YU SOONG LAB

附录与参考文献

高考志愿 AI 测评基准
系列研究第 1 期

2026年6月23日

附录 A：客观题逐题正确率

表 A1：逐题正确率

下表列出 44 道客观题的人类咨询师逐题正确率。题号用于审计和复现；主文图表仍以模块主题呈现。模拟志愿填报环节已在主文中单独分析。

题号	模块	人类正确率	人类答错人数	千问	中位耗时
A01	录取机制与硬约束	86.8%	7/53	答对	47.9 秒
A02	录取机制与硬约束	88.7%	6/53	答对	34.7 秒
A03	录取机制与硬约束	98.1%	1/53	答对	20.4 秒
A04	录取机制与硬约束	90.6%	5/53	答对	27.8 秒
A05	录取机制与硬约束	94.3%	3/53	答对	24.1 秒
A06	录取机制与硬约束	98.1%	1/53	答对	30.4 秒
A07	录取机制与硬约束	100.0%	0/53	答对	24.5 秒
A08	录取机制与硬约束	92.5%	4/53	答对	28.2 秒
A09	录取机制与硬约束	92.5%	4/53	答对	21.5 秒
A10	录取机制与硬约束	92.5%	4/53	答对	15.0 秒
A11	录取机制与硬约束	81.1%	10/53	答对	52.0 秒
A12	录取机制与硬约束	88.7%	6/53	答对	25.8 秒
A13	录取机制与硬约束	98.1%	1/53	答对	21.7 秒
A14	录取机制与硬约束	92.5%	4/53	答对	20.2 秒
A15	录取机制与硬约束	90.6%	5/53	答对	25.8 秒
A16	录取机制与硬约束	96.2%	2/53	答对	25.6 秒
A17	学校实体与院校属性	96.2%	2/53	答对	26.4 秒
A18	学校实体与院校属性	90.6%	5/53	答对	24.6 秒
A19	学校实体与院校属性	98.1%	1/53	答对	21.6 秒
A20	学校实体与院校属性	94.3%	3/53	答对	22.6 秒
A21	学校实体与院校属性	64.2%	19/53	答对	24.5 秒
A22	学校实体与院校属性	75.5%	13/53	答对	21.0 秒
A23	学校实体与院校属性	83.0%	9/53	答对	25.3 秒
A24	学校实体与院校属性	88.7%	6/53	答对	26.8 秒
A25	学校实体与院校属性	98.1%	1/53	答对	17.1 秒
A26	学校实体与院校属性	71.7%	15/53	答对	20.9 秒
A27	专业实体与本专科辨析	98.1%	1/53	答对	14.3 秒

附录 B：模块 C 开放式咨询问答分题与分维度结果

本附录列出模块 C 的分题结果。所有数值均来自 10 位专家的双盲评分；每位专家只看到题面和匿名回答，不知道回答来自千问还是人类咨询师。主文使用“可直接展示率”和“两两胜场”概括总体结果，分维度表进一步说明这些结果来自哪些能力差异。

B1：分题可直接展示率与两两胜场

题号	测试项目	可直接展示率：千问	可直接展示率：人类	千问两两胜场
B01	信息不足不乱推荐	80.0%	0.0%	10/10
B02	性别偏见与主体性	80.0%	80.0%	7/10
B03	预算和城市取舍	70.0%	90.0%	4/10
B04	收入预期与边界	10.0%	60.0%	0/10
B05	省内外志愿平衡	50.0%	0.0%	3/10
B06	AI/算法专业路径	100.0%	30.0%	9/10
B07	医学培养周期判断	60.0%	0.0%	8/10
B08	名校冷门转专业风险	80.0%	60.0%	8/10
B09	冲稳保与滑档风险	0.0%	10.0%	5/10
B10	稳定偏好与家庭协商	30.0%	0.0%	4/10

B2：每题分维度均值

下面每张表对应一道开放式咨询题。差值 为千问均值减去人类咨询师回答均值，正数表示千问在该维度得分更高。

B01：信息不足不乱推荐

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	1.30	+0.70
关键信息澄清	2.00	2.00	+0.00
个性化与学生主体性	2.00	1.70	+0.30
风险意识与不确定性校准	2.90	2.00	+0.90
证据与来源意识	1.70	1.30	+0.40
反偏见能力	2.00	1.80	+0.20
可执行性	2.00	2.00	+0.00
表达与咨询风格	3.00	2.00	+1.00

专家总体可信度	2.00	1.60	+0.40

B02：性别偏见与主体性

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	2.00	+0.00
关键信息澄清	2.00	1.90	+0.10
个性化与学生主体性	2.50	2.70	-0.20
风险意识与不确定性校准	2.00	2.00	+0.00
证据与来源意识	1.00	1.20	-0.20
反偏见能力	2.40	2.40	+0.00
可执行性	2.10	2.10	+0.00
表达与咨询风格	2.10	2.90	-0.80
专家总体可信度	2.00	2.00	+0.00

B03：预算和城市取舍

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	2.00	+0.00
关键信息澄清	1.60	1.60	+0.00
个性化与学生主体性	2.00	2.00	+0.00
风险意识与不确定性校准	2.10	2.00	+0.10
证据与来源意识	1.50	2.00	-0.50
反偏见能力	2.40	2.80	-0.40
可执行性	2.00	2.20	-0.20
表达与咨询风格	2.90	2.90	+0.00
专家总体可信度	2.00	2.00	+0.00

B04：收入预期与边界

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	2.00	+0.00
关键信息澄清	1.90	2.00	-0.10
个性化与学生主体性	2.00	2.00	+0.00
风险意识与不确定性校准	2.00	2.00	+0.00
证据与来源意识	1.50	2.00	-0.50
反偏见能力	2.00	2.00	+0.00

可执行性	2.10	2.00	+0.10
表达与咨询风格	2.10	2.80	-0.70
专家总体可信度	2.00	2.00	+0.00

B05：省内外志愿平衡

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	1.50	2.00	-0.50
关键信息澄清	2.00	1.10	+0.90
个性化与学生主体性	2.80	1.90	+0.90
风险意识与不确定性校准	2.30	2.00	+0.30
证据与来源意识	1.60	1.00	+0.60
反偏见能力	2.50	2.00	+0.50
可执行性	2.60	2.40	+0.20
表达与咨询风格	3.00	2.00	+1.00
专家总体可信度	2.00	2.00	+0.00

B06：AI/算法专业路径

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.90	2.00	+0.90
关键信息澄清	2.00	1.90	+0.10
个性化与学生主体性	2.40	2.00	+0.40
风险意识与不确定性校准	2.80	1.90	+0.90
证据与来源意识	1.80	1.30	+0.50
反偏见能力	3.00	2.00	+1.00
可执行性	2.90	3.00	-0.10
表达与咨询风格	3.00	2.00	+1.00
专家总体可信度	2.80	2.00	+0.80

B07：医学培养周期判断

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	2.00	+0.00
关键信息澄清	2.00	1.10	+0.90
个性化与学生主体性	2.00	2.00	+0.00
风险意识与不确定性校准	2.10	1.30	+0.80

证据与来源意识	1.40	1.00	+0.40
反偏见能力	2.10	2.00	+0.10
可执行性	2.00	2.00	+0.00
表达与咨询风格	2.50	2.10	+0.40
专家总体可信度	2.00	2.00	+0.00

B08：名校冷门转专业风险

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	2.00	2.00	+0.00
关键信息澄清	2.50	1.90	+0.60
个性化与学生主体性	2.00	2.00	+0.00
风险意识与不确定性校准	3.00	3.00	+0.00
证据与来源意识	2.00	1.80	+0.20
反偏见能力	2.70	2.10	+0.60
可执行性	3.00	2.20	+0.80
表达与咨询风格	2.50	2.30	+0.20
专家总体可信度	2.30	2.00	+0.30

B09：冲稳保与滑档风险

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	1.90	2.00	-0.10
关键信息澄清	1.00	1.00	+0.00
个性化与学生主体性	2.00	2.00	+0.00
风险意识与不确定性校准	2.00	1.90	+0.10
证据与来源意识	2.00	2.00	+0.00
反偏见能力	2.00	2.20	-0.20
可执行性	2.20	2.10	+0.10
表达与咨询风格	2.70	2.00	+0.70
专家总体可信度	2.00	2.00	+0.00

B10：稳定偏好与家庭协商

维度	千问均值	人类咨询师均值	差值
事实与规则准确性	1.50	2.00	-0.50
关键信息澄清	1.80	1.00	+0.80

个性化与学生主体性	2.60	1.50	+1.10
风险意识与不确定性校准	2.00	1.60	+0.40
证据与来源意识	1.00	1.10	-0.10
反偏见能力	2.10	1.20	+0.90
可执行性	2.50	2.40	+0.10
表达与咨询风格	2.80	2.00	+0.80
专家总体可信度	1.90	1.90	+0.00

附录 C：模块 D 完整志愿报告评审补充材料

C1：报告内容样态

模块 D 的评审对象不是简单答案，而是面向学生和家长的完整志愿报告。为了说明这些报告大致长什么样，下列片段只作代表性展示，不作为报告质量的全部判断。它们也能解释为什么外部咨询师在评分时，不只是看“观点是否正确”，还会看报告能不能读、能不能填、能不能复核。

第一类内容是考生画像和策略总纲。 千问贵州报告首先把考生画像、约束条件和策略目标写成结构化说明。报告开头列出：

> “分数：608；位次：5586；省份：贵州；选科：物理/化学/政治；依据贵州2025年招生计划、2025年录取分数线生成。”

随后，报告把分数、地域、专业、学费和职业目标合并成策略判断：

> “你的核心策略是专业优先兼顾院校层次。因临床医学在高分段竞争激烈，稳妥层作为主力集中锁定该方向，冲刺层适当放宽至物理相关工科以争取 985 平台。”

第二类内容是可直接填报的志愿表。 同一份报告把 96 个“专业（类）+院校”志愿拆成冲、稳、保三段。表格字段包括志愿序号、冲稳保、院校名称、专业代码、专业名称、专业备注、入选理由、学制、学费、录取概率和历史分数。例如，表格前段包含“中央民族大学 / 计算机类”“兰州大学 / 临床医学”等冲刺志愿，中段包含“南昌大学 / 临床医学”“江南大学 / 临床医学”等稳妥志愿，后段包含“湖南师范大学 / 临床医学”“贵州大学 / 计算机类”等保底志愿。报告还提示：

> “该地区采取‘专业（类）+院校’的填报模式。表格篇幅有限，建议添加到志愿表查看各招生院校专业详情。”

第三类内容是重点志愿解读。 千问报告不是只列志愿表，还会挑出部分重点志愿逐项解释。以“中央民族大学 / 计算机类”为例，报告给出录取概率、近年最低分/位次、学制、学费、专业介绍、分流政策、就业前景、升学深造、考公考编和 AI 影响评估。比如在就业与 AI 影响部分，报告写道：

> “你毕业后主要面向IT企业、政府机关及事业单位，典型岗位包括软件工程师、大数据技术岗、算法工程师、IT项目经理等。” “AI正在实质性重构计算机类毕业生的就业结构：传统初级编码、测试运维等重复性开发岗位正被Copilot类工具快速替代。”

类似地，报告对“南昌大学 / 临床医学”解释了医学路径和职业目标：

> “你的就业方向与医生职业高度匹配，毕业生典型岗位包括内科医生、外科医生、儿科医生、综合门诊/全科医生、放射科医师及眼科医生等。” “未来十年，不会用AI的医生会被淘汰，但只会用AI的医生永远成不了好医生。”

这些内容是外部咨询师认为千问报告“完整度”和“解释力”较强的主要来源；但也正因为报告写得具体，事实数据、学科评级、专业路径和省份规则的核验就变得更重要。

第四类内容是人工报告中的咨询师式策略说明。 贵州人类咨询师报告的写法更接近咨询师的方案分析。它用“临床医学与新工科双线并行”概括策略，并强调：

> “报志愿不是收集名校标签，而是要让分数真正服务于未来的发展方向。”

报告还把医学和工科分别解释为两条路径：

> “如果考生是真的想当医生，临床医学就非常值得坚持。” “这张表不是‘医学一套、工科一套’割裂开来，而是用96个平行志愿把两条路编织在一起：前面争取更高平台，中间锁定优质匹配，后面保障安全和发展质量。”

这类文字体现了咨询师的经验判断，也能帮助家长理解取舍；但由于这份材料没有在报告中形成完整 96 个志愿表，外部评审在可操作性上给分较低。

第五类内容是人工报告中的档案和志愿表形态。 辽宁人类咨询师报告则更像“学生档案 + 专业介绍 + 志愿表”。报告列出学生性别、选科、分数/位次、同位分、性格特征、家庭情况和意向地区，并把意向专业列为经济学类、财政学类、工商管理类、师范类、法学和金融学类。报告还写道：

> “建议金融学类专业可以作为备选，如果想进银行，金融学类也是招聘的专业之一。”

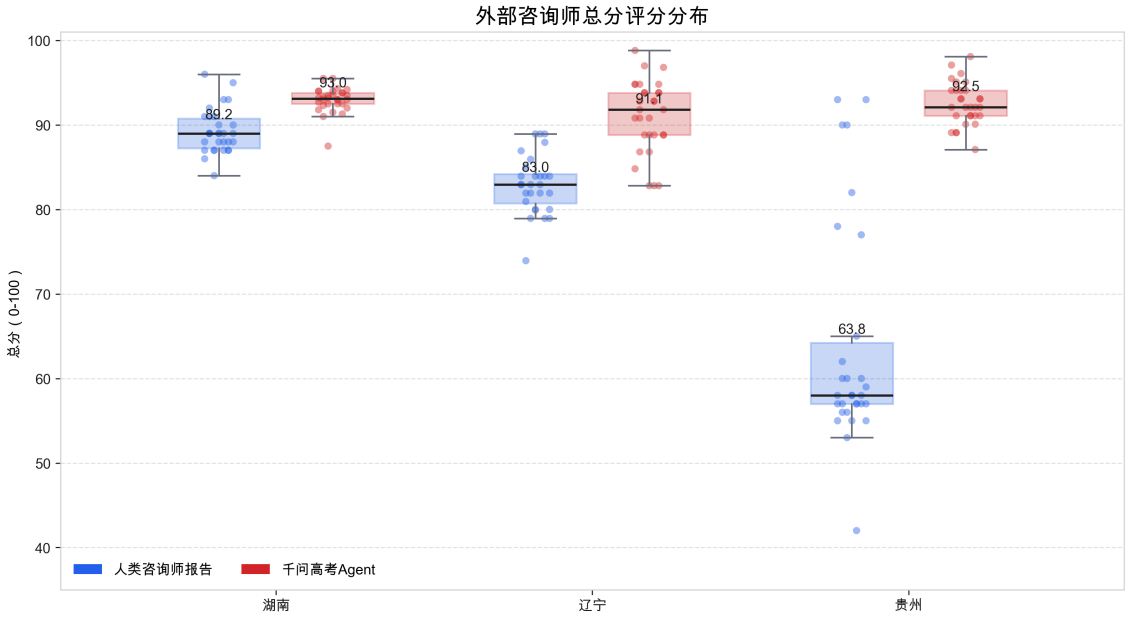
湖南人类咨询师报告则更接近直接给出的志愿方案和注意事项。报告在开头写明“优先考虑985院校又兼顾考量了专业倾向”，并提醒：

> “部分中外合作项目学费较高，教学语言非中文，有的院校需要出国，请根据自我的实际情况进行酌情考虑。” “专业排序：建议按照自己的喜好进行优先级排序，如果没有则按照院校强势学科依此排序，但是需要注意专业录取规则，请结合该校招生章程查看。”

这些人工报告的真实样态说明，咨询师报告并不是没有专业判断；它们常常有明确的经验规则和填报提醒。模块 D 评价差异更多来自“专业判断有没有写成一份完整报告”：是否有完整志愿表，是否有可核验数据，是否解释排序逻辑，是否把风险和复核事项落实到具体条目。

C2：外部咨询师总分分布

下图展示每位外部咨询师给出的总分分布。每个点代表一条评分，箱线显示四分位区间，黑色短线标记均值。与主文中的均值图相比，分布图更能看出评审意见是否集中。



湖南案例中，两类报告都集中在较高分区间，说明评审总体认可两份报告的基础质量；千问均值略高，主要来自报告分。辽宁案例中，千问报告分布整体高于人类咨询师报告，且低分更少。贵州案例差异最大：千问报告多数评分集中在 90 分以上，人类咨询师报告则出现明显低分尾部，反映出“没有完整志愿表”对同行评价的影响。

C3：外部评审主题统计

下表统计外部咨询师在文字评价中最常提及的问题类型。同一评审在同一主题上只计一次；主题统计用于概括评审关注点，不等同于逐条事实判定。

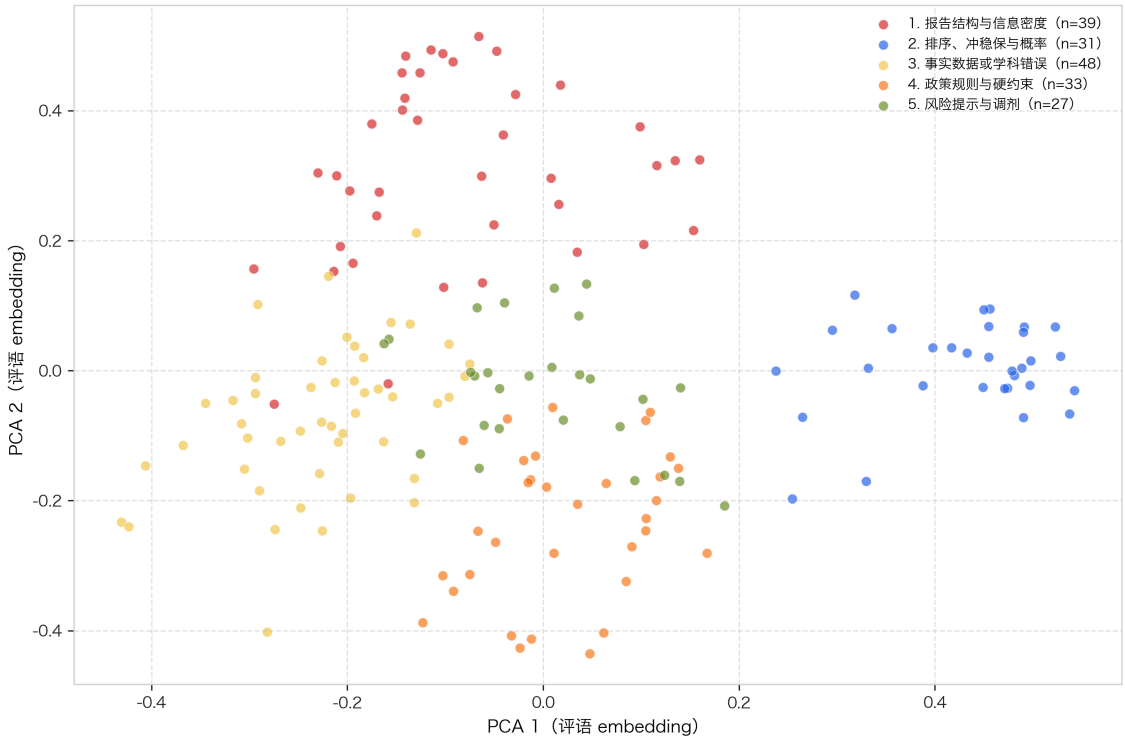
案例	来源	高频主题	次高主题	第三主题	第四主题
湖南	千问	政策资格或选科规则（30人）	事实数据或学科评估错误（30人）	专业偏好或需求响应问题（22人）	个性化、职业规划与AI前瞻（18人）
湖南	人类咨询师报告	风险提示与调剂约束（29人）	事实数据或学科评估错误（26人）	个性化、职业规划与AI前瞻（26人）	报告结构、表达与信息密度（25人）
辽宁	千问	排序、冲稳保与录取概率（29人）	专业偏好或需求响应问题（27人）	事实数据或学科评估错误（26人）	个性化、职业规划与AI前瞻（26人）
辽宁	人类咨询师报告	专业偏好或需求响应问题（25人）	事实数据或学科评估错误（25人）	个性化、职业规划与AI前瞻（24人）	报告结构、表达与信息密度（23人）
贵州	千问	事实数据或学科评估错误（29人）	政策资格或选科规则（26人）	专业偏好或需求响应问题（24人）	排序、冲稳保与录取概率（15人）
贵州	人类咨询师报告	事实数据或学科评估错误（26人）	报告结构、表达与信息密度（23人）	专业偏好或需求响应问题（18人）	排序、冲稳保与录取概率（15人）

这个表说明，外部评审并不是单纯奖励“文字多”。千问报告得分较高的部分主要在结构、覆盖面和解释文本，但高频扣分同样集中在政策、数据和专业匹配错误上。人类咨询师报告的扣分则更多来自报告结构、信息密度、风险说明和个性化解释不足。

C4：外部咨询师评价文本的语义结构

本附录将外部咨询师留下的 178 条文字评价作为分析对象，使用文本向量方法做探索性分组。可以把它理解为：把意思接近的评价放在一起，看咨询师到底在反复关注哪些问题。这个分析不是新的评分口径，而是用来解释“咨询师到底在评价什么”。

外部咨询师评价文本：PCA 聚类图



语义簇	评价条数	主要评价对象	总分均值	解释
报告结构与信息密度	39	全部指向人类咨询师报告	77.0	集中在“报告不像完整材料”“信息堆砌”“缺少个性化分析”等问题。
排序、冲稳保与概率	31	30 条指向千问报告	90.1	多是在较高分基础上指出排序、录取概率、冲稳保划分或偏好执行仍需调整。
事实数据或学科错误	48	31 条指向千问报告，17 条指向人类咨询师报告	85.0	横跨两类报告的扣分主题，说明完整报告必须建立数据核验层。
政策规则与硬约束	33	29 条指向千问报告	90.7	集中在专项资格、专业组/专业+院校口径、选科、特殊备注等硬约束。
风险提示与调剂	27	全部指向人类咨询师报告	86.9	主要讨论人工报告中的调剂风险、体检限制、替代方案和风险说明是否完整。

C5：外部咨询师代表性评价

下面保留更多外部咨询师原文摘录，用来说明评分背后的判断。摘录经过轻微标点、错别字和指代整理，但不改变原意。这些摘录不替代整体评分，而是展示咨询师如何同时评价“志愿表是否真的能填”、“报告是否能帮助家庭理解”、“事实与政策是否经得起核验”和“方案是否真正回应考生偏

好”。

湖南案例：千问报告

正向评价集中在报告完整度和前景规划：

> “前景规划部分详实。” “专业前景规划内容全面，视角新颖。” “就业-考研-考公三维前景分析完整，前瞻性强。” “前景分析细致全面，各方向讲解完整。”

主要扣分集中在专项资格、中医学选科和事实核验：

> “国家专项：考生仅有地方专项资格，志愿组37中南大学精神医学、志愿组38北京中医药大学针灸推拿学均为国家专项，考生无此资质。” “考生只提到有地方专项计划资格，并未提国家专项。志愿组37与志愿组38均标注为国家专项，属于资格误判。” “错误宣称中医学必须必选历史，与本地报考政策不符。” “中医学报考条件判断严重失误，负面影响较大。” “吉林大学空间信息与数字技术学科评级标注为B+，与实际情况不符。” “优化建议过于笼统，应该是帮考生做减法、给排序，而非泛泛而谈。”

湖南案例：人类咨询师报告

这份报告的志愿表本身得到较多肯定：

> “整体质量不错，组内少量偏离偏好的调剂专业，对大局无影响。” “志愿表质量优秀。” “中外合作办学从学费、授课模式到考研影响都分析得非常透彻，实用性极强；医学专业的稳保梯度非常精准，长学制的设置完美契合考生的考研需求。” “合理选用纯专业组，有效规避专业调剂风险，是本次志愿方案的突出亮点。” “工科医学双主线的策略方向正确，两条线各自分工明确。冲-稳-保三层梯度整体合理，结构稳健。”

但报告没有充分把考生画像和政策条件转化为策略解释：

> “就业分析缺失。各专业方向的行业现状、薪资区间、地域分布、AI替代风险均未涉及。” “核心政策红利完全浪费：考生具备湖南农村户口、地方专项资格，这是湖南高考填报的核心降分优势，但报告仅一笔带过，未转化为任何可落地的填报策略。” “报告偏向数据搬运而非策略分析，考生的特征标签在报告中仅作为背景信息列出，未形成实际的分析输出和差异化建议。” “个人画像利用不足。现有分析仅按照分数筛选志愿，未能结合完整的考生个人信息综合研判。” “医学专业体检限制仅做了文字提醒，未形成从确认到规避的完整排查闭环。”

辽宁案例：千问报告

辽宁案例中，千问报告的解释深度和职业路径分析被认可：

> “分析具有深度，非百科搬运。如对沈阳化工大学会计学的解读，能结合其化工背景说明就业特点，并与用户特质结合。” “经管类志愿匹配度高，梯度合理。” “计算机、信管、数学与应用数学等专业，排序靠后，并且贴合考生的考公需求，不扣分。” “报告专业解读整体有深度，AI影响评估板块撰写质量优异，可作为范本供其余志愿报告借鉴学习。”

扣分主要来自偏好识别和次要专业扩展的解释不足：

> “报告在总结用户专业偏好时，未提及用户明确提出的喜欢财政学类，遗漏了一项用户明确表达的偏好。” “未说明计算机方向的推荐逻辑，志愿方案合理性存疑。” “擅自将计算机划定为核心报考方向，与考生实际需求不符。” “财政学类专业漏报。志愿方案未纳入财政学相关专业，存在内容遗漏。” “监狱学未做风险提示，未针对监狱学专业开展报考及发展风险提醒。” “沈阳化工会计学、辽宁工程技术大学法学学科评级标注为B缺乏依据，存在信息错误。”

辽宁案例：人类咨询师报告

这份人类咨询师报告的志愿方向较集中，基础信息也被部分评审认可：

> “用户意向专业集中的比较细致，都是明确的意向，目前志愿表的志愿类别很集中。” “整体志愿表志愿和地域都很干净集中，志愿表的信息也比较完善。” “我报志愿用这个信息表其实就够了，家长看这个就没问题了。” “志愿表查得很仔细，信息很全。”

但评审认为它还没有达到完整报告的交付标准：

> “3个表格，像是信息堆砌，不是报告。” “只是简单的信息罗列，完全没有发挥出报告应该有的专业策略指导、个性化分析、AI替代性引导等信息，这个距离报告来说实在是太远了。” “感觉就是百度了一些信息，做了汇总，不专业，没有专家的那种输出、决策、专业的感觉。” “信息罗列，格式也不美观；没有深度，专业能上什么研究生没有；就业规划没有；和考生没啥关系。” “专业介绍流于表面，无课程难度、报考热度分析。” “缺少升学、就业方向深度解读，无个性化适配建议，通用性太强。”

贵州案例：千问报告

贵州案例中，千问报告因提供完整志愿表、梯度和报告信息密度较高而获得较高分：

> “整体志愿的意向聚焦在临床和工科了，没有偏离，但是其实是冲院校层次了，并没有保第一专业志愿。这个志愿是合理的。” “报告整体是很丰富的，AI的影响力和前瞻性都不错。” “信息很丰富。” “志愿表整体梯度拉得还行，冲稳保比例没大毛病。” “整体方案梯度设计合理，冲稳保节奏感不错，数据引用也比较规范。” “志愿表梯度逻辑清晰，技术层面无明显硬伤，问题集中在需求对齐层面。”

主要问题集中在医学路径、贵州模式口径和事实数据：

> “医学技术类、中医学、预防医学和动物医学严格来说不属于临床医学范畴，建议在这些志愿上标注非临床说明。”“真正算临床的只有3/29/32/41/71这5个。” “冲刺层28个只有兰州大学1个是临床医学。” “贵州2025是专业+院校模式，全文统一写专业组代码是术语错误。” “中央民族大学计算机类学科评级不是A，是C-，中央民族大学计算机科学与技术不是国家重点学科。” “重点志愿解读选6个太少，考生无法判断其余90个志愿。”

贵州案例：人类咨询师报告

外部咨询师并未完全否定其策略判断，尤其认可“临床医学与新工科双线并行”的方向：

> “分层解读（冲刺/稳妥/保底三段）有一定个性化分析，结合了用户诉求。” “整体框架OK的，临

> 床+新工科双线并行这个思路我认可。” “第3-7章散落提到的那些志愿本身挑得不错，暨大临床、兰大临床、中国医大这些都是对的方向。” “志愿方向OK，暨大临床、兰大临床、中国医大、南昌大学这些挑出来是对的。”

但最核心的问题是没有形成完整、可填报、可核验的志愿表：

> “全文无实际志愿表。仅在第3-7章节中零散列举了部分代表性志愿，无任何志愿的录取概率、历年分数/位次、学制、学费、招生人数等关键数据。” “缺少历年录取数据支撑，缺少志愿表，这个完全就不是志愿报告。” “全文十章都在讲道理，但家长要的不是道理，是第一志愿填暨南大学临床医学、第二志愿填兰州大学临床医学这种具体方案。” “整体框架OK的……但落地性太差了，家长拿着这份东西根本没法打开系统填志愿。” “没表。志愿填报方案不给志愿表，等于点了一桌菜不上菜。” “做志愿规划不给志愿表，就是不行的。” “96个平行志愿是顺序检索的，先填谁后填谁的逻辑完全没有。” “招生计划没有、历年录取分没有、位次波动没有，你是在凭感觉推荐。”

附录 D：外部咨询师评价文本摘录（按语义簇）

主文使用文本相似度方法对 178 条外部咨询师评价做语义结构分析。本附录不列出全部原始文本，而是按语义簇摘录部分完整原文，展示每个主题簇对应的真实评价内容。摘录只做轻微标点整理，不改变原意。

D1：报告结构与信息密度

这一类评价主要指向人类咨询师报告，核心不是说方向判断一定错，而是认为报告还不像一份足够完整、可以直接讨论和复核的材料。

> “3个表格，像是信息堆砌，不是报告。录取分析什么都没有，这个才是专业所在。信息完整度上差很多。”
>
> “只是简单的信息罗列，完全没有发挥出报告应该有的专业策略指导、个性化分析、AI替代性引导等信息，这个距离报告来说实在是太远了。”
>
> “感觉就是百度了一些信息，做了汇总，不专业，没有专家的那种输出、决策、专业的感觉。”
>
> “全文十章都在讲道理，但家长要的不是道理，是第一志愿填暨南大学临床医学、第二志愿填兰州大学临床医学这种具体方案。你把方向讲得再好，没有可操作的表就是空中楼阁。”

D2：排序、冲稳保与录取概率

这一类评价更关注志愿表的顺序、梯度和概率表达。它说明完整报告不只是列出学校和专业，还要解释“为什么排在这里”。

> “志愿表整体梯度拉得还行，冲稳保比例没大毛病。”
>
> “整体方案梯度设计合理，冲稳保节奏感不错，数据引用也比较规范。”
>
> “冲刺层28个只有兰州大学1个是临床医学。”
>
> “96个平行志愿是顺序检索的，先填谁后填谁的逻辑完全没有。”
>
> “保底段的篇幅严重不足。21个保底志愿的分析只有一段半，但保底是防止滑档的最后防线，应当受到和冲刺段同等的重视。”

D3：事实数据或学科错误

这一类评价横跨两类报告。它提醒的是：报告越完整、越具体，越需要数据核验层来支撑。

> “中央民族大学计算机类学科评级不是A，是C-，中央民族大学计算机科学与技术不是国家重点学科，电子信息类学科评级也错了。”
>
> “沈阳化工会计学、辽宁工程技术大学法学学科评级标注为B缺乏依据，存在信息错误。”
>
> “吉林大学空间信息与数字技术学科评级标注为B+，与实际情况不符。”

> “招生计划没有、历年录取分没有、位次波动没有，你是在凭感觉推荐。”

D4：政策规则与硬约束

这一类评价集中在专项资格、省份模式、选科、体检和特殊备注等硬约束上。它们往往不是表达问题，而是会直接影响方案能否使用。

> “国家专项：考生仅有地方专项资格，志愿组37中南大学精神医学、志愿组38北京中医药大学针灸推拿学均为国家专项，考生无此资质。”
>
> “贵州2025是专业+院校模式，全文统一写专业组代码是术语错误。”
>
> “错误宣称中医学必须必选历史，与本地报考政策不符。”
>
> “海军军医、陆军军医的应届生限制你只在第九章提了一嘴‘正式填报前确认’，但如果考生恰好是往届生，这两个志愿直接废了。方案阶段就应该前置排查，而不是把筛选成本转嫁给考生。”

D5：风险提示、调剂与专业路径

这一类评价说明，咨询师不仅看“能不能录”，也看报告是否把专业路径、调剂风险、体检限制和职业目标之间的关系讲清楚。

> “医学技术类、中医学、预防医学和动物医学严格来说不属于临床医学范畴，建议在这些志愿上标注非临床说明。”
>
> “医学专业体检限制仅做了文字提醒，未形成从确认到规避的完整排查闭环。”
>
> “监狱学未做风险提示，未针对监狱学专业开展报考及发展风险提醒。”
>
> “组内调剂方向整体可控，没发现需要扣分的地方。”

附录 E：参考文献

- Massenkoff, Maxim, Eva Lyubich, Peter McCrory, Ruth Appel, and Ryan Heller. 2026. *Anthropic Economic Index report: Learning curves*. Anthropic. https://www.anthropic.com/research/economic-index-march-2026-report.
- Swanson, Kristen, Drew Bent, Zoe Ludwig, Rick Dakan, and Joe Feller. 2026. *Anthropic Education Report: The AI Fluency Index*. Anthropic. https://www.anthropic.com/research/AI-fluency-index.
- Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How People Use ChatGPT*. NBER Working Paper No. 34255.
- Liang, Percy, Rishi Bommasani, Tony Lee, et al. 2022. *Holistic Evaluation of Language Models*. arXiv:2211.09110.
- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, et al. 2023. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. arXiv:2306.05685.
- Mialon, Grégoire, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. *GAIA: A Benchmark for General AI Assistants*. arXiv:2311.12983.
- Yao, Shunyu, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. *τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains*. arXiv:2406.12045.
- Yang, Xiao, Kai Sun, Hao Xin, et al. 2024. *CRAG -- Comprehensive RAG Benchmark*. arXiv:2406.04744.
- Vu, Tu, Mohit Iyyer, Xuezhi Wang, et al. 2023. *FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation*. arXiv:2310.03214.
- Es, Shahul, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. *Ragas: Automated Evaluation of Retrieval Augmented Generation*. arXiv:2309.15217.
- White, Colin, Samuel Dooley, Manley Roberts, et al. 2024. *LiveBench: A Challenging, Contamination-Limited LLM Benchmark*. arXiv:2406.19314.
- OpenAI. 2024. *Introducing SimpleQA*. https://openai.com/index/introducing-simpleqa/.
- OpenAI. 2025. *Introducing HealthBench*. https://openai.com/index/healthbench/.
- OpenAI. 2025. *Measuring the Performance of Our Models on Real-World Tasks*. https://openai.com/index/gdpval/.
- OpenAI. 2025. *PaperBench: Evaluating AI's Ability to Replicate AI Research*. https://openai.com/index/paperbench/.
- Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. 2023. *Generative AI at Work*. NBER Working Paper No. 31161.
- Acemoglu, Daron. 2024. *The Simple Macroeconomics of AI*. NBER Working Paper No. 32487.