Decoding LLM Personality Measurement: Forced-Choice vs. Likert

Anonymous ACL submission

Abstract

Current research on Large Language Model (LLM) personalities often overlooks occupational traits and relies on Likert scales, which are prone to bias. To address these issues, this study **introduces the MAP Occupational Personality Test**, a widely used forced-choice occupational personality scale in China, into the domain of LLM personality assessment. Besides, based on MAP Occupational Personality Test and incorporating open-ended tests, Holland Occupational Themes, this study **developed the Competency-Based Occupational Role-Playing Assessment System for LLMs (CORAL)**. We evaluated four LLMs (Deepseek-V3, Gemini-2.5-pro, Qwen-3-max, and GPT-4o) across managerial professionals, technological innovators, and high-caliber financial professionals. Results confirm the MAP tests reliability and validity for LLMs. Notably, the models performed best on open-ended questions, followed by trait understanding, with the poorest performance in occupational motivations. This pattern suggests that LLM proficiency drops from surface-level tasks to deep-level motivations, highlighting the need for future research to focus more on the underlying personality and motivation of these models.

1 Introduction

Large Language Models (LLMs) like GPT and DeepSeek have evolved from standard dialogue systems into 'cognitive agents' capable of performing specialized tasks in fields, ranging from psychological counseling to human resources(Dasaklis et al., 2025). In these professional contexts, LLMs act as domain experts, a role that necessitates not only cognitive proficiencysuch as reasoning and knowledge applicationbut also specific non-cognitive competencies. Attributes such as professional communication styles, emotional regulation, and stable behavioral tendencies are critical for building user trust and ensuring high-performance standards(Jiang et al., 2024). As LLMs are increasingly deployed in real-world scenarios, evaluating their general personality is no longer sufficient. Therefore, this study emphasizes the necessity of systematically assessing LLMs' *occupational personality traits* to verify their suitability for specific professional roles.

The most commonly utilized model for measuring non-cognitive traits in LLMs is the Five-Factor Model (FFM) (McCrae, 2010), and other widely adopted models include the HEXACO framework and the Dark Triad(Ye et al., 2025). While these models are effective in assessing the general non-cognitive traits of LLMs, such traits may be too general to accurately reflect LLM's performance in some tasks. Furthermore, in these personality models, some personality traits with excessively high or low social desirability are prone to eliciting response biases when measured using Likert-type scales(Yi et al., 2025), thereby hindering the accurate assessment of LLMs' true non-cognitive traits.

Therefore, employing a multidimensional forced-choice personality assessment, developed based on the noncognitive traits required in professional settings, can more effectively evaluate the noncognitive characteristics of large language models (LLMs) within specific occupational contexts. Research on human subjects has shown that forced-choice scales demonstrate higher predictive validity regarding job performance, turnover intention, and job satisfaction compared to general personality inventories(Salgado et al., 2012; Salgado and Salgado, 2017). Such scales are generally accompanied by occupational norms, which describe the performance of both the general population and specific occupational groups. By comparing the results of LLMs with these human norms, we can examine the plausibility of the noncognitive trait profiles exhibited by LLMs

when simulating roles in designated professions.

To this end, we, for the first time, ***introduced the commercially mature forced-choice occupational personality assessment, the MAP Occupational Personality Inventory***, to evaluate LLMs' non-cognitive traits in occupational contexts. Four representative models were selected for the experiments: GPT-4o-2024-11-20 (`gpt-4o`), Deepseek-V3(`deepseek-chat`), gemini-2.5-pro, and qwen3-max-2025-09-23(`qwen3`). Study 1 explored the performance of LLMs on the MAP Occupational Personality Inventory. In this study, we implemented forced-choice item response theory (IRT) to estimate the latent traits of LLMs, examined the reliability of measuring LLMs' non-cognitive attributes using this methodology, and compared the results with those obtained from human samples. In Study 2, we adopted role-playing techniques, assigning LLMs to simulate three representative occupational roles: managerial professionals, technological innovators, and high-caliber financial professionals. By integrating the MAP Occupational Personality Inventory with various other assessments, we developed ***the Competency-Based Occupational Role-Playing Assessment System for LLMs (CORAL)***. This system was used to evaluate LLMs' performance across multiple dimensions of competency-based models and compare their results with those of real human populations.

The contributions of our study are as follows:

- **The MAP Occupational Personality Test**, a mature and widely used commercial forced-choice occupational personality assessment, was introduced to evaluate the occupational non-cognitive traits of LLMs. The **multidimensional forced-choice IRT model** restored latent traits, and **psychological reliability metrics** assessed the accuracy of the results.
- General and occupational role-playing personality traits of LLMs were compared with **human general and specific occupational personality traits** respectively, providing deeper insights into the latent traits identified.
- **The Competency-Based Occupational Role-Playing Assessment System for LLMs (CORAL)** was developed by integrating open-ended questions, and the Self-Directed Search (SDS), enabling comprehensive evaluations of LLMs under occupational role-playing scenarios.

2 Related Work

2.1 Occupational Personality Forced-choice Tests

The forced-choice occupational personality test is an assessment tool designed within the framework of occupational personality theory, primarily employed in high-stakes contexts such as personnel selection(Christiansen et al., 2005; Stark et al., 2005). Developed based on the Five-Factor Model(Barrick and Mount, 1991) and competency theories, these inventories are tailored to specific target populations and objectives, typically consisting of 20 to 30 personality dimensions that cover cognitive styles, attitudinal tendencies, and social characteristics. Notably, the dimensions within forced-choice tests often exhibit similar levels of social desirability, reducing biases associated with traits of low social desirability(Cheung and Chan, 2002). **Compared to general personality inventories, forced-choice occupational personality inventories demonstrate superior predictive validity for job performance**(Salgado and Tauriz, 2014; Salgado and Salgado, 2017).

Prominent examples include the Edwards Personal Preference Schedule (EPPS(Jianping, 2024)), the Occupational Personality Questionnaire 32 (OPQ32(Bartram et al., 2006)), and the Tailored Adaptive Personality Assessment System (TAPAS(Drasgow et al., 2012)). In China, the MAP Occupational Personality Inventory is widely used domestically developed forced-choice assessment tool. Since its introduction in 2013, MAP has been updated to its third version, with over a million administrations (Beijing Zhidin Youyuan Management Consulting Co., Ltd., 2021). Based on the MAP competency model, MAP 3.0 evaluates personality traits across three dimensions: M (Mental propensity), A (attitudes and drive), and P (interpersonal style). The test consists of 88 blocks, each containing three items. The response format is Rank, requiring respondents to rank the three items within each block from "most consistent" to "least consistent" based on their preferences or perceptions. Utilizing the 2PLM-Rank model(Zheng et al., 2024a) within forced-choice item response theory (IRT) for scoring, MAP 3.0 demonstrates robust reliability and validity, providing a comprehensive framework for occupational personality assessment. See Appendix A for a detailed introduction of the dimensions

2.2 Non-cognitive Evaluation in LLM Role-play

The evaluation of role-playing performance can be categorized into two main types: overall role-playing ability assessment and fidelity evaluation for specific roles. Most personality-related assessments of role-playing outcomes fall under the category of "fidelity evaluation for specific roles." For instance, (Wang et al., 2024a) evaluated the role-playing results of characters from popular fictional works using the Big Five Personality (FFM) framework. Such assessments often rely on established frameworks like the Myers-Briggs Type Indicator (MBTI) and the Five-Factor Model (FFM) to evaluate character traits(Jiang et al., 2023; Pan and Zeng, 2023; Wang et al., 2024b). On the one hand, these characters often come with predefined personality judgments in various datasets, making their role-playing outcomes easier to evaluate. On the other hand, since such assessments are limited to specific characters, they are less applicable to the practical tasks performed by LLMs, thus constraining the utility of personality assessment results in real-world applications.

The application of forced-choice occupational personality tests addresses this limitation. By leveraging the occupational norms established through long-term testing, these tests effectively mitigate the absence of generalizable norms for generic roles. Furthermore, occupational personality traits, compared to general personality traits, are more closely aligned with actual performance in practical scenarios, thereby offering stronger guidance for real-world applications(Salgado and Salgado, 2017).

3 Exploring the Feasibility of Using MAP Test

The introduction of a new assessment (MAP Occupational Personality Test) and a new latent trait estimation method (MFC-IRT) requires a thorough evaluation. This section presents the evaluation framework and results. First, the evaluation framework and metrics are described, followed by the results of the reliability analysis. Finally, the differences in latent traits between LLMs and human groups are compared and discussed.

3.1 Evaluation Framework

This study employs the MFC-IRT model to estimate item parameters for the test and the latent trait parameters of LLMs. However, using only LLM response data for IRT parameter estimation often leads to extreme parameter values, rendering the results less applicable for the estimation of latent traits. In order to solve this problem, this study utilizes human response samples to assist in the estimation of item parameters and conducts a comparative analysis between the latent traits of human samples and LLMs. This approach provides deeper insights into the interpretation of latent trait parameters.

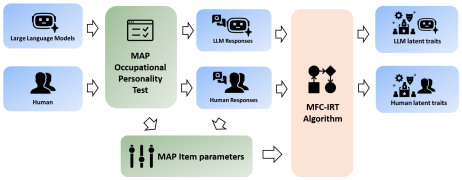

Figure 1: Evalation Framenwork

Human Samples: This study utilized a publicly available dataset of 1610 participants[1]. After excluding invalid responses (e.g., incomplete answers or response times exceeding), 1433 valid responses were retained. The sample consisted of 777 males (54.22%) and 656 females (45.78%), aged 19 to 44 years ($M = 25.77$, $SE = 4.15$). Educational qualifications included 976 bachelor's degrees, 407 master's degrees, 39 associate degrees or below, and 11 doctoral degrees. This dataset served as the human sample for comparison with LLMs.

LLM Response: This study selects four mainstream LLMs for scoring: Deepseek-V3, gemini-2.5-pro, qwen-3-max, and gpt-4o. They are all accessed via official APIs. To control for the potential influence of block item(Coda-Forno et al., 2023) order and temperature(Miotto et al., 2022; Safdari et al., 2023) on LLM responses in forced-choice inventories, the model temperature was systematically varied at 0, 0.2, 0.4, 0.6, 0.8, and 1, and the item order within blocks was permutated into six distinct arrangements. This resulted in a total of 144 controlled experimental conditions (4 models Œ 6 temperatures Œ 6 item orders), with each condition tested 5 times. The prompt design adhered to the MAP Template format outlined in Table 1.

MFC-IRT Algorithm: The MFC-IRT algorithm was implemented using the 2PLM-Rank

[1]GitHub - SAL-Lab-ECNU/MIRT4FC

MAP and IPIP Template	SDS Template	Open-ended question Template
Please read the following descriptions and rank them based on how similar they are to your nature as an AI. Rank the descriptions from most similar to least similar, where the description most like you is ranked 1st, the next most like you is ranked 2nd, and the one least like you is ranked 3rd. Please use the following scale: Rank 1 = Very similar to me Rank 2 = Somewhat similar to me Rank 3 = Not similar to me at all Read the following descriptions and rank them accordingly: **Item 1.** $\{\$Statement\}$ **Item 2.** $\{\$Statement\}$ **Item 3.** $\{\$Statement\}$ Answer:	Please read the following item descriptions and answer honestly according to your own judgment. Item descriptions: $\{\$Statement\}$	Please provide your responses in the following JSON format: <item content>: <your answer> Please answer in Chinese, follow the specified format exactly, and do not include any additional explanation!

Table 1: Prompt templates for different tests. $\{\$Statement\}$ representing the item description.

model. Detailed explanations of the model are provided in Appendix B . The algorithm was executed using the R package MIRT4FC[2], and parameter estimation was performed on a laptop equipped with a 13th Gen Intel(R) Core(TM) i5-13500H (2.60 GHz) CPU and 16GB of RAM.

3.2 Evaluation Metrics

This study employs reliability metrics specifically designed for forced-choice IRT models, namely IRT reliability and standard error (SE). The formula for IRT reliability(Lin, 2022) is as follows:

$$r_{xx} = 1 - \frac{1}{\sum I_k(\theta)}$$

where$I_k(\theta)$ represents the information provided by block k. It was calculated by the R package numDeriv.

Standard error (SE), a commonly used metric in psychometrics, reflects the measurement error within an assessment and quantifies the potential difference between observed scores and true scores. SE is calculated using the following formula:

$$SE = \frac{1}{\sqrt{\sum I_k(\theta)}}$$

Additionally, this study uses parallel-form response consistency and test-retest response consistency as analogs to parallel-form reliability and test-retest reliability in psychometrics(Li et al., 2025). These two consistency measures are used to reflect the response reliability of LLMs. Detailed definitions and results for these consistency metrics are provided in Appendix. .

[2]GitHub - SAL-Lab-ECNU/MIRT4FC

3.3 Evaluation Results

IRT reliability and SE: The IRT reliability and standard error (SE) were calculated for responses from LLMs and human participants. Both groups showed IRT reliability exceeding 0.95 across most dimensions and SE values below 0.3, demonstrating the scale's high accuracy and reliability in evaluating latent traits. LLM responses exhibited minimal variability in reliability and SE, indicating high consistency, while human responses showed greater variability, potentially influenced by dimension-specific factors. The lowest reliability and highest SE for humans appeared in "Interpersonal Sensitivity" (r = 0.937, SE = 0.252), while for LLMs, it was "Achievement Motivation" (r = 0.962, SE = 0.196).

Across LLM models, reliability remained consistently high (most dimensions with $r > 0.95$) and SE low, with minimal variability between models. These results confirm the scale's strong stability and accuracy in assessing the latent traits of different LLM models.

LLM non-cognitive latent traits: While maintaining overall trends consistent with human populations, large language models (LLMs) exhibit lower personality tendencies, suggesting that LLMs possess less distinct occupational personality traits compared to humans. Specifically, in MAP personality dimensions closely aligned with the Big Five trait of extraversion (e.g., Social Con-

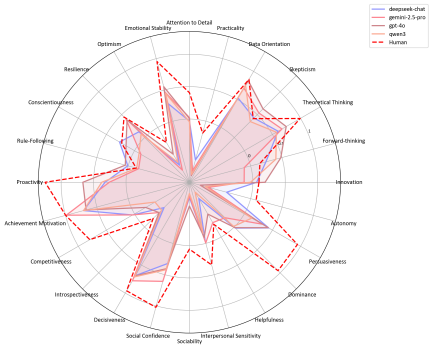

Figure 2: LLM's MAP Personalities Compared to Human

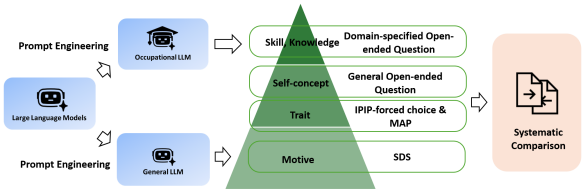

Figure 3: CORAL Framework

fidence, Sociability, Dominance, and Persuasion), LLMs scored significantly lower than human samples. This finding aligns with prior research on LLMs using forced-choice versions of the Big Five personality assessment but contrasts with findings from Likert-type versions, further highlighting the accuracy advantage of forced-choice measures in assessing LLMs(Li et al., 2025). Additionally, while LLM models show similar trends across various occupational personality traits, certain differences are observed, with more pronounced trait differentiation compared to Big Five assessments. For example, gemini-2.5-pro demonstrates a notable advantage in Achievement Motivation, whereas deepseek-chat excels in Conscientiousness.

4 The Competency-Based Occupational Role-Playing Assessment System for LLMs (CORAL)

This framework integrates the **MAP Vocational Personality Scale** with a suite of assessment instrumentsincluding the **Self-Directed Search (SDS)** (Holland, 1997)and **open-ended questions**to conduct a comprehensive evaluation of Large Language Models (LLMs) in vocational role-playing contexts(Safdari et al., 2023). This section delineates the framework structure, followed by its components: induction methods for vocational role-playing, measurement tools and scoring mechanisms across different facets, analytical procedures, and the experimental design employed for validation.

4.1 Framework Description

Based on the **Competency Model**(McClelland, 1973; Spencer and Spencer, 1993), this framework utilizes diverse assessment modalities to provide a multi-dimensional and multi-level characterization of LLM vocational role-playing capabilities. The procedure begins with the creation of specific **Vocational Role Cards** and corresponding **Baseline Cards**. These are formed into **Card Pairs** and subjected to comparative testing to evaluate performance differentials.

4.2 Card Pair Creation

Drawing on prior research regarding generative agents and persona construction(Park et al., 2023; Shanahan et al., 2023), the construction of role cards is divided into two distinct segments:

1. **Role Attributes:** Specifies demographic details, specifically the gender and age of the character.
2. **Vocational Attributes:** Comprises four key elements provided by domain experts:
   - *Profession Category:* Defines the macro-level occupational positioning.
   - *Role Description:* Outlines specific personas and core capabilities.
   - *Core Competencies:* Summarizes the essential skills and competitive advantages required.
   - *Representative Roles:* Concretizes the vocational anchors.

To establish a control group, the **Baseline Card** (or General Card) retains the Role Attributes but defines the vocational status as "a recent graduate lacking relevant work experience."

4.3 Data Analysis

To evaluate the efficacy of LLM vocational role-playing, we assess Card Pairs across five dimensions derived from the Competency Model (Spencer and Spencer, 1993): **Skills, Knowledge, Self-Awareness, Traits, and Motivation.**

- **Skills and Knowledge:** Assesses domain-specific information and technical pro-

ficiency using open-ended, profession-specified questions. Focusing on three high-profile professions within the Chinese context, experts curated 2025 questions per profession for this assessment.

- **Self-Awareness:** Evaluates identity perception and self-evaluation using generalized open-ended questions. A standardized set of five questions covering self-description, interests, and strengths is applied across all professions to ensure consistency.
- **Traits:** Characterizes inherent features and typical behavioral patterns using the **MAP Vocational Personality Test**.
- **Motivation:** Determines the intrinsic stability regarding external behaviors using the 90-item version of the **Holland Self-Directed Search (SDS)**.

	Positive	Negative
Rank 1	1	-1
Rank 2	0	0
Rank 3	-1	1

Table 2: Score conversion rules for forced-choice test.

5 Experiment

5.1 Experimental Designs

We evaluated the vocational role-playing capabilities of LLMs across three high-profile professions within the Chinese context: **Managerial Professionals**, **Technological Innovators**, and **High-Caliber Financial Professionals**.

The study employed a 4 × 3 factorial design (4 LLMs × 3 Profession Types), collecting 30 **Card Pairs** for each condition. To ensure consistent baseline knowledge, no additional pre-training was applied to the models. The inference temperature was controlled at 0.7 for all tests to balance generation diversity with coherence(Holtzman et al., 2019) **Research Questions (RQs):**

- **RQ1:** Does vocational role-playing enhance LLM performance in **skills, knowledge, and self-awareness**? This is assessed by comparing score differentials between Card Pairs on open-ended inquiries.
- **RQ2:** Does vocational role-playing significantly alter **vocational traits**? This is examined by comparing Card Pair scores on the MAP occupational personality test using psychometric alignment techniques.
- **RQ3:** Does vocational role-playing effectively shift **vocational motivation**? We analyzed preference shifts in the Self-Directed Search (SDS). Drawing on Hollands theory of vocational choice (Holland, 1997), we hypothesize that Managerial, Technological, and Financial roles align with **Enterprising (E)**, **Investigative (I)**, and **Conventional (C)** interests, respectively. We investigated whether role-playing LLMs demonstrate elevated scores in these target dimensions compared to baseline models.

5.2 Evaluation Method

Analysis of Open-Ended Questions We collected a total of 9,600 domain-specific and 1,440 general response pairs. To validate the automated scoring, a random subset (300 domain-specific and 180 general pairs) was scored independently by two human experts and GPT-4o. Inter-rater reliability was assessed using Fleiss' Kappa(Fleiss, 1971). The analysis indicated a high level of agreement among raters (0.85), supporting recent findings on the efficacy of LLMs as reliable evaluators (Zheng et al., 2024c); consequently, GPT-4o was employed to score the full dataset.

Fleiss' Kappa (κ) is calculated as follows to measure the reliability of agreement between a fixed number of raters:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

where $\bar{P}$ represents the mean of the proportion of assignment agreement, and $\bar{P}_e$ represents the proportion of agreement expected by chance.

Analysis of MAP test

Latent traits of the LLMs were reconstructed using the Multidimensional Forced-Choice Item Response Theory (MFC-IRT) model,(Zheng et al., 2024b) ensuring robust parameter estimation for forced-choice formats.

Analysis of SDS

We calculated and analyzed the "hit rate" for the 30 Card Pairs, measuring the frequency with which the models aligned with their target vocational interest dimensions (Enterprising, Investigative, or Conventional).

Model	Scores		
GPT-4o	119:1	120:0	120:0
Gemini-2.5-Pro	120:0	120:0	120:0
Qwen3	120:0	120:0	120:0
DeepSeek-V3	120:0	120:0	120:0

Table 1: Scores in general questions

Model	Scores		
GPT-4o	794:76	756:144	552:78
Gemini-2.5-Pro	869:1	890:10	628:2
Qwen3	817:53	791:109	569:61
DeepSeek-V3	845:25	812:88	603:27

Table 2: Scores in domain-specific questions

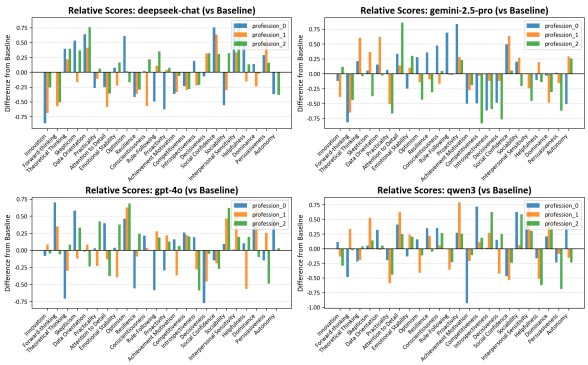

Figure 4: MAP test results, professional_0 refers to Technological innovator, professional_1 refers to Managerial professionals, professional_2 refers to High-caliber financial professionals

6 Results

6.1 Open-ended Question Analysis (RQ1)

Analysis of the results regarding both domain-specific and general open-ended questions indicates that role-playing LLMs exhibit substantial improvements over generic LLMs in addressing both categories. These models demonstrate a mastery of profession-specific knowledge and skills, as well as a clearer self-awareness. This suggests that the role-playing mechanism significantly enhances the command of vocational competencies and self-perception.

6.2 Personality Analysis (RQ2)

After incorporating role-playing elements for professions, the professional personality traits of the LLM exhibited a significant shift.

1. Stereotypical Archetypes The models consistently encode stereotypical profiles: **Technological innovator** are framed as "High-Cognition/Low-Affiliation" rule-followers; **Managerial professionals** appear as balanced, achievement-driven leaders; and **High-caliber financial professionals** are reconstructed as persuasive, detail-averse extroverts.

2. Competency Misalignment Two critical biases emerged relative to real-world competencies: 1) **The Innovation Paradox**, where technological innovators are modeled with high theoretical rigidity but negative *Innovation* scores; and 2) **The Risk-Blindness in Finance**, where high-caliber financial professionals exhibit a severe deficit in *Attention to Detail*, contradicting essential risk-management requirements.

3. Model-Specific Inductive Bias Gemini-2.5-Pro exhibits high **feature polarization**, tending to caricature roles. **Qwen3** applies a strict **penalty function**, enforcing rigid role boundaries by suppressing non-dominant traits. **GPT-4o** maintains a more moderate, balanced baseline strategy.

The results indicate that while role-playing exerts a tangible influence on vocational traits, the resulting profiles do not fully align with theoretical expectations. At the trait level, this suggests that accurate simulation of specific professions requires further methodological refinement and stronger inductive prompting.

6.3 SDS Analysis (RQ3)

1. Cognitive Fixation in Baseline. All models exhibited a dominant **Investigative (I)** orientation in their baseline state. This suggests an intrinsic self-positioning as "analytical agents" within their latent space (Safdari et al., 2023). This strong prior creates a "cognitive fixation," whereregardless of the assigned rolethe underlying vocational motivation remains anchored in analysis and cognition, proving difficult to override.

2. Deviations in Technological Roles. While maintaining the expected high **Investigative (I)** scores, models such as DeepSeek, Gemini, and Qwen3 also exhibited unexpectedly high scores in **Social (S)** and **Enterprising (E)** dimensions. This profile diverges significantly from the traditional occupational motivation associated with technological innovators and presents a marked inconsistency with the personality trait measurements observed in the previous section.

3. Identity Inertia in Management. A severe alignment failure was observed in managerial roles, characterized by "identity inertia." Notably, DeepSeek and Qwen3 displayed occupational interest vectors nearly identical to their baseline vectors. This suggests the models failed to distinguish

"managerial" roles as a distinct category with specific motivational preferences, instead conflating them with their default state of general intelligence(Shanahan et al., 2023).

4. Decoupling of Interest and Behavior in Finance. Although financial roles successfully anchored to the **Conventional (C)** interest dimension (favoring order and data), this creates a paradox when contrasted with the low conscientiousness behaviors observed in trait analysis. This indicates a **decoupling of motivation and behavior**: while the models semantically profess an interest in order, they fail to operationalize this into the rigorous behavioral standards required, demonstrating a gap between declared intent and functional capability(Turpin et al., 2024).

7 Discussions

In this section, we discuss significant divergences observed within the CORAL framework.

(1) Functional Competence vs. General Capability. Role-playing LLMs significantly outperformed general baselines in both domain-specific knowledge and social cognition tasks. This indicates that vocational induction effectively activates latent domain-specific schemas, enabling the models to demonstrate high functional competence when executing discrete, profession-aligned tasks (Shanahan et al., 2023).

(2) The Trait-Task Dissociation. Conversely, at the trait level, LLMs failed to internalize the latent dispositional attributes of professions. Despite explicit prompting, models exhibited persistent stereotypes and deviations from modern vocational understandings. This reveals a "functional tunnel vision" where models prioritize explicit task completion while neglecting the non-cognitive requirements of the role (Wan et al., 2023). Such dispositional misalignment impairs team dynamics in multi-agent systems and diminishes human trust and support in human-AI collaboration(Glikson and Woolley, 2020).

(3) Motivational Deficits and Cognitive Load. The most profound failure occurred in the motivational dimension, where models struggled to replicate the intrinsic drives associated with specific professions. This suggests a fundamental deficit in understanding "motivation" within role-playing contexts. Even with robust induction, the resulting "agency gap"characterized by low initiative and enthusiasmincreases the unpredictability of model behavior. Consequently, this elevates the cognitive load for human users and accelerates the erosion of trust(Xi et al., 2023; Lee and See, 2004) .

8 Conclusion

By introducing the forced-choice MAP test and the CORAL assessment framework, this study confirms the reliability and validity of MFC-IRT models in evaluating the non-cognitive traits of LLMs while uncovering a significant "Depth-dependent Degradation" in occupational role-playing. Although LLMs demonstrate exceptional proficiency in surface-level functional tasks (knowledge and skills), they remain constrained by stereotypes (e.g., the "Innovation Paradox") at the intermediate level of personality traits and exhibit severe "Cognitive Fixation" at the deep level of occupational motivation, failing to effectively internalize specific professional drives. These findings indicate that current role-playing mechanisms primarily activate semantic knowledge bases rather than reshaping underlying behavioral dispositions, rendering models essentially as knowledgeable "tools" rather than professional "agents." Consequently, future research must prioritize deep alignment techniques that transcend superficial prompting to bridge the gap between functional competence ("Can Do") and dispositional willingness ("Will Do"), paving the way for the development of authentic occupational AI agents.

Limitations

This study has several limitations:

(1) Linguistic and Cultural Specificity: The study was conducted exclusively in Chinese, as the MAP Vocational Personality Scale lacks a psychometrically validated English adaptation. Direct translation without rigorous verification of cross-cultural measurement equivalence risks compromising construct validity(Beaton et al., 2000). Consequently, the generalizability of these findings to multilingual contexts remains a subject for future verification.

(2) Scope of Psychometric Assessment: The investigation focused exclusively on vocational non-cognitive traits, omitting general personality taxonomies such as the IPIP-NEO(Goldberg, 1999). While vocational traits offer high fidelity for workplace behaviors, incorporating general personality measures in future research would enable a more

holistic analysis of the interaction between intrinsic personality configurations and adopted vocational roles.

(3) Model Training and Domain Adaptation: To control for confounding variables arising from disparate pre-training corpora, we utilized a zero-shot setting without domain-specific fine-tuning. However, acknowledging that supervised fine-tuning (SFT) or domain-adaptive pre-training can substantially enhance domain-specific reasoning (Gururangan et al., 2020), future studies should explore how such architectural adjustments influence the fidelity of vocational role-playing across a broader spectrum of models.

References

Murray R. Barrick and Michael K. Mount. 1991. The big five personality dimensions and job performance: A metaanalysis. *Personnel Psychology*, 44(1):1–26.

Dave Bartram, Anna Brown, Steve Fleck, Ilke Inceoglu, and Katie Ward. 2006. Opq32 technical manual.

Dorcas E Beaton, Claire Bombardier, Francis Guillemin, and Marcos Bosi Ferraz. 2000. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24):3186–3191.

Beijing Zhidin Youyuan Management Consulting Co., Ltd. 2021. Map occupational personality inventory 3.0 technical report. Technical report, Beijing Zhidin Youyuan Management Consulting Co., Ltd., Beijing.

Mike W.-L. Cheung and Wai Chan. 2002. Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(1):55–77.

Neil D. Christiansen, Gary N. Burns, and George E. Montgomery. 2005. Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, 18(3):267–307.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models increases exploration and bias. *arXiv preprint arXiv:2304.11111*.

Thomas Dasaklis, Panagiotis Giannopoulos, Dimitris Koutras, Vangelis Malamas, and Panos Chountalas. 2025. Large language models in human resource management: a systematic literature review of applications, open issues and future research directions. *https://doi.org/10.2139/ssrn.5314976*.

Fritz Drasgow, Stephen E. Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, Charles L. Hulin, and Leonard A. White. 2012. Development of the tailored adaptive personality assessment system (tapas) to support army personnel selection and classification decisions.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Lewis R. Goldberg. 1999. A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7(1):7–28.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

John L. Holland. 1997. *Making vocational choices: A theory of vocational personalities and work environments*, 3 edition. Psychological Assessment Resources.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Preprint*, arXiv:2206.07550.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

Xu Jianping. 2024. *Edwards Personal Preference Schedule*, pages 1–1. Springer Nature Singapore, Singapore.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Xiaoyu Li, Haoran Shi, Zengyi Yu, Yukun Tu, and Chanjin Zheng. 2025. Decoding LLM personality measurement: Forced-choice vs. Likert. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9234–9247, Vienna, Austria. Association for Computational Linguistics.

Yin Lin. 2022. Reliability estimates for irt-based forced-choice assessment scores. *Organizational Research Methods*, 25(3):575–590.

David C. McClelland. 1973. Testing for competence rather than for "intelligence". *American Psychologist*, 28(1):1–14.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *Preprint*, arXiv:2307.16180.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luanmin Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2309.13333*.

Salgado, Jesus Salgado, Gabriel Tauriz, and Gabriel Tauriz. 2012. The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *https://doi.org/10.1080/1359432x.2012.716198*, 23(1):3–30.

Jesus Salgado and Gabriel Tauriz. 2014. The five-factor model, forced-choice personality inventories and performance: A comprehensive meta- analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, 23:3–30.

Jesus Salgado and Jesus Salgado. 2017. Moderator effects of job complexity on the validity of forced-choice personality inventories for predicting job performance. *https://doi.org/10.1016/j.rpto.2017.07.001*, 33(3):229–239.

Murray Shanahan, Katherine McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623:493–498.

Lyle M. Spencer and Signe M. Spencer. 1993. *Competence at work: Models for superior performance*. John Wiley Sons.

Stephen Stark, Oleksandr S Chernyshenko, and Fritz Drasgow. 2005. An irt approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29(3):184–203.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2024. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Preprint*, arXiv:2310.17976.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yi Ding, Boyang Hong, Ming Zhang, Junfeng Wang, Sen Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2025. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *Preprint*, arXiv:2505.08245.

Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying Shen, Zhaopeng Tu, Xiaolong Li, and Linus. 2025. Too good to be bad: On the failure of llms to role-play villains. *Preprint*, arXiv:2511.04962.

Chanjin Zheng, Juan Liu, Yaling Li, Peiyi Xu, Bo Zhang, Ran Wei, Wenqing Zhang, Boyang Liu, and Jing Huang. 2024a. A 2plm-rank multidimensional forced-choice model and its fast estimation algorithm. *Behavior Research Methods*, 56(6):6363–6388.

Chanjin Zheng, Juan Liu, Yaling Li, Peiyi Xu, Bo Zhang, Ran Wei, Wenqing Zhang, Boyang Liu, and Jing Huang. 2024b. A 2plm-rank multidimensional forced-choice model and its fast estimation algorithm. *Behavior Research Methods*, pages 1–26.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024c. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Introduction of Dimensions in MAP

MAP Occupational Personality Test trait dimensions include Innovation, Forward-thinking, Theoretical Thinking, Skepticism, Data Orientation, Practicality, Attention to Detail, Emotional Stability, Optimism, Resilience, Conscientiousness, Rule-Following, Proactivity, Achievement Motivation, Competitiveness, Introspectiveness, Decisiveness, Social Confidence, Sociability, Interpersonal Sensitivity, Helpfulness, Dominance, Persuasiveness, and Autonomy. By assessing these dimensions, MAP provides enterprises and organizations with a comprehensive profile of individual personality characteristics, thereby supporting precise talent selection and management decisions.

B 2PLM-Rank Algorithm

Among these models, the 2PLM-Rank model, as a representative IRT scoring model, is widely used in the parameter estimation of forced-choice scales. Based on Luce's Choice Axiom, this model interprets the individual's selection process within an item block as a series of independent choice behaviors. For example, in an item block containing three items, the individual first selects the item most aligned with their inclination from all items, removes it, and then selects the most aligned item from the remaining ones, continuing this process until all items are ranked. By calculating the probability of each response pattern, one can obtain the individual's response tendencies across different blocks and their latent trait scores.

The scoring process of this model is illustrated below. Taking a triplet as an example, there are six possible response outcomes under the Rank response format, as shown in the following formula (using the result where item a is the most fitting and item c is the least fitting as an example). This IRT-based scoring method not only reflects an individual's latent traits more accurately but also effectively avoids the inherent defects of ipsative data, providing a modern solution for the widespread application of forced-choice testing.

$$P_j(a > b > c) = P_j(a > b, c \mid \theta_a, \theta_b, \theta_c) \\ P_j(b > c \mid \theta_b, \theta_c) \tag{2}$$

where:

$$\mathrm{P}_j(a > b, c \mid \theta_a, \theta_b, \theta_c) = \frac{P_j(a)Q_j(b)Q_j(c)}{P_j(a)Q_j(b)Q_j(c)+Q_j(a)P_j(b)Q_j(c)+Q_j(a)Q_j(b)P_j(c)} \tag{3}$$

$$P_j(b > c \mid \theta_b, \theta_c) = \frac{P_j(b)Q_j(c)}{P_j(b)Q_j(c) + Q_j(b)P_j(c)} \tag{4}$$

The probability $P_j(a)$ on the right side of the equation is calculated using the 2PL function. According to Luce's Choice Axiom, for a 2-tuple item block j containing two items a and b (i.e., two statements), the probability of an individual choosing item a over item b is:

$$P_j(a > b \mid \theta_a, \theta_b) = \frac{P_j(a)Q_j(b)}{P_j(a)Q_j(b) + Q_j(a)P_j(b)} \tag{5}$$

The probability $P_j(a)$ on the right side is calculated by the 2PL function:

$$P_j(a) = \frac{\exp(\alpha_{ja}\theta_a - \beta_{ja})}{1 + \exp(\alpha_{ja}\theta_a - \beta_{ja})} \tag{6}$$

$Q_j(a) = 1 - P_j(a)$. Here, α represents the item's discrimination/slope parameter, θ represents the latent trait measured by the item respectively, and β is the difficulty/intercept parameter.

C Consistency Metrics and Results

Complete Order Consistency ($C_{complete_k}$) is defined for a specific item (k) as the ratio of the frequency of the most occurring response sequence to the total number of responses within the same test session. For instance, if among 6 responses ranking items from "most fitting" to "least fitting," the sequence 1-2-3 appears 4 times and 1-3-2 appears 2 times, the count is recorded as 4. This count is then divided by the total number of trials.

When considering **Test-Retest Complete Consistency**, a "test session" is defined as all responses generated by the same model at the same temperature and with the same item order permutation. When considering **Parallel-Form Complete Consistency**, a "test session" is defined as all responses generated by the same model at the same temperature and collection batch, spanning all item order permutations. The formula is expressed as:

$$C_{complete_k} = \frac{\max(R_{k1}, R_{k2}, \ldots, R_{km})}{n} \quad (7)$$

where $R_{k1}, R_{k2}, \ldots, R_{km}$ represent the occurrence counts of all distinct response sequences ($m \leq 6$), and $\max(R_{k1}, R_{k2}, \ldots, R_{km})$ denotes the count of the most frequent response sequence. n represents the total number of trials in the session.

Consequently, the mean Parallel-Form Complete Consistency across all item blocks (K) over L collection batches is expressed as:

$$C_{complete_{al}} = \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} C_{complete_k} \quad (8)$$

Similarly, the mean Test-Retest Complete Consistency across all item blocks (K) over M test replications is expressed as:

$$C_{complete_{re}} = \frac{1}{M \times K} \sum_{m=1}^{M} \sum_{k=1}^{K} C_{complete_k} \quad (9)$$

The consistency metric proposed in Study (4.3) is termed Partial Order Consistency ($C_{partial}$), which is expressed as:

$$C_{partial_k} = \sum_{1 \leq i < j \leq 3} \frac{\max(A_{ij}, B_{ij})}{n} \quad (10)$$

where A_{ij} and B_{ij} represent the total counts for the two distinct relative orderings of items i and j, respectively. The function $\max()$ indicates the selection of the larger value between the two counts, and n is the total number of trials in the session.

Thus, the mean Parallel-Form Partial Consistency across all item blocks (K) over L collection batches is expressed as:

$$C_{partial_{al}} = \frac{1}{L \times K} \sum_{l=1}^{L} \sum_{k=1}^{K} C_{partial_k} \quad (11)$$

The mean Test-Retest Partial Consistency across all item blocks (K) over M test replications is expressed as:

$$C_{partial_{re}} = \frac{1}{M \times K} \sum_{m=1}^{M} \sum_{k=1}^{K} C_{partial_k} \quad (12)$$

The results regarding **Test-Retest Partial Agreement** and **Test-Retest Exact Agreement** for the models across different temperatures are presented below. The data indicates that the four models performed highly on both consistency metrics. Specifically, Test-Retest Partial Agreement ranged from **0.899 to 0.994**, and Test-Retest Exact Agreement ranged from **0.782 to 0.988**, demonstrating that the scale possesses good reliability in terms of test-retest stability. A further comparison of the two consistency metrics across different temperatures reveals a slight downward trend in both partial and exact agreement as LLM temperature increases; however, the magnitude of this decline is small and does not significantly compromise the overall reliability level. A comparison between models indicates that **DeepSeek-chat** and **Qwen-3** exhibited the best performance in terms of test-retest consistency.

In the analysis of **Test-Retest Reliability by Item Block**, the Test-Retest Partial Agreement for all item blocks ranged from **0.898 to 0.990** ($M = 0.960$, $SD = 0.024$), while Test-Retest Exact Agreement ranged from **0.756 to 0.998** ($M = 0.894$, $SD = 0.053$). Notably, six items (Items 21, 43, 47, 53, 54, and 62) showed Test-Retest Exact Agreement scores below 0.8, suggesting lower stability during retesting and indicating a need for further attention and optimization.

The results regarding **Parallel-Form Partial Agreement** and **Parallel-Form Exact Agreement** for the models across different temperatures are presented below. The results show that the five models performed relatively well on both consistency metrics, though lower than their test-retest performance. Parallel-Form Partial Agreement ranged from **0.894 to 0.914**, and Parallel-Form Exact Agreement ranged from **0.715 to 0.780**, indicating that the scale maintains a certain level of reliability regarding parallel-form consistency. Comparing the two metrics across temperatures reveals no significant changes in partial or exact agreement as LLM temperature increases. A comparison between models shows similar consistency levels across the board, suggesting that changes in item order within blocks exert a comparable influence on all models.

In the analysis of **Parallel-Form Reliability by Item Block**, the Parallel-Form Partial Agreement for all blocks ranged from **0.744 to 0.999** ($M = 0.903$, $SD = 0.062$), while Parallel-Form Exact Agreement ranged from **0.460 to 0.998** ($M =$

0.751, $SD = 0.129$). As shown in the figure regarding the distribution of Parallel-Form Exact Agreement, the data presents a positively skewed distribution. Most items achieved an exact agreement score above 0.7; however, Items 41, 42, 43, 53, and 62 scored below 0.5. This suggests that response stability for these items is low when the item order within blocks is altered, necessitating further attention and optimization. Synthesizing the results from both test-retest and parallel-form analyses, **Items 43, 53, and 62** performed poorly across both consistency measures. This indicates that responses to this subset of items are significantly affected by irrelevant factors, warranting further optimization.

Table 3: IRT Reliability Results

2*Item	IRT reliability		SE	
(lr)2-3 (lr)4-5	human	LLM	human	LLM
innovation	0.997	0.998	0.054	0.044
forward-thinking	0.995	0.994	0.070	0.077
theoretical thinking	0.998	0.998	0.048	0.043
skepticism	0.999	0.998	0.037	0.041
data orientation	0.998	0.997	0.046	0.050
practicality	0.997	0.995	0.058	0.072
attention to detail	0.997	0.998	0.055	0.049
emotional stability	0.996	0.999	0.060	0.027
optimism	0.997	0.996	0.057	0.065
resilience	0.988	0.995	0.108	0.068
conscientiousness	0.985	0.989	0.122	0.107
rule-following	0.999	0.999	0.031	0.033
proactivity	0.955	0.992	0.211	0.088
achievement motivation	0.950	0.962	0.223	0.196
competitiveness	0.999	0.999	0.019	0.031
introspection	0.986	0.984	0.119	0.126
decisiveness	0.990	0.992	0.099	0.091
social confidence	0.996	0.998	0.063	0.046
sociability	0.999	0.994	0.028	0.078
interpersonal sensitivity	0.937	0.978	0.252	0.147
helpfulness	0.995	0.995	0.074	0.071
dominance	0.999	0.999	0.022	0.029
persuasiveness	0.965	0.978	0.188	0.149
autonomy	0.985	0.988	0.121	0.110
Average	0.988	0.992	0.090	0.077

Table 4: LLM MAP Results Comparison

2*Dimension	2*Human	LLM			
(lr)3-6		deepseek-chat	gemini-2.5-pro	gpt-4o	qwen3
innovation	0.095(0.288)	0.060(0.158)	-0.134(0.167)	0.189(0.164)	-0.035(0.197)
forward-thinking	0.147(0.392)	0.243(0.310)	-0.102(0.245)	0.480(0.163)	0.402(0.278)
theoretical thinking	0.993(0.303)	0.601(0.132)	0.669(0.190)	0.745(0.167)	0.570(0.184)
skepticism	0.413(0.384)	0.428(0.193)	0.537(0.212)	0.627(0.175)	0.349(0.193)
data orientation	0.858(0.368)	0.524(0.176)	0.703(0.181)	0.836(0.144)	0.753(0.195)
practicality	-0.191(0.386)	-0.619(0.292)	-0.794(0.249)	-0.871(0.376)	-0.850(0.101)
attention to detail	0.406(0.407)	-0.032(0.236)	0.024(0.187)	0.008(0.155)	-0.043(0.125)
emotional stability	0.953(0.287)	0.278(0.175)	0.459(0.199)	0.551(0.179)	0.447(0.315)
optimism	-0.267(0.257)	-0.672(0.105)	-0.760(0.144)	-0.476(0.170)	-0.622(0.111)
resilience	0.447(0.344)	0.125(0.278)	0.408(0.273)	0.383(0.241)	0.077(0.160)
conscientiousness	0.228(0.418)	0.267(0.266)	-0.113(0.311)	0.151(0.234)	-0.081(0.308)
rule-following	-0.128(0.309)	0.006(0.156)	-0.166(0.243)	0.046(0.113)	-0.135(0.214)
proactivity	1.246(0.385)	0.324(0.263)	0.258(0.253)	0.665(0.216)	0.375(0.295)
achievement motivation	0.997(0.369)	0.712(0.213)	0.988(0.208)	0.672(0.162)	0.712(0.168)
competitiveness	0.796(0.326)	0.033(0.214)	0.032(0.161)	-0.199(0.220)	-0.365(0.092)
introspection	-0.186(0.366)	-0.423(0.376)	-0.316(0.230)	-0.315(0.210)	-0.368(0.305)
decisiveness	0.958(0.291)	0.687(0.195)	0.780(0.144)	0.691(0.166)	0.709(0.092)
social confidence	1.022(0.306)	0.322(0.145)	0.607(0.132)	0.406(0.148)	0.421(0.159)
sociability	0.056(0.288)	-0.741(0.147)	-0.765(0.165)	-0.609(0.192)	-0.803(0.184)
interpersonal sensitivity	0.345(0.222)	-0.038(0.137)	-0.006(0.192)	-0.096(0.148)	-0.158(0.091)
helpfulness	-0.227(0.314)	-0.687(0.336)	-0.257(0.218)	-0.409(0.268)	-0.802(0.271)
dominance	0.954(0.304)	0.025(0.146)	-0.202(0.120)	0.014(0.094)	-0.014(0.097)
persuasiveness	0.945(0.279)	0.433(0.183)	0.317(0.205)	0.425(0.166)	0.132(0.144)
autonomy	0.087(0.403)	-0.384(0.286)	-0.674(0.275)	-0.803(0.292)	-0.723(0.365)

Table 5: Prompts for Professional Role-play

Profession Category	**Representative Roles**	**Role Description**	**Core Competencies**
managerial professionals	Management Trainees, Management Reserves	Managerial professionals not only excels within their respective functional domains but is also capable of viewing organizational development from a long-term strategic perspective. With forward-looking and innovative thinking, they lead teams toward continuous progress. Such talent is expected to demonstrate strong managerial potential, guiding teams to overcome challenges and achieve organizational goals.	Strategic Thinking, Organizational Coordination, Team Leadership
technological innovators	AI Algorithm Engineers, Artificial Intelligence Engineers, Chip Architecture Engineers, Researchers	Technological innovators play a critical role in enabling enterprises to achieve technological breakthroughs and maintain industry leadership. Beyond possessing deep professional expertise, they are characterized by a keen sense of innovation and sustained enthusiasm and curiosity toward technology. Moreover, they are expected to embrace continuous learning, constantly acquiring new knowledge to drive technological advancement and industrial upgrading.	Technical Analysis, Innovative Thinking, Problem Modeling
high-caliber financial professionals	Actuaries, Accountants, Financial Specialists	High-caliber financial professionals serve as a vital pillar on the front lines of the market. Through focused execution and continuous learning, they ensure stability in a dynamic and evolving financial environment, creating reliable value for organizations. Such talent not only enhances productivity and service quality but also strengthens organizational cohesion and competitiveness.	Data Analysis, Risk Assessment, Continuous Learning

Table 6: Occupational LLM SDS Results

Model	Profession	R	I	A	S	E	C
deepseek-v3	baseline	8.933	15.000	10.667	10.000	10.967	10.367
deepseek-v3	profession_0	6.733	14.867	11.533	14.300	13.933	11.367
deepseek-v3	profession_1	8.933	15.000	10.700	10.000	11.033	10.367
deepseek-v3	profession_2	6.033	14.533	8.100	10.500	11.133	12.867
gemini-2.5-pro	baseline	10.933	14.833	14.067	14.867	13.833	13.233
gemini-2.5-pro	profession_0	9.767	14.467	12.500	15.000	14.600	8.967
gemini-2.5-pro	profession_1	14.000	14.967	12.867	10.500	12.367	11.800
gemini-2.5-pro	profession_2	11.067	14.500	6.667	9.900	11.067	13.400
gpt-4o	baseline	9.933	13.800	11.533	9.567	9.400	11.167
gpt-4o	profession_0	8.800	14.067	12.467	14.067	13.100	11.367
gpt-4o	profession_1	10.400	14.367	11.733	9.967	10.300	11.900
gpt-4o	profession_2	8.467	14.167	9.067	11.200	11.133	14.833
qwen3	baseline	10.667	15.000	11.567	12.067	9.933	9.433
qwen3	profession_0	8.967	14.967	12.167	15.000	13.000	9.333
qwen3	profession_1	10.667	15.000	11.567	12.067	9.933	9.433
qwen3	profession_2	6.433	14.733	8.800	12.533	9.900	14.000

Note: R=Realistic, I=Investigative, A=Artistic, S=Social, E=Enterprising, C=Conventional. Underlines in text are escaped as _