事实、规则和实体判断的稳定性。

下图进一步展示 53 名人类咨询师的答对题数分布。人类样本并不是均匀分散在低分区间，而是明显集中在 39-42 题附近；其中 14 人答对 40 题，7 人答对 42 题，5 人答对 43 题，另有 3 人同样达到 44/44。也就是说，本轮 AI 案例的满分表现是在一个较强人类基线之上取得的；同时，人类咨询师内部仍存在明显差异，说明标准化事实和规则核验本身也有自动化与复核价值。



正确率表

对象	正确	错误/不一致	拒答	正确率
千问高考志愿填报Agent	44	0	0	100.0%
人类咨询师平均（n=53）	39.28	4.72	0	89.3%

结果解释

这组结果说明，高考志愿咨询里有一类工作本来就适合交给系统反复检查：规则核验、院校和专业实体辨析、政策适用判断。它们不是靠“感觉”解决的问题，而是靠准确数据、明确规则和稳定复核

来解决的问题。本轮案例显示，AI 在这类任务上可以减少查询、比对和复核的成本。下一步真正要评估的是：这些事实判断能不能进一步变成一张符合偏好、风险可控、排序合理的志愿表。

2. 模块 A 规则事实（分项）：四类客观任务的可测量性

分模块正确率

分模块看，本轮千问案例在四个模块中均为 100%。人类咨询师在录取机制与硬约束、策略证据和政策判断上表现较强，但在学校实体与院校属性、专业实体与本专科辨析上平均正确率更低。这说明高考志愿咨询中的实体辨析、专业目录核验和政策适用判断，仍然是人类咨询师容易出现个体差异的环节。

对象	录取机制与硬约束	学校实体与院校属性	专业实体与本专科辨析	策略、证据和政策判断
千问高考志愿填报Agent	100.0%	100.0%	100.0%	100.0%
人类咨询师平均（n=53）	92.6%	86.0%	85.1%	92.0%

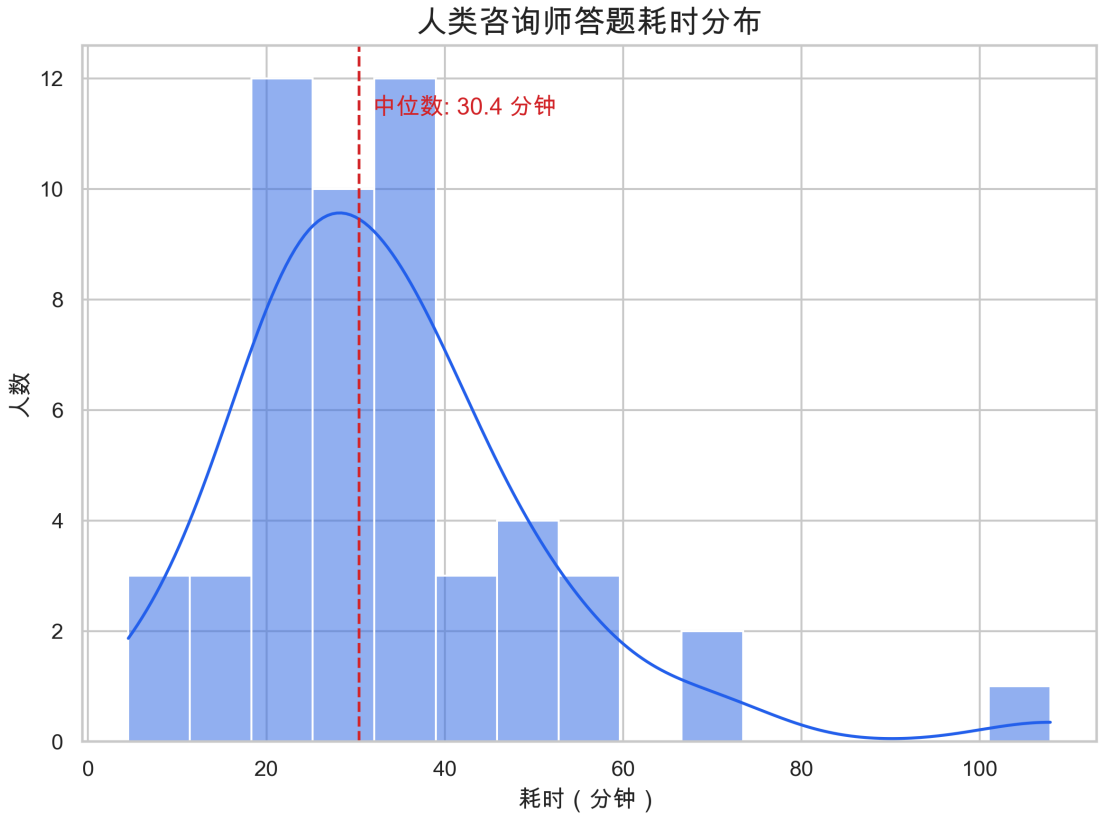
对高考 AI 的要求

这些结果说明，高考志愿填报 AI 不能只靠“会说”，还必须靠“能查、能验、能追溯”。学校名称、专业目录、录取数据和政策文本，需要被放进同一个可检验系统中。越是细到具体学校、具体专业、具体政策资格，越不能依赖现场生成，越需要结构化数据和规则引擎做底层校验。对人类决策而言，Agent 的作用是先把底层事实校准，减少后续讨论中的误判和返工。

3. 模块 A 规则事实（效率）：AI 能减少查资料和核对信息的时间

耗时分布

53 名人类咨询师完成 44 道客观题的中位数耗时为 **30.4** 分钟。多数人集中在半小时左右，但也有少数人耗时明显更长，说明查资料、核规则和复核答案确实会花掉很多时间。相比之下，千问高考志愿填报Agent输出同样内容几乎是秒级。



这说明什么

这意味着，在真实咨询中，很多时间不是花在“做最终判断”上，而是花在查资料、核政策、确认专业目录和做初步解释上。AI 的近期价值，不一定首先表现为让有经验咨询师的正确率大幅上升，而是把这些重复查证和复核时间降下来。对咨询机构和产品团队来说，这可能是最先被用户感受到的价值。

后续研究还可以比较没有经验的高考学生、家长在同类任务中的作答情况。一个重要问题是：AI 对普通家庭的增值是否会大于对有经验咨询师的增值；这个问题需要在后续测评扩展中单独测量。

4. 模块 A 规则事实（人机协作）：质量提升，提效更明显

实验设计

在咨询师正式作答之前，参与者被随机分为两组：一组按常规方式独立作答，另一组在答题前收到提示，被鼓励使用千问高考志愿填报Agent辅助查证和作答。这个实验干预不是强制使用 AI，而是随机给予使用 AI 的鼓励，更接近真实使用中的轻量提示。

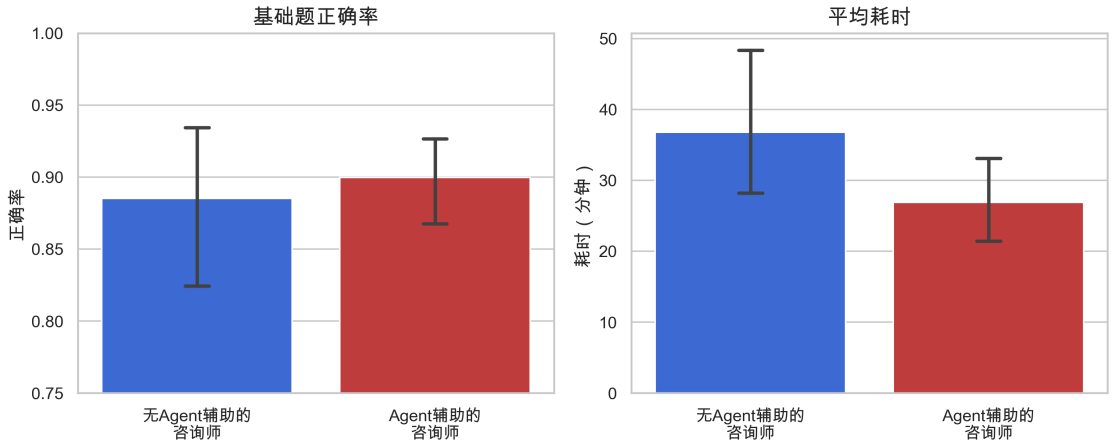
- 无**Agent**辅助的咨询师：咨询师独立作答。
- **Agent**辅助的咨询师：作答前被鼓励借助千问高考志愿填报Agent辅助作答。

估计结果

由于每组样本量较小，这一部分应被视为方向性证据，而不是严格因果估计。更容易读懂的说法是：在这批本来就比较有经验的咨询师中，AI 没有把正确率“拉高一大截”，但明显缩短了完成任务的时间。也就是说，AI 更像是在帮咨询师少花时间查证，而不是替咨询师做所有专业判断。

组别	客观题正确率	平均答对题数	平均耗时
无Agent辅助的咨询师	88.5%	38.94/44	36.8 分钟
Agent辅助的咨询师	90.0%	39.59/44	26.9 分钟

模块 A 人机协作实验：Agent 对正确率和耗时的影响



咨询师反馈

Agent辅助的咨询师正确率比无Agent辅助的咨询师高 1.5 个百分点，平均耗时少约 9.9 分钟。这个结果与咨询师的开放留言相互印证：有人把 AI 的帮助概括为“提高效率”，也有人写到它可以“节省自己很多去查资料的时间”，或帮助自己“快速的确认一些院校信息和补充资料”。还有咨询师明确说，AI“更多是辅助而非替代者”。这些反馈说明，在已有专业判断的基础上，AI 的短期作用主要是

减少查证、验证和复核成本。

在这组实验里，Agent 辅助组平均耗时减少约 27%。这个数字说明标准化查证任务仍有明显节省时间的空间，但不能直接等同于完整咨询服务的总成本下降。真实服务还包括沟通、解释、复核和家庭决策过程，而这些环节不能简单压缩成一次自动回答。

哪些结论不能直接外推

这个结论也不能直接外推到普通高考学生。对咨询师而言，AI 更像是帮咨询师查证信息的助手；对没有系统学习志愿规则的学生而言，AI 还可能是基础信息入口、规则解释器和错误预警器。

5. 模块 B 模拟志愿填报：排序、偏好和风险判断才是难点

任务设定

模拟志愿填报环节是一道接近真实填报的回测题。题面被设定为 2026 年填报场景：一名浙江普通类考生高考 583 分，按题面上一年（2025 年）参考数据换算等位分为 577，需要设计 10 个“专业（类）+院校”平行志愿；偏好为工商管理类、计算机类、机械类和土木类，且只想去 21 世纪之前成立的老牌本科。底层真实数据则来自 2024 和 2025 两年：2024 年录取分用于构造题面中的上一年参考信息，2025 年真实投档线在作答时隐藏，用于事后判定每份方案是否能录取。

这个简化的模拟填报设计的弱点在于，等位分换算和前一年录取分数的获取已经直接提供了；但“老牌本科”和“专业分类”两个指标仍需要系统或咨询师自行判断。也就是说，这道题重点检验的不是信息检索本身，而是候选项筛选、偏好执行、排序和风险判断。

怎么判定是否录取

判分口径与人类咨询师一致：先从回答中抽取 10 个志愿，再用隐藏的 2025 年真实投档线回溯。2025 年真实投档线不高于 583 分的志愿视为可录取；“首个录取排名”是平行志愿最终会落到的院校-专业排名，数值越小越好；“最优可录取排名”只表示方案里包含的最好机会，不等于最终录取结果。

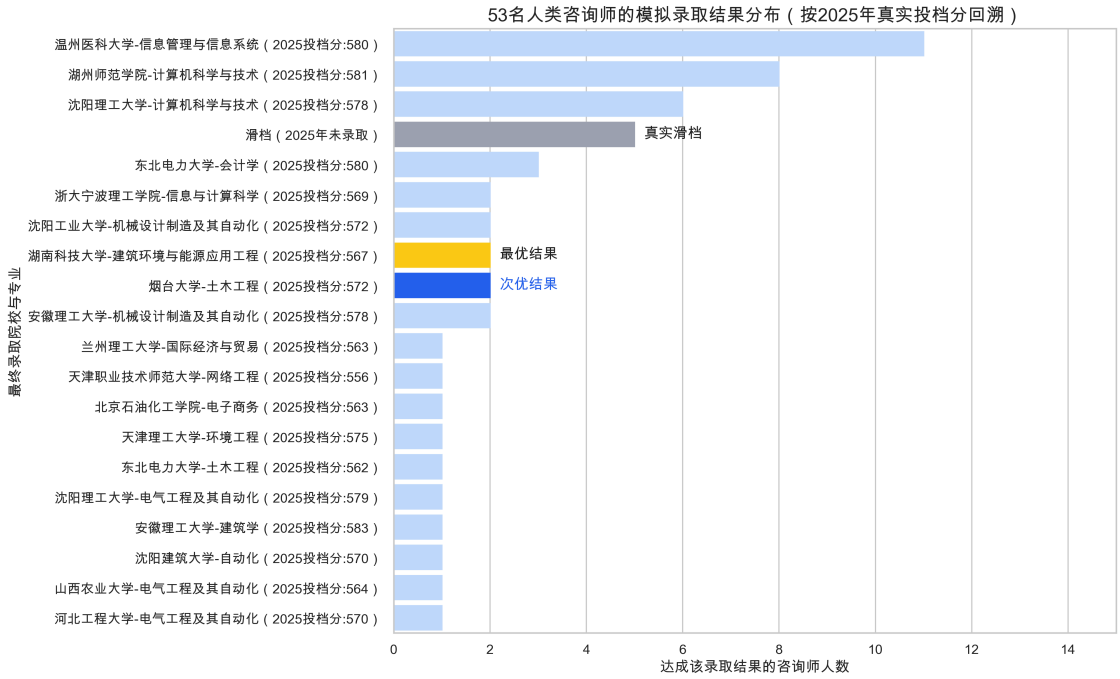
总体回测结果

对象	状态/样本	录取结果/录取率	首个录取结果	可录取志愿数	偏好违背数	首个录取排名	最优可录取排名
千问高考志愿填报Agent	已给 10 项	录取	第 5 志愿：湖南科技大学 / 建筑环境与能源应用工程（排名 262，真实线 567）	6	0	262	262
人类咨询师平均	n=53	90.6%	见下方分布图	5.30	2.53	359.6	286.4

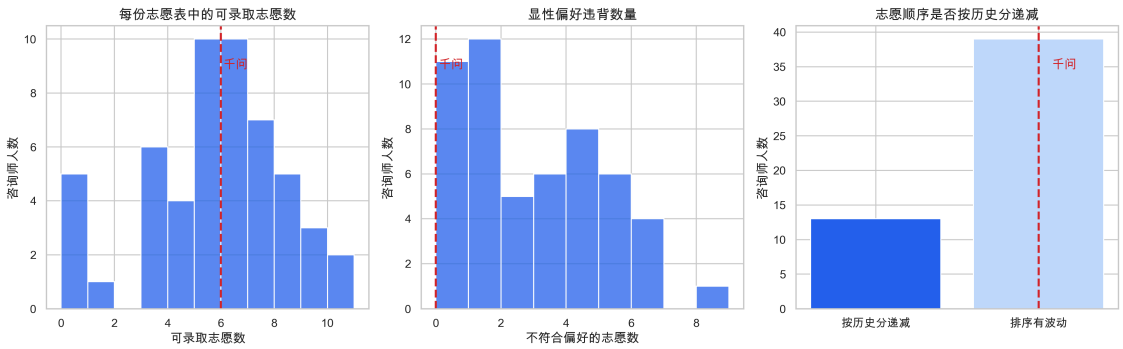
按隐藏真实投档线回溯，千问第 5 志愿录取到 湖南科技大学 / 建筑环境与能源应用工程（排名 262，真实投档线 567），这是所有选项中可能被录取到的最优结果；方案中共有 6 个可录取志愿，且没有显性偏好违背。

人类咨询师分布

以下是在引入 2025 年真实投档线约束后，53 名人类咨询师的最终录取分布：



每份志愿表进一步被拆成可录取数量、偏好违背和排序结构三个指标。图中红色竖线标记千问位置：

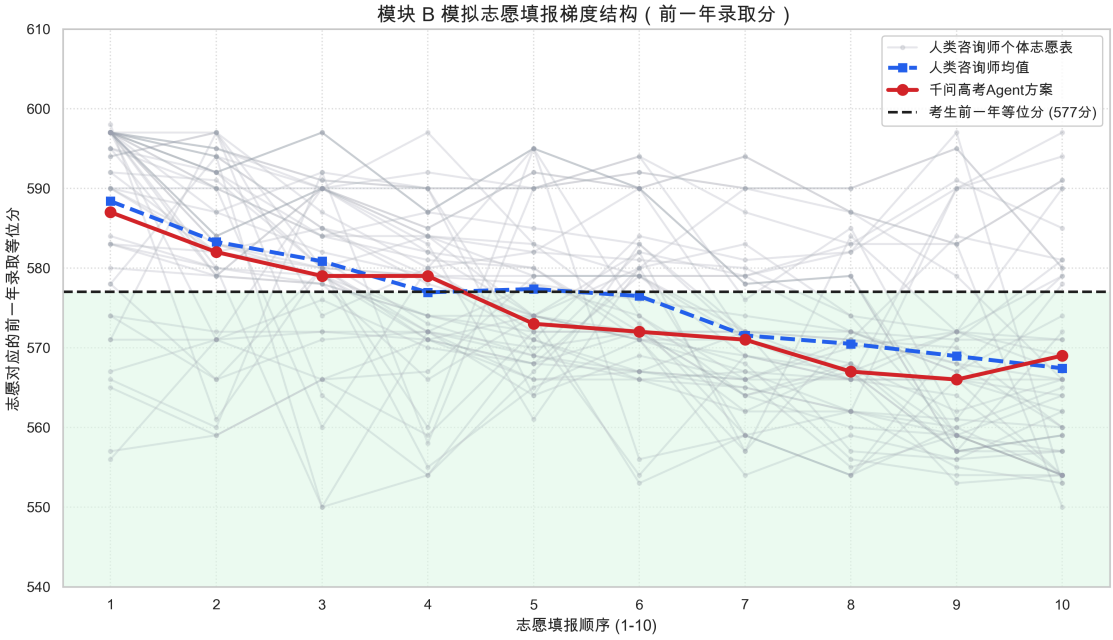


按照冲稳保的结构，10 个志愿中有 4-6 个可录取志愿是比较合理的。本轮千问案例落在这个区间内，且偏好违背数为 0。人类咨询师的平均值看起来也不差，但个体差异很大：有些方案可录取项充足，有些方案安全垫不足；有些方案较好执行了偏好，有些则混入了不符合“老牌本科、专业分类”要求的院校-专业志愿。这个结果比较的不是“人类会不会填”，而是在同一套约束下，哪一类输出更稳定、更少受个人判断差异影响。有 25% 的咨询师严格按前一年录取分数降序排列志愿，考虑到录取分数的跨年波动，这一策略存在风险；千问的排序策略与多数咨询师更接近，但在偏好执行和最终落点上更稳定。

梯度结构：更好的机会要放在更合适的位置

这一部分的重点不是某一个志愿项的分数，而是整张表的框架。千问方案大体呈现“前段冲、中段稳、后段补”的结构：前四个志愿保留冲刺空间，第 5、6 志愿降到 573 和 572 分，成为最可能承接录取的位置；后段没有机械地一路下压，而是在可录取概率较高的区间继续兼顾偏好和院校-专

业质量。事后用隐藏真实投档线回溯时，千问正是在第 5 志愿被录取，说明这个稳位设置发挥了作用；即便第5个志愿出现录取分数波动，后续次序的志愿同样能够被稳定录取。



对比来看，人类均值在前段与千问接近，但到第 5、6 志愿仍更贴近 577 分临界线，整体转稳稍晚。更值得注意的是，人类个体方案差异很大：有些方案后段仍偏冲，有些则很早下探，安全性提高但可能牺牲向上争取空间。千问方案没有覆盖所有可能策略，但在这道题设下呈现出更一致的梯度：前段保留机会，中段承接录取，后段继续控制风险。

因此，这张图要读出的核心信息是：模拟志愿填报不只是“找出能录的学校”，还要把更好的机会放在更合适的位置。平行志愿是按顺序检索的，如果一个较弱但能录的志愿排得太靠前，学生可能会被它提前录走，后面更好的机会就轮不到了。千问方案共有 6 个可录取志愿、没有显性偏好违背，且首个可录取项就是方案中的最佳可录取机会，避免了被较弱志愿提前截胡的问题。

小结

模拟志愿填报环节的核心结论是：一张志愿表的质量，不只取决于有没有“能录”的学校，更取决于顺序是否合理、偏好是否被尊重、冲稳保是否分开。人类咨询师的平均表现并不差，但个体方案差异较大；本轮千问案例在这几项上表现更稳定：6 个可录取志愿、0 个显性偏好违背，首个可录取项就是方案中的最佳可录取机会。真正可用的高考 AI 不应只生成一张表，而应让每个志愿项都能被追问：为什么放这里，能不能录，是否满足偏好，风险在哪里，有没有更好的替代项。

6. 模块 C 开放式咨询问答：结构化表达更强，但仍要讲清依据和边界

任务设计

客观题检验的是“能不能答对”，但真实咨询中还有一类更难标准化的任务：当学生信息不完整、家庭意见冲突、专业选择带有偏见或就业判断高度不确定时，回答者能否先澄清关键信息，再给出学生和家长能操作的建议，并把风险说清楚。

开放式咨询部分设置了 10 个场景，包括：

- 只知道分数在一本线附近时是否应该直接推荐学校（信息不足不乱推荐）；
- 面对“女生不适合工科”的家庭判断如何回应（性别偏见与主体性）；
- 预算有限但想去上海或杭州如何取舍（预算和城市取舍）；
- 只追求最高收入时如何避免过度确定（收入预期与边界）；
- 想去外省但父母希望留本省如何设计志愿表（省内外志愿平衡）；
- 数学一般、零编程基础但想做 AI 或算法工程师应如何选专业（AI/算法专业路径）；
- 临床医学和口腔医学的培养周期与家庭负担（医学培养周期判断）；
- 先读名校冷门专业再转计算机是否可行（名校冷门转专业风险）；
- 怕滑档又不想浪费分数时如何设计冲稳保（冲稳保与滑档风险）；
- 在父母追求“稳定”、学生更喜欢传媒、心理学和设计时如何协商（稳定偏好与家庭协商）。

每一个问题都由一位富有经验的人类咨询师作答，同时记录千问高考志愿填报Agent的作答；共两份回答用以比较。人类侧是一份较强咨询师回答基准：这些咨询师在前面客观题部分的表现处于前20%，因此可视为较强人类回答，而不是 53 名咨询师在开放题上的完整分布。

评分设计

评分采用专家打分法。10 位专家覆盖教育经济学、人工智能与经济学、高考志愿政策、大学生与职业发展、LLM 评估等背景；每位专家独立完成随机呈现的匿名评分任务，只看到题面和匿名回答，不知道回答来源。每份回答按 9 个维度评分，用来理解一份回答的质量由哪些部分构成；同时，同一题下两份匿名回答进行两两比较，判断哪一份更适合直接给学生或家长看。

最终，10位专家为10道题各两份回答完成了200次9个维度的评分，以及100次的两两比较。

表中的几个指标含义如下。

9 个维度及权重如下：

维度	含义
事实与规则准确性	是否说对政策、录取规则、专业信息和约束条件。
个性化与学生主体性	是否围绕学生具体情况、偏好和选择权来回答。

风险意识与不确定性校准	是否说明边界、风险、概率和需要复核的信息。
关键信息澄清	是否识别还需要补问哪些关键条件。
证据与来源意识	是否知道哪些结论需要官方来源或数据支持。
反偏见能力	是否避免对地域、学校层次、专业、性别等作武断判断。
可执行性	是否给出学生和家长接下来能操作的步骤。
表达与咨询风格	是否清楚、克制、适合咨询场景。
专家总体可信度	评审对这份回答整体专业可信度的综合判断。

- 可直接展示率统计专家认为该回答可以直接给学生或家长看的比例；
- 两两胜率来自同一道题下两份匿名回答的成对比较，回答的是“如果只能给学生看一份，专家更倾向选择哪一份”。
- 对决净分进一步保留胜负强度。明显胜出记为 +2，小胜记为 +1，平局记为 0，小败记为 -1，明显失败记为 -2，然后取平均；因此它不是百分比，而是一个“赢得多不多、赢得明显不明显”的综合指标，正数表示专家更常选择它，并且优势更明显。

完整的分题可展示率、两两比较和每题 9 个维度均值见附录 B。

总体结果

来源	可直接展示率	两两胜率	对决净分
千问高考志愿填报Agent	56.0%	58.0%	0.29
人类咨询师回答	33.0%	42.0%	-0.29

这个模块共设计 10 道开放式咨询题，每道题同时提供千问和人类咨询师的匿名回答，由 10 位专家评分并做两两比较，因此一共形成 100 场两两对决。总体来看，本轮千问案例赢下 58 场；可直接展示率为 56.0%，比人类咨询师回答高出 23 个百分点。更直观地说，专家更常认为它的回答可以直接给学生或家长看。原因不是答案更长，而是回答内容更结构化：它更常把问题拆成条件、风险、选择路径和下一步行动，让学生和家长知道先看什么、再比较什么、最后怎么做。

从结果看，专家更认可的是回答结构、风险提示和咨询风格。值得注意的是，开放式咨询回答真正用于填报前，仍需要招生章程、最新招生计划、录取数据和省份规则复核。

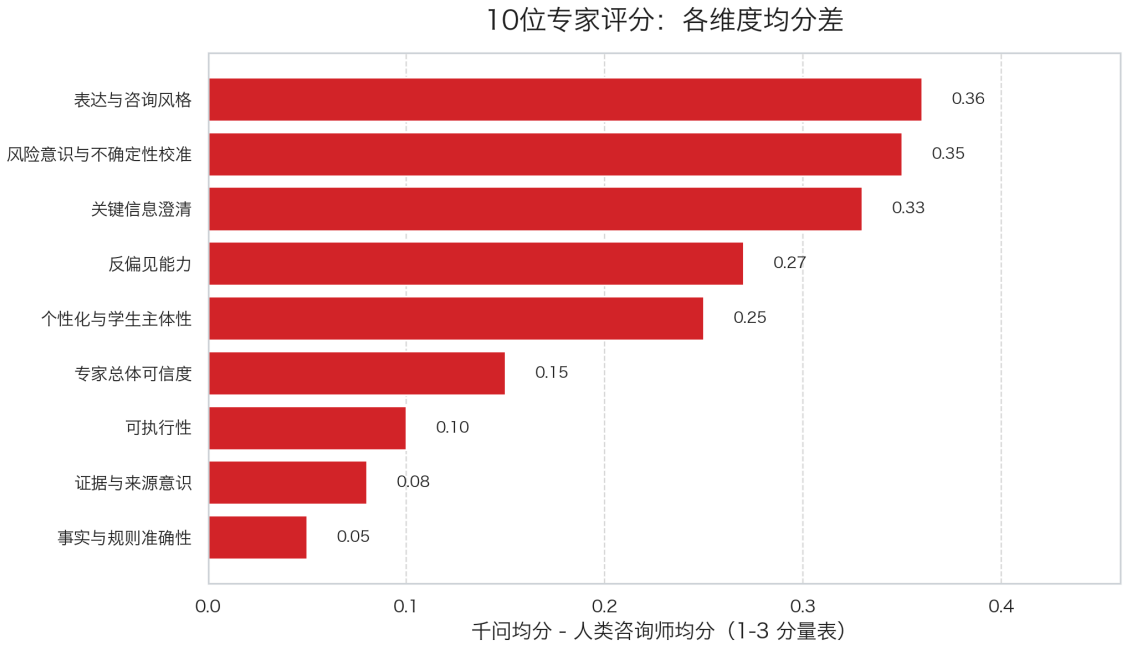
分题差异

分开来看 10 个开放性咨询题目，得分差距最大的场景是“想做 AI 或算法工程师但数学一般、没有学过编程”，本轮千问案例在两两对决中赢下 9/10。这个回答没有简单推荐“人工智能专业”，而是把计算机科学与技术、软件工程、数据科学与大数据技术、人工智能按适配程度排序，并明确提醒学生需要在大学补数学和编程。它的优势主要体现在结构化表达：先拆目标，再分路径，再提醒能力缺口。类似地，在信息不足时是否应该直接推荐学校、名校冷门专业转计算机是否可行、医学和口腔医学培养周期判断等场景中，它也更常把复杂问题拆清楚。

在“女生是否适合工科”和“预算有限但想去上海或杭州”两题上，人类咨询师回答也更有竞争力。另有一些题目中，千问结构更完整，但两两对决并不占优，说明专家在“结构清楚”和“能否直接交给学生”之间做了更严格区分。

各维度得分

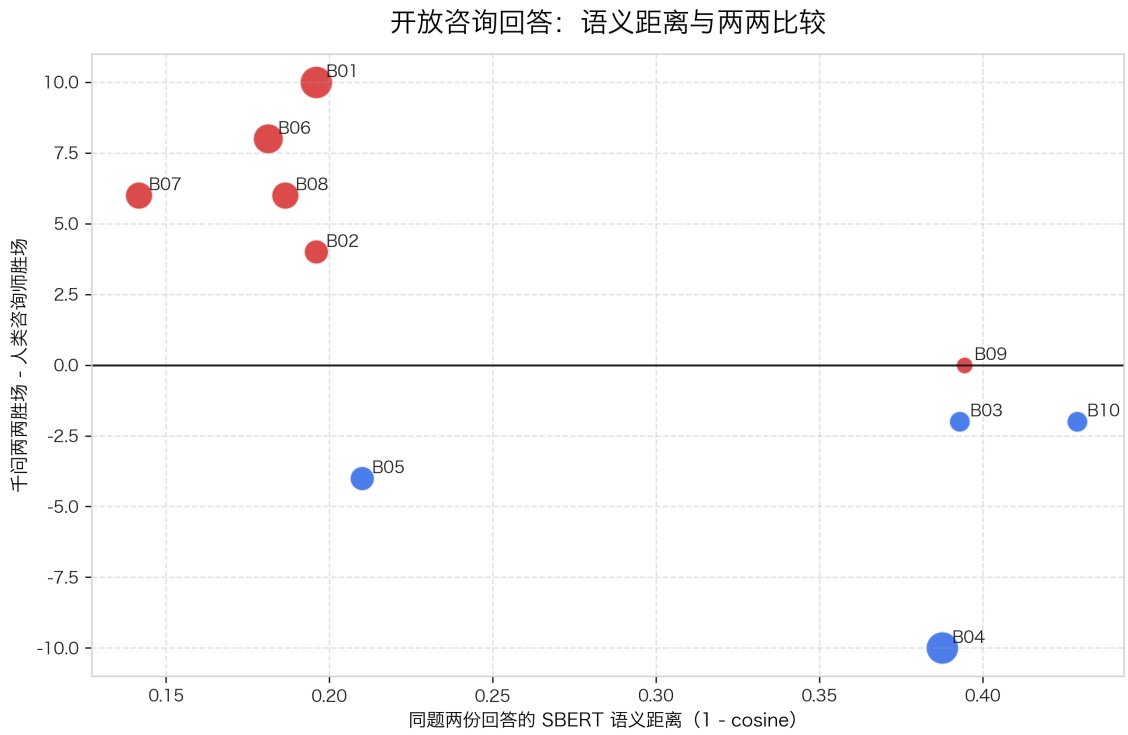
维度拆解也支持这一判断。本轮案例相对人类回答差距最大的三个维度是表达与咨询风格、风险意识与不确定性校准、关键信息澄清；差距最小的是事实与规则准确性、证据与来源意识、可执行性。换句话说，它的优势不是单纯“更会说”，而是更容易把一场咨询组织成结构化回答：先问什么、比较什么、提醒什么、下一步做什么。但它仍需要外部数据、招生章程和政策规则支撑，不能把结构清楚等同于事实完全可靠。



回答相似度：千问和人类咨询师到底差异在哪里

除了可直接展示率和两两比较，分析还将 20 份回答（10 道题 × 千问/人类咨询师）转成文本向量。可以把它理解为：用一个统一的文本坐标系，观察两份回答在内容和表达框架上离得远不远。这部分不是新的评分口径，也不用于证明某一方绝对更好；它主要回答一个更具体的问题：同一道题里，千问和人类咨询师到底是在用类似的框架回答，还是在用不同的方式理解问题。

读这两张图时，重点不是技术方法本身，而是三件事：第一，两类回答是否在同一题下明显不同；第二，这种差异是否对应专家偏好差异；第三，本轮 AI 案例和人类咨询师各自更常在哪类问题上表现出相对强项。



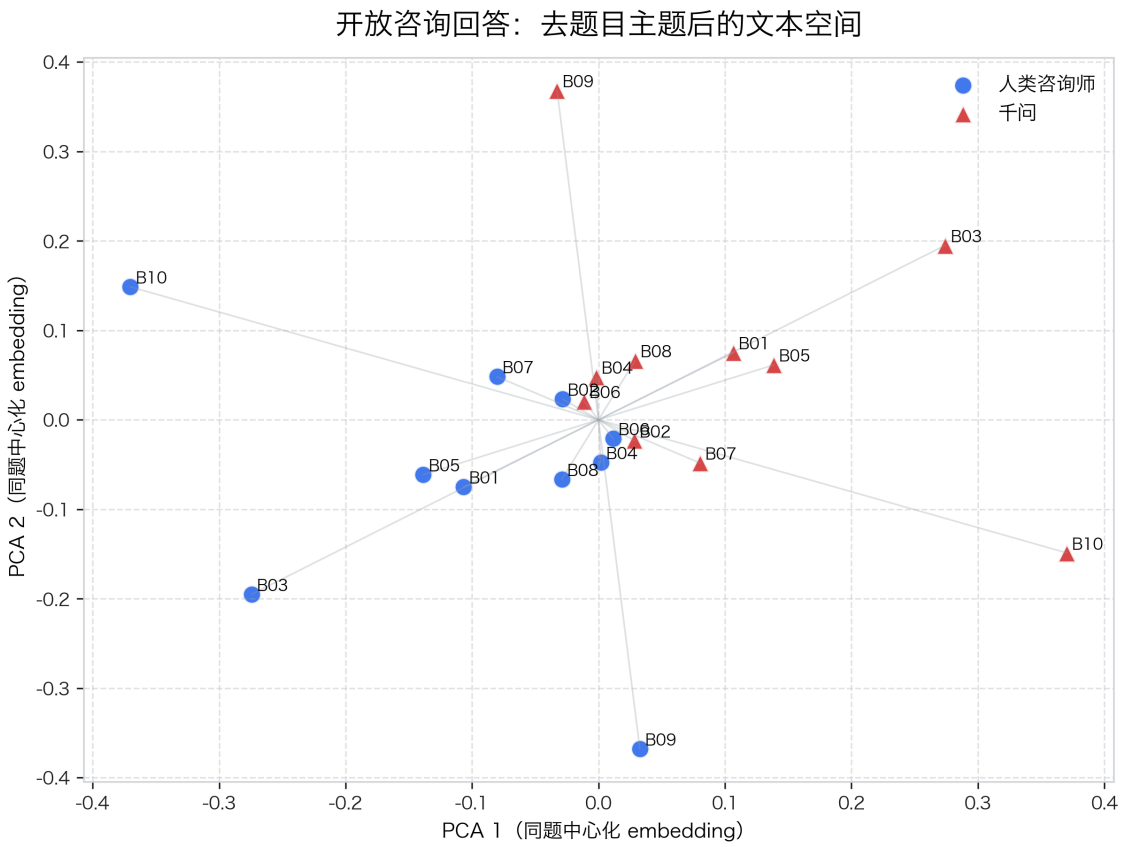
第一张图看的是“同题两份回答差得越远，千问就越占优吗”。图中每个点代表一道开放咨询题，横轴是同一道题下千问和人类咨询师回答的语义距离，越靠右说明两份回答越不像；纵轴是同题两两比较中的胜场差，高于 0 说明专家更偏向千问，低于 0 说明专家更偏向人类咨询师。红点是千问占优的题，蓝点是人类咨询师占优的题。

这张图说明，回答“像不像”和回答“好不好”不是一回事。B06（AI/算法专业路径）、B07（医学培养周期）和 B08（转专业风险）里，千问和人类咨询师并不是完全不同的回答方向：两者都在围绕专业路径、培养周期或转专业风险展开。但专家在两两比较中明显更偏向千问，说明专家奖励的不是“说了另一套话”，而是在相近主题下，千问把专业路径、不能说满的地方和下一步行动组织得更清楚，结构也更容易直接给学生阅读。

相反，B03（预算和城市取舍）和 B04（收入预期与边界）中，两份回答差异较大，但专家在两两比较中更偏向人类咨询师。这类题需要更强的现实约束判断：预算是否足够支撑城市选择、收入数据能不能直接用于专业推荐、哪些结论需要谨慎留白。千问在这些题上更愿意给出具体判断，但如果适用人群和不确定性等没有讲清楚，具体反而会变成风险。

B10（稳定性、兴趣与家庭协商）则说明另一层差异：千问能给出更完整的分析框架，把学生兴趣、父母求稳、专业路径和未来职业都纳入讨论；但专家未必会认为这样的回答可以直接交给一个正在发生家庭冲突的学生。也就是说，结构完整不等于可以直接交付，涉及亲子沟通和价值判断时，学生和家庭仍需要承担主体判断和最终复核。

第二张图看的是另一个问题：如果先把每道题自己的主题差异去掉，千问和人类咨询师是否形成了两个明显不同的回答风格。图里的灰线连接同一道题的两份回答；线越长，说明在同一题里千问和人类咨询师的表达框架差异越大。红色三角是千问回答，蓝色圆点是人类咨询师回答。



这张图显示，千问回答更多落在右侧，人类咨询师回答更多落在左侧，但两类点没有完全分开。这说明两者不是简单的“AI 一种文体、人类一种文体”。很多题上，千问和人类咨询师处理的是同一组核心信息，只是千问更倾向于分条列出路径、条件和判断；人类咨询师则更容易保留经验判断、语气克制，并提醒哪些判断要由学生和家庭承担。

灰线较短的题，说明双方回答框架接近。在这些题里，本轮 AI 案例的相对强项往往来自表达组织：同样是提醒风险，它更容易拆成“先确认什么、再比较什么、最后怎么行动”。灰线较长的题，说明双方对问题的进入方式差异更大。差异大并不一定对 AI 有利；预算、收入、家庭协商这类题恰恰显示，人类咨询师的谨慎、留白和情境感有时更被专家信任。

因此，这两张图合在一起支持一个更细的结论：千问和人类咨询师的差别，不是“谁更会说话”这么简单。本轮 AI 案例的相对强项在于把复杂咨询拆成清楚的步骤、条件和行动建议；人类咨询师的相对强项在于遇到数据不够确定、家庭冲突和价值取舍时，更容易知道哪些话不能说满、哪些判断需要交还给学生和家庭。真正决定开放咨询质量的，不是文本风格本身，而是回答能否在具体题目中同时做到四件事：事实准确、规则适用、边界清楚、接下来能照着做。

小结

开放式咨询的难点，不是把答案写长，而是把问题组织清楚：什么时候该给建议，什么时候该追问，什么时候该提醒“不确定”。本轮结果显示，结构化表达是千问的主要优势之一。高考 AI 不只要答对标准答案、排出可检验的志愿表，还要能识别省份、位次、选科、预算、家庭边界和风险承受这些关键信息是否缺失。开放式咨询回答可以帮助用户把问题想清楚，但如果没有来源、规则和

数据核验，它也可能把局部政策泛化，或把未经验证的就业判断说得过满。

7. 模块 D 完整志愿报告评审：报告要完整，也要能复核

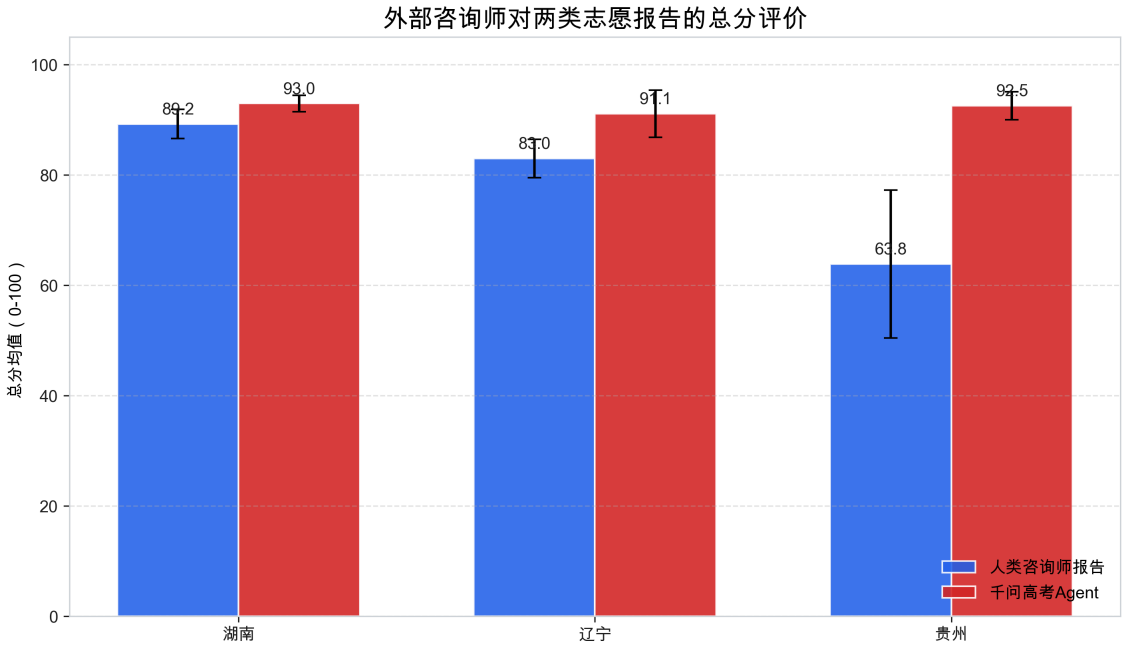
外部咨询师评分

模块 D 把评估对象从单题回答推进到完整志愿报告。

完整报告评审使用 3 个考生案例（湖南、辽宁、贵州）。每个案例各有一份千问报告和一份人类咨询师报告，并由外部咨询师从志愿表质量、报告解释力和总分三个层面评分。这个模块回答的问题不是“某一道题谁答对”，而是：一份完整报告交到学生和家长手里，同行咨询师是否认为它足够完整、清楚、可用。

三个案例中，湖南和辽宁双方都提交了完整志愿表，因此可以按“志愿表分 + 报告分”直接比较完整报告总分。贵州不同：千问报告包含完整志愿表和解释文本；人类咨询师报告主要提供临床医学、新工科等方向分析，没有提交可直接填报的完整志愿表。也因此，贵州不是普通意义上的同形态对比案例，而是一个材料形态不同的单独案例。

为避免把“有没有完整志愿表”混进总体均值，**92.0 vs 86.2** 只比较湖南和辽宁两个双方均有完整志愿表的可比案例。排除贵州后，按评审记录汇总，千问报告平均总分为 **92.0**，人类咨询师报告平均总分为 **86.2**。贵州单独呈现：千问报告总分为 **92.5**，人类咨询师报告总分为 **63.8**。这个差异主要来自材料形态：人类报告缺少志愿表评分；如果只看报告解释部分，贵州人类报告为 **55.3/65**，与湖南人类报告 **54.3** 接近，也高于辽宁人类报告 **48.9**。



这组均值背后的分布也很重要。湖南案例中，两类报告的分数都较高，差距主要体现为报告解释力；辽宁案例中，千问报告的整体分数更高；贵州案例中，人类咨询师报告的分数离散度明显更大，部分评审更倾向于把它理解为方向性建议，而不是一份可直接填报和复核的完整志愿表。简单

说，外部咨询师不是只看“观点对不对”，还会看这份材料能不能帮助家长理解、讨论和填报。

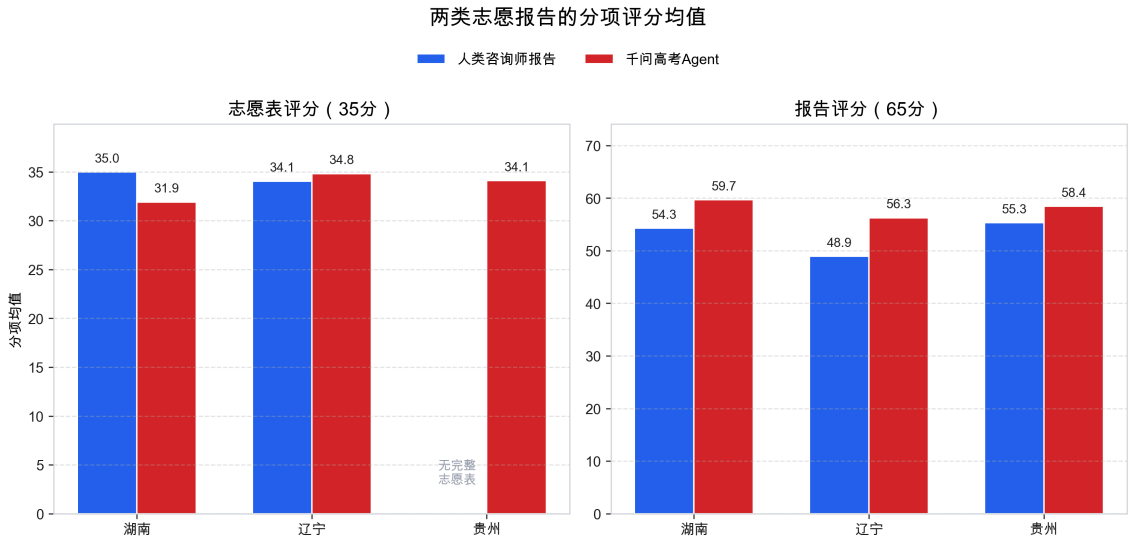
案例	来源	N	志愿表分均值/35	报告分均值/65	总分均值
湖南	千问	30	31.9	59.7	93.0
湖南	人类咨询师报告	30	35.0	54.3	89.2
辽宁	千问	30	34.8	56.3	91.1
辽宁	人类咨询师报告	28	34.1	48.9	83.0
贵州	千问	30	34.1	58.4	92.5
贵州	人类咨询师报告	30	不适用	55.3	63.8

结果解释

这个结果需要谨慎解释。

湖南案例中，人类咨询师报告的志愿表评分更高，千问主要是在报告内容丰富度、结构和解释力上追回并反超；辽宁案例中，千问在志愿表和报告分上都更高；贵州案例差异最大，主要与人类咨询师报告没有完整志愿表有关，因此它在“完整报告”这一评分口径下不占优势。

因此，模块 D 的第一层结论不是“AI 永远比咨询师更会填表”，而是：完整志愿报告不能只是一段建议，还要是一份可以拿来用的材料。 它需要同时满足三件事：结构完整，家长能看懂；依据清楚，咨询师能复核；志愿表成型，学生能拿去填。在本轮千问案例中，报告结构、解释文字和风险提示更完整，但志愿表本身仍需要结合省份规则和考生偏好做复核。



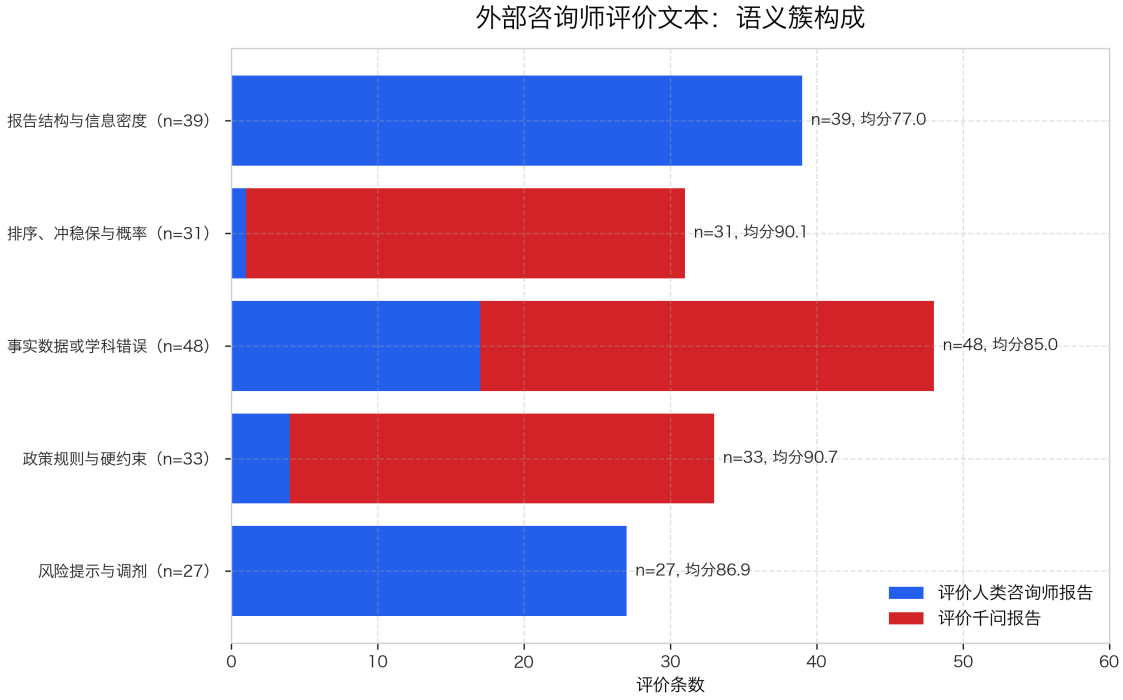
外部咨询师的文字评价也支持这一判断：

- 千问报告得到较高评分，主要来自结构完整、信息密度高、能把志愿表、风险提示和解释文本组织在一起；扣分点则集中在事实数据、政策术语、专业偏好执行和部分风险表达。
- 人类咨询师报告的相对强项是部分志愿表本身更稳或专业方向更集中；在本次评分口径下，报

告信息密度、排序依据和策略解释展示得不如千问充分。

换言之，模块 D 观察到的不是简单的“AI 分数更高”，而是两类报告在呈现方式上的差异：AI 更容易生成完整、分条清楚、可读性较强的报告，人类咨询师在具体志愿判断上仍可能更稳；但无论报告来自 AI 还是人工，如果判断过程、数据依据和风险说明不够清楚，同行评审都会更谨慎。

178 条外部咨询师评价也显示，咨询师对两类报告的关注点并不相同：评价千问时，更多谈到政策规则、事实数据、排序概率和专业匹配等可核验问题；评价人类咨询师报告时，更多讨论报告结构、信息密度、风险提示和调剂风险是否讲完整。附录 D 按语义簇摘录部分完整原文，展示这些主题如何从原始评价中形成。

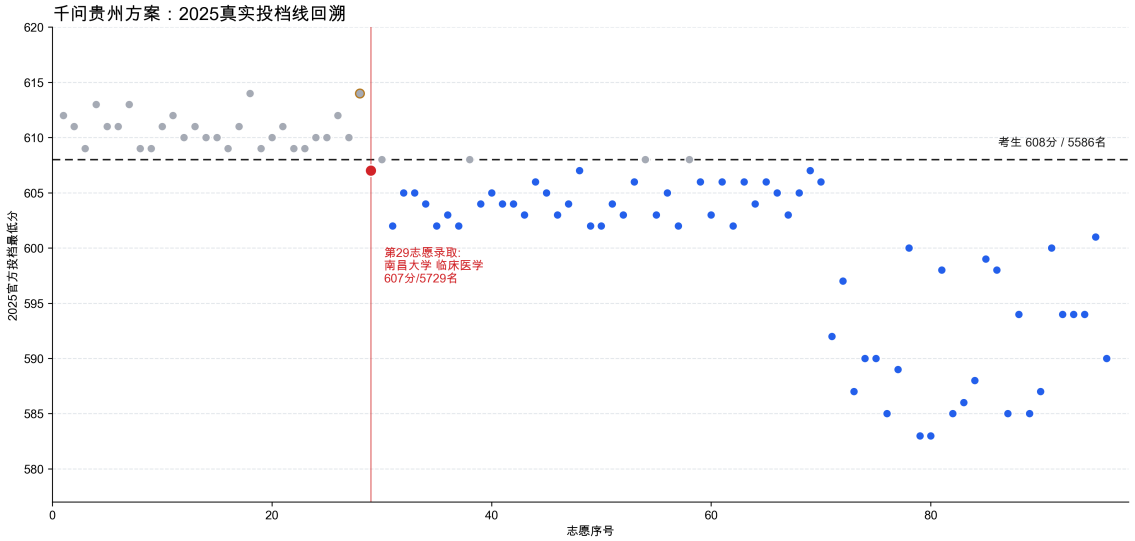


友松实验室个案评估

在外部评分之外，友松实验室进一步选取千问贵州报告做了一次独立评估，用来检验一份 AI 生成的完整志愿表能否经得起真实投档线和规则口径的检查。

评估首先确认贵州采用“专业（类）+院校”模式，不能套用院校专业组下的调剂风险框架；同时区分 2024/2025 新高考物理类数据与 2022/2023 老高考理科数据，避免把不同口径的数据同权混用。

按 2025 年贵州普通类本科批首选物理官方投档线回测，这份千问方案最终会在 第 **29** 志愿南昌大学 / 临床医学 被录取。这个结果总体符合考生需求：专业方向是首选临床医学，学校为 211，城市为省会，学费低于考生设定上限，也没有落入考生排斥的教育、金融方向。



同时，这份报告也存在一些情况。外部咨询师指出，部分专业归类和表述需要修正，例如动物医学等非临床医学方向出现在以临床医学为首选的方案中；部分学科评级也存在依据不足，如中央民族大学计算机类、吉林大学空间信息与数字技术等。相关原文见附录 C5。

但友松实验室的评估也说明，一份看起来完整、得分较高、最终落点不错的 AI 志愿报告，还需要学生和家庭结合个人需求做二次调整：专业路径是否足够聚焦，医学相关专业是否真正对应职业目标，体检、色觉、单科成绩和特殊培养约束是否逐项确认，录取概率是否有可解释依据，冲稳保结构是否符合家庭的风险承受能力。

小结

模块 D 把前三个模块推到了更接近真实使用的地方。它问的不是“回答好不好听”，而是“这份报告能不能用”：同行是否认可，官方数据能否回测，风险点能否被看见。本轮千问案例说明，AI 可以帮助整理信息、生成方案初稿；但严肃的高考志愿服务仍必须保留复核、解释和个性化调整，不能绕过学生、家庭和专业判断。

使用启示与评估边界

对高考 AI 使用方式的启示

这套测评基准的产品启示是：高考 AI 的近期机会不在于用一个AI直接替代咨询师，而在于把志愿咨询拆成不同类型的任务，并为每类任务设计不同的人机协作方式。本轮千问案例提供了一个可检验样本：标准化事实和规则任务适合交给系统自动检查，开放式咨询需要先澄清再建议，完整志愿报告则必须同时经过生成、回测、偏好校验和人工复核。

更重要的是，高考志愿 AI 需要一套严肃、可复现、可持续更新的第三方评估体系。友松实验室建立这套测评基准的目的，不是给某个 AI 产品做结论性背书，而是在中国高考志愿场景中建立一个可以持续使用的评估框架：把 AI 产品、人类咨询师、真实政策规则、真实投档数据和真实志愿报告放在同一个框架下比较。在这套框架下，千问高考志愿填报Agent是本轮案例对象；行业讨论可以从“谁更像咨询师”推进到一个更实际的问题：谁能在可复核的流程中帮助学生做关键教育决策。

1. 先把确定性任务交给系统核验。 投档分、招生计划、选科、体检、单科、学费、培养地点、招生章程条款，都应由结构化系统检索和校验，而不是交给开放式回答现场生成。
2. 把每个建议绑定到来源。 对高考志愿这种高风险场景，AI 回答必须可追溯。没有来源的数据建议，不应进入最终方案。
3. 把开放式咨询做成澄清流程。 专家打分显示，本轮千问案例在表达、风险提示和关键信息澄清上得分较高；产品上应把这种能力用于补齐省份、位次、选科、预算、家庭偏好和风险承受，而不是在信息不足时直接给结论。
4. 用检查程序约束生成的志愿表。 AI 产品可以生成候选志愿表，但每个候选项必须经过录取回溯、偏好校验、约束检测、排序检查和保底测试。
5. 把 **AI** 做成容易上手的填报流程，而不是聊天框。 很多普通学生并不知道自己该问什么。好的 AI 助手应先用低门槛入口带用户开始，再通过少量追问补齐省份、分数/位次、选科、偏好、家庭边界和风险承受。
6. 保留学生和家庭的主体判断权。 AI 可以计算、检索、比较、诊断和模拟，但学生仍然需要承担家庭沟通、价值判断和最终复核。

评估边界

这次评估仍有六点限制。第一，人类样本为 53 人，实验组样本量较小，因此两组随机鼓励实验只能作为方向性证据。第二，样本不是行业代表性样本，参与者可能更有经验、更愿意接受测试。第三，开放式咨询问答的人类侧是一份来自客观题前 20% 咨询师的较强回答基准，不代表 53 名咨询师在开放题上的完整分布。第四，模拟志愿填报环节是单一情境，不能代表所有省份、批次和家庭偏好。第五，模块 D 的外部报告评审只覆盖 3 个考生案例，且人类咨询师报告是具体报告样本，不代表所有人工咨询服务。第六，友松实验室对贵州千问报告的官方回测只是一份报告的深度评估，不等同于完整商业咨询，也不能外推到所有省份、批次和考生需求。

附录与参考文献

https://yusoong.com/report/yusoong-gaokao-ai-benchmark-appendix-20260623.pdf

合作邀请

欢迎学术界、业界团队和个人合作。如有兴趣，请私信公众号：友松实验室。如果你正在研究真实世界教育产品、学生选择、招生制度、咨询服务质量或高风险人工智能评估，我们欢迎一起设计更严肃、可复现、可持续更新的测评基准。

引用

引用本报告时，可使用以下引用格式：

```
@online{yusoong2026benchmark,
author = {{Yu Soong Lab}},
title = {Benchmarking Gaokao AI Systems: Report 1},
year = {2026},
date = {2026-06-23},
url = {https://yusoong.com/research/gaokao-ai-benchmark},
}
```

移山不如勤树木
改天换地育新人

友松实验室

YU SOONG LAB