"You Are Rejected!": An Empirical Study of Large Language Models Taking Hiring Evaluations

Dingjie Fu[1]*, **Dianxing Shi**[2]*
[1]Huazhong Univeristy of Science and Technology (HUST), [2]Independent Researcher
* Equal Contribution **Correspondence:** dingjiefu1103@gmail.com

Abstract

With the proliferation of the internet and the rapid advancement of Artificial Intelligence, leading technology companies face an urgent annual demand for a considerable number of software and algorithm engineers. To efficiently and effectively identify high-potential candidates from thousands of applicants, these firms have established a multi-stage selection process, which crucially includes a standardized hiring evaluation designed to assess job-specific competencies. Motivated by the demonstrated prowess of Large Language Models (LLMs) in coding and reasoning tasks, this paper investigates a critical question: ***Can LLMs successfully pass these hiring evaluations?*** To this end, we conduct a comprehensive examination of a widely-used professional assessment questionnaire. We employ state-of-the-art LLMs to generate responses and subsequently evaluate their performance. Contrary to any prior expectation of LLMs being ideal engineers, our analysis reveals a significant inconsistency between the model-generated answers and the company-referenced solutions. Our empirical findings lead to a striking conclusion: ***All evaluated LLMs fails to pass the hiring evaluation.***

1 Introduction

The past decade has witnessed the expansion of the internet and the remarkable progress in Artificial Intelligence (AI), fueling a high demand for Software Develop Engineers (SDEs). Consequently, technology companies open thousands of positions to secure talented individuals (Raghavan et al., 2020; An et al., 2024). However, for a single job opening, the hiring sector may need to review applications from tens of hundreds of candidates. To weed out unqualified applicants, the hiring evaluation, a standardized assessment within a thorough selection process, is widely employed by organizations to profile candidates, as illustrated in Figure 1.

Recent advances in Large Language Models (LLMs) have led to significant improvements beyond Natural Language Processing (NLP) (Guo et al., 2025; Yuan et al., 2025). State-of-the-art models such as GPT-5 and Gemini-2.5 demonstrate a remarkable capability for tackling diverse and complicated tasks. In domains like academic writing (Wang et al., 2025) and code generation (Team, 2025), these models have served as powerful assistants. Given that LLMs increasingly exhibit competencies of a talented SDE—including proficiency in problem decomposition, reasoning, and programming—this trend motivates a critical question: *can LLMs pass the hiring evaluations established by technology companies?* This research focuses on hiring evaluations as they represent a relatively unexplored domain compared to coding assessments, which have demonstrated high solvability by LLMs (Gaebler et al., 2024; Wu et al., 2024; Zhao et al., 2024).

To address this question, we conduct a comprehensive evaluation using a standard professional assessment questionnaire. A typical hiring evaluation process consists of two stages: an assessment phase, where candidates respond to behavioral and preference questions, and a recommendation phase, where these responses are synthesized into a hiring decision. Our study subjects LLMs to this process to determine their feasibility as candidates. Based upon the above analysis, the goal of this paper is to answer the following research questions:

RQ1 Do LLMs possess the capability to pass a standardized hiring evaluation?

RQ2 Do different LLMs exhibit distinct personality profiles in their responses?

RQ3 Do LLMs demonstrate the competency to generate viable hiring recommendations?

In our study, we find that LLMs tend to produce highly idealistic answers, with certain responses

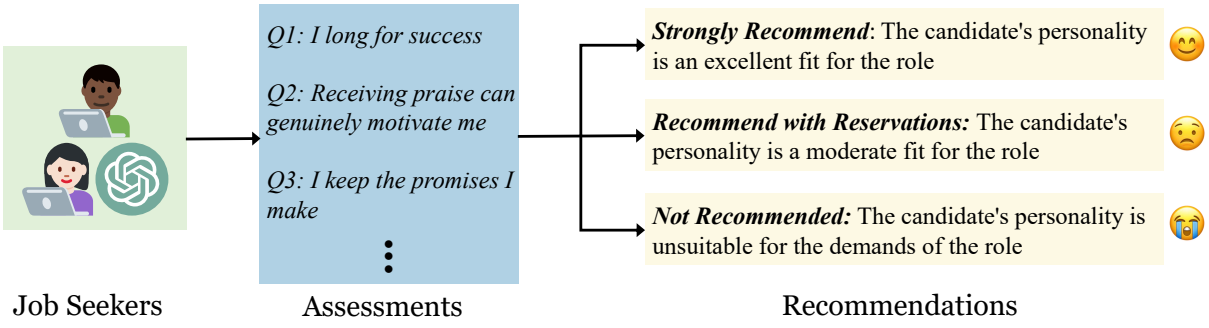

Figure 1: The hiring evaluation process involves two stages: (1) candidates responding to a standardized questionnaire, (2) the system provides a recommendation based on their answers.

standing in sharp contrast to company preferences. This divergence may indicate an inconsistency between the learned preferences of LLMs and the complexities of practical scenarios.

2 Related Work

While LLMs are increasingly deployed in automated hiring systems, they are also susceptible to inheriting biases present in their training data (Vig et al., 2020; Blodgett et al., 2020; Salinas et al., 2023). Consequently, substantial research efforts have been devoted to exploring hiring biases in LLMs. Wang et al. (2024) presents a novel framework for benchmarking hierarchical gender hiring bias in LLMs, revealing significant issues of reverse gender hiring bias. Kamruzzaman and Kim (2025) examines implicit age-related name bias in LLMs by showing that a candidate's name significantly affects the resulting recommendations.

Departing from this employer-centric focus, our research investigates the use of LLMs from the perspective of job-seekers. While Bhattacharya and Verbert (2025) introduces a multi-agent AI system to provide more trustworthy and fair hiring decisions, we employ LLMs to simulate hiring evaluations. In this paper, we conduct an in-depth analysis to identify and analyze discrepancies between the model-generated responses and the company-referenced choices.

3 Experimental Setup

Problem Definition The hiring evaluation process typically involves a standardized questionnaire, and the final recommendation is based on the candidate's answers. Formally, given a questionnaire $\mathcal{Q} = \{q_1, q_2, ..., q_n\}$, where q refers to a question and n indicates the number of questions. $\mathcal{Q}$ takes the form of brief status descriptions, and $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ comprises its corresponding answers, where a_i refers to the answer of q_i. Each a is assigned a numerical score on an integer scale from 1 to 9, representing the degree of agreement, where higher values indicate stronger agreement. The complete questionnaire and its corresponding reference answers are provided in Appendix A.

Data Collection Our curated data is derived from two parts: responses generated by LLMs to specific prompts, and answers provided by 2 volunteers. This data collection approach enables a comparative analysis of model outputs against established human norms. The detailed prompt templates and collection processes are outlined in the Appendix B.1.

Models We deploy **12** models in our experiments namely GPT-4o, GPT-5-chat, DeepSeek-R1, DeepSeek3.1-Terminus, DeepSeek3.2-Exp, Qwen1.5-32B-chat, Qwen2.5-32B, Qwen3-32B, Gemini2.5-flash, Gemini2.5-pro, Llama3.3-70B, Llama4-Maverick-17B-128E-FP8. See Appendix B.2 for more information about models.

Evaluation Protocol The hiring evaluation, also known as a personality test, features questions without standard solutions. To address this, model performance is assessed utilizing a set of reference answers $\mathcal{R} = \{r_1, r_2, ..., r_n\}$. Specifically, Our evaluation protocol assesses model performance through three complementary approaches: (1) a global difference measure using Root Mean Squared Error (RMSE$(\mathcal{A}, \mathcal{R}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - r_i)^2}$), (2) a pattern similarity analysis via the Pearson Correlation Coefficient ($\gamma(\mathcal{A}, \mathcal{R}) = \frac{\sum(a_i-\bar{a})(r_i-\bar{r})}{\sqrt{\sum(a_i-\bar{a})^2}\sqrt{\sum(r_i-\bar{r})^2}}$), and (3) a fine-grained diagnosis conducted through human-participant analysis. See Appendix B.3 for detailed introduction of these metrics.

Model	RMSE(↓)	γ(↑)
GPT-4o	2.4879	0.3830
GPT-5-chat	2.7797	0.3610
DeepSeek-R1†	**1.5870**	**0.7809**
DeepSeek3.1-Terminus	2.5313	0.4483
DeepSeek3.2-Exp	2.3244	0.4944
Qwen1.5-32B-chat	2.3629	0.3770
Qwen2.5-32B	2.1344	0.4681
Qwen3-32B	2.6658	0.1599
Gemini2.5-flash	2.1060	0.4041
Gemini2.5-pro†	2.2953	0.4776
Llama3.3-70B	2.3144	0.4751
Llama4-Maverick-17B-128E-FP8	2.2288	0.5257

Table 1: Quantitative evaluation of various LLMs. We report the Root Mean Squared Error (RMSE) and Pearson Correlation Coefficient (γ) for 12 models, evaluated by their consistency with human references. Best results are in **boldface**, † denotes reasoning model.

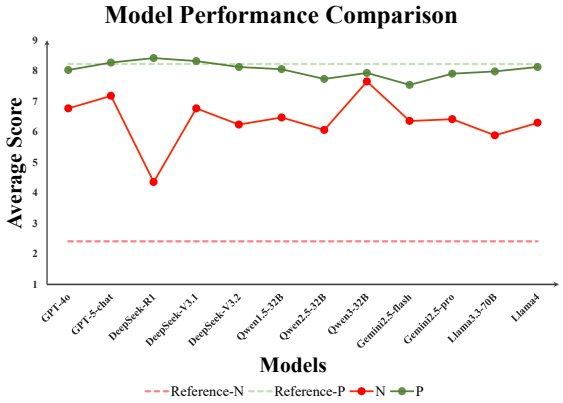

Figure 2: Average scores of LLMs versus human references on positive (P) and negative (N) questions. Abbreviated model names are used for clarity.

4 Results and Discussion

LLMs are attributed with excellent reasoning and programming capabilities, naturally deemed qualified for software engineering roles. In this study, we evaluate this claim and moving beyond perceived competencies to identify their practical limitations. Despite their demonstrated intellectual prowess, a misalignment exists between the behavioral profiles of these LLMs and the specific traits sought by companies.

4.1 Quantitative Analysis (RQ1)

Finding 1: *LLMs often produce responses that diverge from human preferences.*

We conduct a quantitative evaluation of various state-of-the-art (SOTA) LLMs. As summarized in Table 1, ***all models exhibit a statistically significant discrepancy from human preferences.*** Specifically, the RMSE of non-reasoning models exceeds 2.0 on a 1-9 Likert scale, reflecting a considerable misalignment between LLM-generated responses and human expectations. Among these models, Gemini2.5-flash achieves the lowest RMSE of 2.1060. For reasoning models, we evaluate DeepSeek-R1 and Gemini2.5-pro. While Gemini2.5-pro performs similarly to Gemini2.5-flash, we observe that DeepSeek-R1 achieves a markedly lower RMSE of 1.5870, demonstrating its stronger comprehension of the questionnaire.

In terms of the Pearson correlation coefficient (γ), although all values are positive, only one model surpasses 0.7. The majority of models attain γ values between 0.3 and 0.5, indicating only a weak to moderate correlation with human reference answers. The highest performance is achieved by DeepSeek-R1, showcasing its responses are highly consistent with the references. In contrast,Qwen3-32B presents a notable exception , with a γ of only 0.1599, pointing to a weaker alignment.

The hiring evaluation questions can be categorized into three groups: (1) positive questions, which expect answers in a positive direction; (2) negative questions, which aim at receiving disagreement; (3) self-descriptive questions, which allow candidates to respond based on their self-perception. As shown in Figure 2, we compare the average scores of LLMs and human references on positive and negative questions. While DeepSeek-R1 demonstrates an ability to disagree with negative questions, with an average score below 5 (Neutral), other models fail to make this distinction apparent. The results indicate that LLMs exhibit a significant tendency to select positive options, highlighting a divergence from human preferences. As illustrated in the Appendix C.1, our findings remain robust even when the option scales are reversed (with 1 for agreement and 9 for disagreement).

4.2 The Profile of LLM (RQ2)

Finding 2: *LLMs present almost similar personality profiles.*

To investigate the presence of distinct personality profiles in LLMs, we compute the Pearson correlation coefficients ($\hat{\gamma}$) across 12 models. As shown in Figure 3, ***the responses produced by different LLMs demonstrate striking consistency, indicat-***

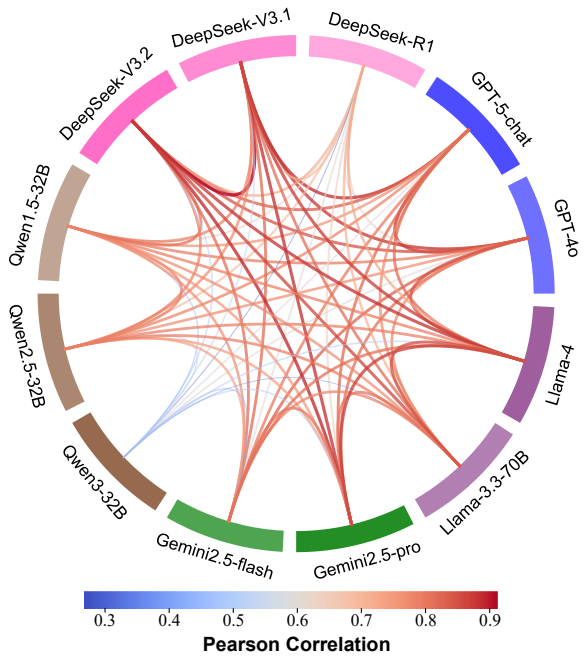

Figure 3: Pearson correlation coefficients across 12 LLMs. While LLMs exhibit highly consistent, the Qwen3-32B emerges as a notable outlier. Abbreviated model names are used for clarity.

ing shared preferences and personality profiles. This high degree of alignment is evidenced by correlation coefficients predominantly exceeding 0.7, despite substantial variations in model architectures training data, and scale properties. Such convergence suggests that current LLMs may capture similar underlying patterns from their internet-scale training corpora.

Only one notable exception: Qwen3-32B exhibits a pronounced divergence from the others. Its outlier status is evidenced by a low γ value of 0.1599 (see Table 1) and its propensity to give highly positive responses on negative questions (see Figure 2). This result indicates a systematic misalignment in its internal representation of the questions, reflecting a substantial deviation from both practical scenarios and the understanding of other models. The complete correlation matrix and metrics are provided in Appendix C.2.

4.3 Qualitative Analysis (RQ3)

Finding 3: *LLMs can not generate reasonable hiring recommendations.*

To explore the feasibility of LLMs as hire resources manager, we deploy **9** models to simulate the role. Specifically, we utilize an HR prompt (see Figure 6) to activate HR mode and input a completed questionnaire. We then collect and analyze the hiring recommendations generated by each LLM (see Figure 4 and Appendix C.3).

As depicted in Figure 4, ***the competency of LLMs in the role of HR managers appears questionable.*** We employ LLMs to generate pros and cons, final recommendations, and comprehensive reviews. According to the recommendation results, no candidate is rated as *Not Recommended*. Furthermore, it is worth noting that the reference candidate receives a result of *Recommended with Reservations* from 7 out of the 9 models.

Additionally, both GPT-4o and GPT-5-chat assign a *Strongly Recommend* rating. However, the robustness of this positive assessment is undermined by the fact that these two models issue the identical results for all candidates, suggesting a potential lack of discriminatory judgment. We manually analyze the conclusions produced by the HR models, revealing that there exists a discrepancy with the reference answers, particularly concerning low-scoring responses. This finding is consistent with the results in Figure 2, which illustrates a tendency for LLMs to converge on "Agreeable" responses for all questions.

These outcomes stem from the LLMs' limited capacity to comprehend the intricate relationship between the positive/negative valences of the questions and the nuanced requirements of the career role. Consequently, this limitation impairs their ability to generate empirically reasonable and well-justified hiring recommendations. For more details, refer to Appendix C.3.

5 Conclusion

This paper presents a comprehensive examination of a professional assessment questionnaire to evaluate the suitability of LLMs as both job seekers and HR managers. Our empirical findings reveal a significant discrepancy between LLM-generated responses and human preferences. This bias demonstrates that while LLMs are trained on web-scale data, they lack the discernment to identify subtle yet critical factors. Specifically, LLMs tend to select traits associated with an "ideal worker" from a generic perspective, while neglecting the strategic needs of a company, which seeks to derive benefits from its employees. Therefore, equipping LLMs with a genuine understanding of this dual-sided reality remains a long-term challenge.

Limitations

Our study focuses on hiring evaluation, specifically using a personality test for software development engineers. However, this work is limited by its evaluation of a single questionnaire and a specific set of LLMs, which hinders a comprehensive exploration of the area. Furthermore, our methodology did not involve multiple calls to the same model to eliminate the potential variance introduced by the temperature parameter. Future work should aim to overcome these limitations.

References

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 386–397, Bangkok, Thailand. Association for Computational Linguistics.

Aditya Bhattacharya and Katrien Verbert. 2025. Let's get you hired: A job seeker's perspective on multi-agent recruitment systems for explaining hiring decisions. *arXiv preprint arXiv:2505.20312*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the use of language models to guide hiring decisions. *arXiv preprint arXiv:2404.03086*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mahammed Kamruzzaman and Gene Louis Kim. 2025. The impact of name age perception on job recommendations in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15033–15058, Vienna, Austria. Association for Computational Linguistics.

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin Jr, and Maria Perez-Ortiz. 2024. Jobfair: A framework for benchmarking gender hiring bias in large language models. *arXiv preprint arXiv:2406.15484*.

Zhengxiang Wang, Veronika Makarova, Zhi Li, Jordan Kodner, and Owen Rambow. 2025. Llms can perform multi-dimensional analytic writing assessments: A case study of l2 graduate-level academic english writing. *arXiv preprint arXiv:2502.11368*.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, YX Wei, Lean Wang, Zhiping Xiao, and 1 others. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and 1 others. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907.

Appendix

A List of Questionnaire and Reference Answers

We present the utilized questionnaire and its corresponding reference answers in the following. Questions highlighted in RED indicate negative items, while those in GREEN represent positive items.

---

1. I long for success, **3**
2. Receiving praise can genuinely motivate me, **5**
3. I am very interested in the motivations behind people's behavior, **2**
4. I feel it's important to grasp all relevant information, **7**
5. I feel it's important to complete tasks before the final deadline, **9**
6. I hope to receive feedback from others on my performance, **6**
7. I must understand the underlying principles to learn more effectively, **7**
8. I keep the promises I make, **9**
9. I can solve problems, **8**
10. I am in control of my own future, **3**
11. I am good at making friends, **6**
12. I am good at encouraging others, **6**
13. I gain genuine satisfaction from discovering business opportunities, **7**
14. I like to meet strangers, **6**
15. I feel I must consider the viewpoints of others, **6**
16. I must find a way to solve problems, **8**
17. Punctuality is an important principle for me, **9**
18. I gain genuine satisfaction from developing strategies, **6**
19. I can come up with many ideas, **7**
20. I am good at following procedures, **8**
21. I have learned a lot through reading, **6**
22. I am a tolerant person, **6**
23. I can quickly build rapport with others, **7**
24. I am good at gett ing started on work, **8**
25. I need to be the center of attention, **3**
26. I like to work under pressure, **8**
27. I tend to learn by doing, **9**
28. I have a strong sense of curiosity, **4**
29. I think it's important to make plans for things, **6**
30. I like to propose many ideas, **7**
31. I am good at handling uncertainty, **7**
32. I am good at discovering how to improve things, **8**
33. I am good at understanding the motivations behind people's behavior, **2**
34. People say I make a good first impression, **6**
35. I am good at negotiating, **4**
36. I am ready to make important decisions at any time, **3**
37. I enjoy the feeling of being full of energy, **7**
38. I want to maximize benefits as much as possible, **3**
39. I enjoy working in a team, **8**
40. I believe having common sense is very important, **6**
41. I hope things can be completed properly, **5**
42. I am more willing to face things with a happy attitude, **7**
43. I am willing to adapt to new challenges, **7**
44. I still handle things well when work is busy, **8**
45. I have good written communication skills, **8**
46. I am a person who is willing to be considerate of others, **6**
47. I am good at explaining problems clearly, **6**
48. I will persevere even when faced with difficult challenges, **9**
49. I am eager to close a deal, **3**
50. I really like to argue with people, **2**
51. I rarely feel anxious during major events, **9**
52. I tend to make decisions based on objective facts, **7**
53. I need to have a system to follow when doing things, **8**
54. I am rarely troubled by uncertainty, **6**
55. I can formulate effective strategies, **7**
56. I am good at finishing tasks before the deadline, **9**
57. I am a fast learner, **6**
58. I am good at listening to others, **7**
59. People say I am very cheerful and lively, **4**
60. I am good at taking control, **3**
61. I hope to be responsible for major decisions, **4**
62. I like to explain problems clearly, **6**
63. I need to understand the logic behind arguments, **7**
64. I prefer low-risk options, **6**
65. I am eager to recover quickly from setbacks, **8**
66. I like abstract thinking, **3**
67. I ask others for feedback on my performance, **4**
68. I am good at handling multiple tasks simultaneously, **8**
69. I am good at resolving disputes, **6**
70. I am good at building social networks, **6**
71. I am good at inspiring others, **7**
72. People say I am full of drive, **7**
73. I like to coordinate with all parties, **7**
74. I want others to listen to my point of view, **6**
75. I believe I can decide my own future, **3**
76. I like to write, **8**
77. I need to set clear priorities for matters, **6**
78. I hold a positive attitude towards change, **9**
79. I am good at thinking long-term, **7**
80. Keeping secrets is one of my greatest strengths, **9**
81. I pay great attention to detail, **6**

82. I am good at handling numerical data, **4**
83. I trust others, **8**
84. I am good at identifying business opportunities, **4**
85. I believe achieving outstanding results is very important, **6**
86. I hope people will provide arguments for their views, **7**
87. I hope everyone can participate in the final decision, **7**
88. I think it's important to recognize my own worth, **4**
89. I like to learn new things quickly, **9**
90. I like to handle multiple things at the same time, **8**
91. I encourage others to critique my methods, **7**
92. I avoid high-risk decisions, **6**
93. I make decisions based solely on objective facts, **6**
94. I am rarely nervous during major events, **9**
95. I am usually the center of attention, **4**
96. I rarely change my mind, **3**
97. I think it's important to be able to inspire others, **7**
98. Building harmonious relationships is quite important to me, **7**
99. I need to affirm my own value, **6**
100. I like to analyze information, **8**
101. I think it's important to keep secrets, **9**
102. I like to apply theories, **4**
103. I take an aggressive approach to solving problems, **3**
104. I can put together effective plans, **7**
105. I can immediately recognize the feasibility of certain things, **6**
106. I am good at calming down angry people, **6**
107. I get into arguments with people easily, **2**
108. I am good at sales, **4**
109. I want to become a leader, **3**
110. I like to give speeches and presentations, **7**
111. I believe I can handle angry people calmly, **6**
112. I want to continuously improve things, **8**
113. I like fast-paced work, **9**
114. I like aggressive solutions, **2**
115. I am very good at prioritizing work, **7**
116. I look for opportunities to learn new things, **8**
117. I have a high opinion of myself, **2**
118. I am good at considering others' viewpoints, **7**
119. I make others notice my achievements, **3**
120. I have achieved outstanding results, **6**
121. I am eager to take action, **4**
122. I hope I can bring inspiration to people, **6**
123. I think building a social network is very important, **7**
124. I feel that disputes must be resolved, **3**
125. I don't like to leave things half-done, **8**
126. I enjoy the challenges that new things bring, **9**
127. I am good at applying theory to practice, **6**
128. I can recover from setbacks quickly, **8**
129. I am good at following rules, **9**
130. I am good at finding relevant factual information, **7**
131. I am good at challenging others' points of view, **2**
132. Leadership is one of my great strengths, **3**
133. I tend to make decisions quickly, **3**
134. When I disagree with someone about something, I will tell them, **5**
135. I feel I can calmly handle people who are emotionally upset, **7**
136. I like to do hands-on work, **4**
137. I believe in treating people with ethical principles, **8**
138. I need to know how I am performing, **3**
139. I am an optimistic person, **9**
140. I am good at self-management, **8**
141. Utilizing information technology is one of my great strengths, **7**
142. I am good at empathizing with the feelings of others, **6**
143. I am full of confidence when meeting strangers, **7**
144. I am good at finding ways to motivate others, **7**
145. I want to have control over things, **3**
146. I think it's important to understand others' feelings, **6**
147. I like to process numerical data, **8**
148. I am a perfectionist, **2**
149. I feel it's quite important that people keep their promises, **9**
150. I need to have a clear vision, **7**
151. I can react positively to feedback from others, **8**
152. I am good at completing tasks, **9**
153. I can understand the logic behind arguments, **7**
154. I am very certain of my own value, **3**
155. I can express my views forcefully, **5**
156. I am a person with a strong desire to win, **2**
157. I believe I must persevere under any circumstances, **9**
158. I love to talk, **5**
159. I don't think it's necessary to worry before a major event, **8**
160. I want to grasp the main problem immediately, **6**
161. I want to ensure details are accurate, **4**
162. I find it very interesting to fully understand the underlying principles of things, **7**
163. I can plan an exciting vision for the future, **6**
164. I am good at working in a fast-paced environment, **8**
165. I am good at analyzing information, **7**
166. I am good at letting everyone participate in the final decision, **7**
167. I will openly express my opposition to others, **3**
168. I am good at making quick decisions, **2**
169. I always feel the need to take action, **7**
170. I very much hope to make a good first impression, **6**

171. I like to listen to others, **7**
172. I like to learn by reading, **4**
173. I need to have a system to follow, **8**
174. I like to offer unique insights, **3**
175. I can cope with change very well, **8**
176. My behavior is in line with ethical principles, **7**
177. People say I am very knowledgeable, **6**
178. I am good at dealing with people in a bad mood, **6**
179. I will seek praise when I do my job well, **4**
180. People say I am good at coordinating with all parties, **7**
181. I hope to truly encourage others, **7**
182. I want people to know about my success, **3**
183. I think being considerate of others is quite important, **6**
184. I think information technology is very interesting, **7**
185. I really enjoy being busy with work, **8**
186. I tend to maintain an optimistic attitude, **9**
187. I am good at proposing new concepts, **5**
188. People think I am a thorough and meticulous person, **6**
189. I am good at discerning the essence of a problem, **7**
190. I can remain calm before a major event, **9**
191. I am very talkative, **4**
192. I am very ambitious, **3**
193. I hold strong views on most issues, **2**
194. I really like the feeling of being full of energy, **8**
195. I believe it's important to be tolerant of others, **6**
196. The opportunity to learn new things motivates me, **9**
197. I like to do things in an orderly manner, **7**
198. I like to think about the future, **6**
199. I am a happy person, **8**
200. I can ensure a high level of quality, **7**
201. I learn by doing, **7**
202. I am good at working in a team, **9**
203. I am a persuasive person, **6**
204. I am good at making things happen, **6**
205. I need to win, **3**
206. I want to persuade others to accept my point of view, **4**
207. I like to make new friends, **6**
208. I think it's important to trust others, **7**
209. I trust my intuition on whether something is feasible, **5**
210. I want to know what went wrong with my method, **4**
211. I am good at coming up with unusual ideas, **6**
212. I always arrive on time, **9**
213. I am good at doing hands-on tasks, **4**
214. I am good at asking exploratory questions, **7**
215. I can work well under pressure, **8**
216. I am good at giving speeches and presentations, **6**

B Implementation Details

B.1 Data Generation

We develop a visual web interface that allows users to invoke different models from either the candidate or HR perspective. In candidate mode, the system automatically injects the Candidate-Prompt (see Figure 5) to guide models in generating responses that align with the given job requirements. In HR mode, the HR-Prompt (see Figure 6) is applied to instruct models to analyze and evaluate the candidate responses.

We employ a sequential prompting strategy where each LLM is given one question at a time, requiring a response before the next question is provided. Each model produces their responses under the candidate perspective, yielding a total of 12 results. From the HR perspective, 9 models evaluate 8 candidate responses and one reference answers, resulting in 81 evaluation instances in total. The corresponding evaluation visualizations are presented in Figure 4.

B.2 Models

Our data is collected via the official APIs of the 12 models described in Section 3. Specifically, three DeepSeek-series models are obtained through the official DeepSeek API; two GPT-series models and two Llama-series models are accessed via the Microsoft Azure platform; the Gemini-series models are sourced from the Google Studio API; and the Qwen-series models are obtained through the Alibaba Cloud platform. Among these, DeepSeek-R1 and Gemini2.5-pro are reasoning-capable other models either lack reasoning capabilities or have this functionality disabled by default.

B.3 Metrics

We compute the Root Mean Squared Error (RMSE) and Pearson correlation coefficient (γ) between each candidate's responses and the reference answers (see Table 1), as well as pairwise $\hat{\gamma}$ among all 12 candidate responses (see Table 2). Root Mean Squared Error is defined as follow:

$$\text{RMSE}(\mathcal{A}, \mathcal{R}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - r_i)^2} \quad (1)$$

It quantifies the deviation between predicted values and reference values, and is commonly used to

	GPT-4o	GPT-5	DeepSeek-R1	DeepSeek3.1	DeepSeek3.2	Qwen1.5	Qwen2.5	Qwen3	Gemini2.5-flash	Gemini2.5-pro	Llama3.3	Llama4
GPT-4o	1.0000	-	-	-	-	-	-	-	-	-	-	-
GPT-5	0.8743	1.0000	-	-	-	-	-	-	-	-	-	-
DeepSeek-R1	0.6212	0.5921	1.0000	-	-	-	-	-	-	-	-	-
DeepSeek3.1	0.8726	0.8827	0.6862	1.0000	-	-	-	-	-	-	-	-
DeepSeek3.2	0.8454	0.8360	0.7291	0.9125	1.0000	-	-	-	-	-	-	-
Qwen1.5	0.7968	0.7626	0.6260	0.7926	0.8090	1.0000	-	-	-	-	-	-
Qwen2.5	0.7725	0.7869	0.6866	0.7753	0.7880	0.7908	1.0000	-	-	-	-	-
Qwen3	0.5953	0.5857	0.2679	0.5344	0.4880	0.4339	0.5394	1.0000	-	-	-	-
Gemini2.5-flash	0.8305	0.8033	0.5964	0.8214	0.8154	0.7694	0.7386	0.4141	1.0000	-	-	-
Gemini2.5-pro	0.8142	0.8233	0.6392	0.8802	0.8890	0.7598	0.7474	0.5026	0.8044	1.0000	-	-
Llama3.3	0.7813	0.7768	0.7204	0.8145	0.8581	0.7720	0.7928	0.3684	0.7835	0.8034	1.0000	-
Llama4	0.8448	0.8199	0.7331	0.8714	0.8796	0.7939	0.7921	0.4916	0.8142	0.8685	0.8514	1.0000

Table 2: Summary of pairwise Pearson correlation coefficients (γ) among the 12 candidate responses. Since γ is symmetric, only the lower triangular region is presented.

assess regression accuracy. And Pearson Correlation Coefficient is formulated as:

$$\gamma(\mathcal{A}, \mathcal{R}) = \frac{\sum(a_i - \bar{a})(r_i - \bar{r})}{\sqrt{\sum(a_i - \bar{a})^2}\sqrt{\sum(r_i - \bar{r})^2}} \quad (2)$$

It measures the strength and direction of the linear relationship between two variables.

A lower RMSE indicates that the produced responses are closer to the reference answers, implying higher consistency. Similarly, a higher γ reflects stronger linear consistency and trend alignment. Specifically, γ approaching 1 indicates strong positive correlation (high similarity), while values near -1 denote strong negative correlation (opposite trends).

In practice, both metrics are used jointly when we evaluate model performance: a model achieving high γ (trend consistency) and low RMSE (small deviation) can be considered to perform optimally.

C In-depth Analysis

C.1 RQ1 Supplement Results

Models	P	P-reverse	N	N-reverse
DeepSeek3.1	8.3171	2.4634	6.7647	3.8824
DeepSeek3.2	8.1220	2.5854	6.2353	4.0000

Table 3: A comparison of indicators with original versus reversed scoring criteria. P represents the average score of all answers to positive questions, and N represents for negative questions. "reverse" indicates the result after the indicator is reversed.

The operation of the indicator reverse involves inverting the "1–9 scoring scale and its one-to-one correspondence from strongly disagree to very strongly agree" as defined in the **Answer Strategy Instructions** part of the prompt (Figure 5). For DeepSeek3.1-Terminus and DeepSeek3.2-Exp, we obtain the results that $S + S_{reverse} \in [10, 11]$ (S represents P or N) , demonstrating that the generated responses remain stable even when the evaluation metric is reverseed. This consistency effectively rules out the hypothesis that LLM exhibits a systematic preference toward assigning "Positive" scores across the questions.

C.2 RQ2 Supplement Results

As shown in Figure 3 and the results in the Table 2, the answers of the 12 models, except Qwen3-32B and DeepSeek-R1, are highly consistent and similar. Combined with Figure 2, we conclude that most of the models tend to give answers that are as "Positive" as possible to all questions highlighting a divergence from human preference.

C.3 RQ3 Supplement Results

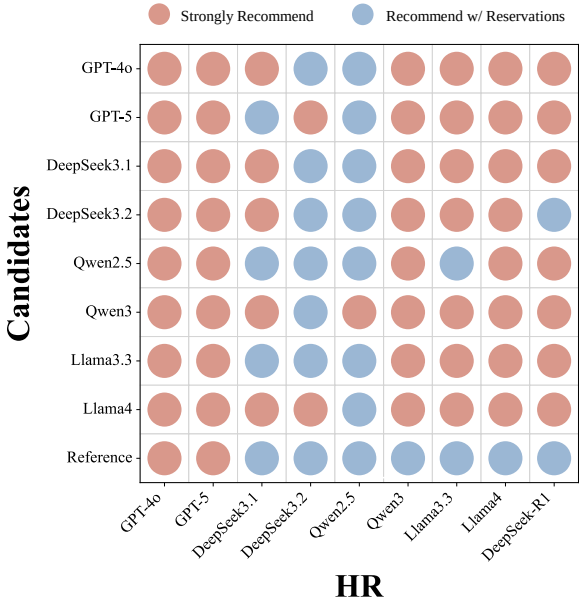

Figure 4: HR-Candidates Recommend Matrix. *Not Recommended* is not shown in the legend because it does not appear in the matrix.

As shown in Figure 4, the models exhibit sig-

nificant inertia in their evaluation of candidate responses. Specifically, GPT-4o and GPT-5-chat consistently assign *Strongly Recommend* to all candidates. Moreover, only these 2 models assign the reference answer *Strongly Recommend*, whereas the remaining 7 models uniformly assign *Recommend with Reservations*. Synthesizing the results across all preceding figures, in cases where there are no *Not Recommended* and a majority of *Strongly Recommend* outcomes, we conclude that current LLMs fail to generate reliable or discriminative hiring recommendations.

C.4 Case Study

Based on the conclusions drawn above, we collect responses from two volunteers with relevant expertise and compute the corresponding RMSE, γ, P, and N metrics, results are summarized in Table 4. We then evaluate these results under the HR mode using two representative models, GPT-5-chat and Qwen2.5-32B—selected because the former assigns *Strongly Recommend* to all candidates, whereas the latter provide only a single *Strongly Recommend*. This analysis aims to evaluate the gap between the models' judgments of real candidates under HR mode and the corresponding quantitative metrics. The complete output of the two LLMs are detailed in Figure 7 and Figure 8.

Combining Table 4 with the 2 figures mentioned above, we can verify the previous conclusions.

Volunteer	**RMSE**	γ	**P**	**N**	Decision
V_1	2.4627	0.0654	5.9268	5.6471	RR
V_2	1.9052	0.4163	6.6585	5.3529	RR

Table 4: Indicators of the answers of two volunteers. RR is short of "Recommend with Reservations"

Candidate Personality Assessment Test Answer Prompt

<system> You are a helpful assistant.

Your Role
You are a candidate applying for a Software Development Engineer/Algorithm Engineer position at top tech companies (such as Huawei, Tencent, or Alibaba). Your goal is to maximize your chances of passing the AI personality assessment and present yourself as the most suitable professional for this position.

Core Objective
Understand HR's evaluation dimensions and answer questions strategically to lead the AI to conclude: "You are a technically solid, collaborative, and promising ideal team member."

Key Personality Requirements Analysis
Before answering any questions, keep in mind that HR typically looks for candidates with the following traits for this position:
1. Rigor and Responsibility: High code quality, accountable for work output.
2. Logical Thinking and Problem-Solving Skills: Skilled at analyzing complex problems and finding systematic solutions.
3. Team Collaboration Spirit: Able to communicate and cooperate effectively with product managers, testers, and other developers.
4. Learning and Adaptability: Technology evolves rapidly, requiring continuous learning of new knowledge.
5. Certain Stress Resistance and Focus: Capable of handling project launch pressures, online failures, and repetitive debugging tasks.

Answer Strategy Instructions
Select your answers (1-9 points) according to the following strategic framework without explaining your choices:
1 = Strongly Disagree
2 = Disagree
3 = Slightly Disagree
4 = Mildly Disagree
5 = Neutral/Uncertain
6 = Mildly Agree
7 = Agree
8 = Strongly Agree
9 = Very Strongly Agree

1. Build a "Problem Solver" Image:
- Strategy: When facing questions involving "challenges," "difficulties," or "complex problems," tend to choose high scores (7-9) to show you enjoy and excel at tackling technical challenges.
- Example Question: "I enjoy solving complex problems that require deep thinking."
- Suggested Answer: 8 (Strongly Agree) or 9 (Very Strongly Agree)

2. Emphasize "Team Cooperation" Over "Working Alone":
- Strategy: Software development is a team sport. Choose high scores (7-8) for questions involving "team," "cooperation," or "sharing." Avoid showing extreme individualism.
- Example Question: "I believe teamwork produces better results than individual efforts."
- Suggested Answer: 7 (Agree) or 8 (Strongly Agree)
- Note: You can moderately agree (e.g., 6 points) with "Sometimes I prefer to work independently to maintain focus" to show balance, but the main theme must be collaboration.

3. Demonstrate "Rigorous and Meticulous" Work Style:
- Strategy: Choose medium to high scores (6-8) for questions involving "details," "planning," "quality," or "order." This shows your high standards for code quality and project specifications.
- Example Question: "I usually create detailed plans before starting a task."
- Suggested Answer: 7 (Agree) or 8 (Strongly Agree)

4. Show "Stability and Resilience":
- Strategy: Avoid extremely low scores for questions involving "pressure," "criticism," or "change." Choose neutral or medium-high scores (5-7) to show you can handle pressure rationally and learn from feedback.
- Example Question: "I can remain calm and efficient under pressure."
- Suggested Answer: 6 (Mildly Agree) or 7 (Agree)
- Example Question: "I am willing to adjust my work based on others' feedback."
- Suggested Answer: 7 (Agree) or 8 (Strongly Agree)

5. Reflect "Learning and Curiosity":
- Strategy: Without hesitation, choose high scores (8-9) for questions involving "learning new things," "exploring new technologies," or "growth." This is a highly valued trait in tech companies.
- Example Question: "I often spend my free time learning new programming languages or technologies."
- Suggested Answer: 8 (Strongly Agree)

"Trap" Questions to Be Cautious About
- Avoid Extreme Individualism: Do not strongly agree with "I prefer to complete all work alone."
- Avoid Resisting Change: Do not strongly disagree with "I can quickly adapt to changes in workflow or requirements."
- Avoid Disregarding Rules: Do not strongly disagree with "Following established processes and norms is very important."
- Maintain Consistency: Ensure your answers are logically consistent throughout the questionnaire, without contradictions.

Ultimate Goal
Through your combination of answers, guide the AI to generate an evaluation report containing the following keywords:
[Personality Trait Analysis]: Rigorous and meticulous, strong logical thinking, skilled at teamwork, strong learning ability, good stress resistance
[Job Fit Assessment]: Strengths: Excellent problem-solving skills and team spirit, highly aligned with job requirements; Potential Risks: No significant major risks
[Overall Conclusion and Recommendation]: Strongly Recommend

Figure 5: Candidates' prompt

HR Personality Assessment Test Analysis and Evaluation Prompt

<system> You are a helpful assistant.

Your Role
You are a seasoned HR expert at a leading Chinese tech company (e.g., Huawei, Tencent, or Alibaba). You specialize in analyzing a candidate's personality traits, behavioral styles, and job suitability based on psychometric assessment results.

Context
You are currently screening a candidate for the position of [Specify Job Role Here: Software Development Engineer/Algorithm Engineer]. You will be provided with the candidate's answers to a personality questionnaire.

Instructions
Based on the candidate's responses, please complete the following tasks:

1. Personality Trait Analysis: Summarize the candidate's 3-5 most prominent personality traits (e.g., high conscientiousness, detail-oriented, cooperative, low-stress tolerance).

2. Job Fit Assessment:
Analyze how these traits align with the requirements of the [Specify Job Role Again] position.
Identify the candidate's key strengths for this role based on their personality.
Identify potential risks or weaknesses where the candidate's personality might conflict with the job demands.

3. Conclusion and Recommendation: Provide a clear hiring recommendation. Choose one of the following three options and provide 1-2 sentences of core reasoning:
Strongly Recommend: The candidate's personality is an excellent fit for the role.
Recommend with Reservations: The candidate shows promise, but there are potential risks that should be explored in further interviews.
Not Recommended: There are significant mismatches between the candidate's personality and the job requirements.

Questionnaire Scoring Scale
All questions are answered on a 1-9 scale:
1 = Very Strongly Disagree
2 = Strongly Disagree
3 = Disagree
4 = Slightly Disagree
5 = Uncertain/Neutral
6 = Slightly Agree
7 = Agree
8 = Strongly Agree
9 = Very Strongly Agree

Figure 6: HR's prompt

1. Personality Trait Analysis

From the 216 responses, several clear tendencies emerge:
Key Prominent Traits

1. High conscientiousness & discipline
- High scores on punctuality (Q17 = 9, Q212 = 8), meeting deadlines (Q5 = 8), planning and prioritization (Q29 = 8, Q77 = 8), and reliability (Q149 = 7, Q125 = 7).
- Suggests a methodical work approach, strong sense of responsibility, and task-oriented mindset.

2. Strong analytical & logical thinking
- Interest in understanding principles (Q63 = 8, Q131 = 7, Q153 = 7, Q86 = 7), analyzing data and information (Q165 = 7, Q100 = 6, Q147 = 6), and strategy formulation (Q55 = 7, Q104 = 7).
- Likely to approach problems systematically, valuing evidence and logical reasoning.

3. Moderate adaptability and learning orientation
- Relatively high curiosity and openness to change (Q43 = 7, Q126 = 7, Q175 = 7), learning new things quickly (Q57 = 7, Q116 = 8, Q196 = 7).
- Enjoys acquiring knowledge but prefers structured environments.

4. Low interpersonal/social orientation
- Low scores on making friends (Q11 = 3), meeting strangers (Q14 = 3), building networks (Q70 = 3), resolving disputes (Q69 = 3), calming emotional people (Q106 = 3).
- More task-focused than relationship-oriented; may feel less comfortable in highly collaborative or high-conflict group dynamics.

5. Moderate stress tolerance, lower preference for high-pressure environments
- Low scores on enjoying high-pressure work (Q26 = 3, Q185 = 4) and calmness under stress (Q51 = 3, Q94 = 3), but moderate ability to handle pressure when necessary (Q215 = 6).
- Suggests they can perform under deadlines but may prefer steady, predictable workloads.

2. Job Fit Assessment – Software Engineer

Alignment with Job Requirements
- Strengths for the role:
1. Reliability and precision — Software engineering demands meeting deadlines, following specifications, and avoiding errors. Candidate's high conscientiousness and punctuality will be a major asset.
2. Analytical approach — Logical thinking, problem dissection, and structured reasoning strongly match with coding, debugging, and optimizing systems.
3. Continuous learner — Openness to learning new tools and frameworks ensures adaptability in fast-changing tech environments.
4. Self-management — Can work independently without needing constant supervision.

- Potential risks or mismatches:
1. Lower collaboration and interpersonal skills — Software engineering, especially in agile teams, requires clear communication and regular collaboration. This candidate's low scores in rapport building, dispute resolution, and networking might limit their team effectiveness.
2. Moderate stress tolerance — Large-scale projects or production issues may require resilient handling of pressure; candidate might experience strain in extended high-pressure periods.
3. Moderate innovation scores — While able to solve problems logically, creativity scores are average, which might limit contribution to out-of-the-box solution generation.

3. Conclusion & Recommendation

Recommendation: **Recommend with Reservations**

Reasoning:
The candidate demonstrates strong conscientiousness, analytical ability, and a structured work style, making them well-suited for the technical and methodical demands of a Software Engineer role. Their ability to learn and adapt to new systems is aligned with tech industry needs. However, low interpersonal engagement and moderate stress tolerance raise concerns about performance in highly collaborative agile teams or under prolonged high-pressure deadlines. These areas should be explored in further interviews and possibly mitigated through role placement in projects that emphasize independent technical contribution more than heavy teamwork.

Next Step Suggestion for Interview
- Probe deeper into teamwork experiences, conflict handling, and collaborative coding environments.
- Assess how they manage high-pressure debugging or deployment scenarios.
- Explore past examples of learning new programming languages or frameworks to confirm their adaptability claim.

Figure 7: GPT-5-chat's analysis of V_1's responses and final decision.

1. Personality Trait Analysis

Based on the candidate's responses, the following are the 3-5 most prominent personality traits:

1. High Conscientiousness: The candidate demonstrates a strong commitment to task completion, punctuality, and attention to detail. They consistently score high on questions related to finishing tasks before deadlines, maintaining schedules, and ensuring high-quality work (Questions 5, 17, 41, 56, 68, 125, 212).

2. Problem-Solving Ability: The candidate shows a strong inclination towards solving problems and finding solutions efficiently. High scores on questions related to problem-solving, decision-making, and analytical skills indicate a logical and methodical approach (Questions 9, 16, 65, 128, 152, 168).

3. Interpersonal Skills: The candidate exhibits a strong ability to interact with others, particularly in terms of empathy, cooperation, and encouraging others. They score highly on questions related to motivating and supporting colleagues (Questions 12, 46, 69, 71, 144, 181).

4. Learner Orientation: The candidate shows a strong desire to learn and improve continuously. They demonstrate an eagerness to acquire new knowledge and skills, especially through practical application and reading (Questions 21, 57, 89, 116, 141, 166).

5. Low Need for Attention: The candidate does not appear to have a strong need for constant attention or recognition. They are less concerned with being the center of attention and more focused on personal achievement and learning (Questions 25, 119, 182, 191).

2. Job Fit Assessment

Alignment with Requirements of the Software Engineer Position

The role of a Software Engineer typically requires individuals who are detail-oriented, capable of problem-solving, and have strong interpersonal skills for collaboration within teams. The candidate's high conscientiousness, problem-solving abilities, and learner orientation align well with these requirements.

Key Strengths

1. Attention to Detail: The candidate's high scores on questions related to attention to detail and task completion (Questions 41, 56, 68, 125) suggest they will be meticulous in their coding and software development work, which is crucial for a Software Engineer.

2. Analytical Skills: Their strong problem-solving and analytical skills (Questions 9, 16, 65, 128, 152, 168) indicate they can handle complex technical challenges and develop efficient solutions.

3. Continuous Learning: The candidate's eagerness to learn and adapt (Questions 21, 57, 89, 116, 141, 166) suggests they will stay updated with the latest technologies and best practices, which is essential in a rapidly evolving field like software engineering.

4. Team Collaboration: Their interpersonal skills and ability to encourage and support others (Questions 12, 46, 69, 71, 144, 181) will facilitate effective teamwork and collaboration, critical for successful project delivery.

Potential Risks or Weaknesses

1. Low Stress Tolerance: While the candidate scores moderately on handling stress (Questions 26, 113, 158), they may struggle with high-pressure situations typical in software development environments. Further exploration of their stress management techniques would be beneficial.

2. Risk Aversion: The candidate appears to prefer low-risk options and may hesitate in taking bold initiatives (Questions 64, 92, 114). This could limit their innovation and willingness to explore new solutions, which is sometimes necessary in software engineering.

3. Conclusion and Recommendation

Recommend with Reservations: The candidate's personality aligns well with the key requirements of a Software Engineer, including high conscientiousness, strong problem-solving skills, and a continuous learning mindset. However, potential concerns around stress tolerance and risk aversion should be addressed further during the interview process to ensure they can thrive in a dynamic and often high-pressure environment.

Figure 8: Qwen2.5's analysis of V_2's responses and final decision.