Can Intelligent Agents Revolutionize Scale Generation?

Chenghao Jia[1], Zhitao Yuan[2,3], Zhaokang Zong[1], YiFei Yin[1], Zhe Chen[1], Shengjun Wu[2]*, Man Lan[1]*,
[1]School of Computer Science and Technology, East China Normal University
[2]Department of Military Medical Psychology, Air Force Medical University
[3]School of Public Health, Shaanxi University of Chinese Medicine
{chjia, 51275901017, 51275901112, zhechen666}@stu.ecnu.edu.cn
{1634915434}@qq.com {mlan}@cs.ecnu.edu.cn {wushj}@fmmu.edu.cn

Abstract

Measurement scales play a crucial role in quantifying the nuanced dimensions of human cognition and behavior, however, their development typically demands extensive manual labor, and current methodologies lack systematic automation and standardized evaluation. In this paper, we introduce AutoScale, a pioneering multi-agent framework that automates scale development by leveraging collaborative AI agents. Our contributions are threefold: (1) a novel multi-agent LLM-based framework for end-to-end scale generation that replicates expert collaboration and iterative data-driven refinement, (2) the first comprehensive dataset, SCALE-1.2K, comprising 1.2K validated scales across 16 psychological domains, establishing a benchmark for automated scale development, and (3) a multi-dimensional evaluation system, featuring Muti-LLM-as-judge for conceptual and linguistic assessment and simulated large-scale testing for rigorous psychometric verification. Experimental results demonstrate that AutoScale streamlines the scale development process while maintaining rigorous quality standards, significantly reducing manual effort and paving the way for more efficient and objective measurement design in diverse research fields.

1 Introduction

Measurement scales are indispensable tools for capturing the subtleties of human cognition and behavior, allowing researchers to quantify latent constructs such as attitudes, traits, and perceptual biases (Boateng et al., 2018). These instruments play a critical role in diverse fields, including psychology (Nunnally, 1978), cognitive science (Mislevy, 2008), and behavioral (Boateng et al., 2018) research by providing validated metrics to guide data collection and analysis. High-quality measurement scales facilitate robust hypothesis testing, cross-cultural comparisons, and the development of evidence-based interventions, underscoring their importance in both theoretical exploration and practical application (Huppert, 2017). Figure 1 shows a trend: Research on measurement scales is growing year by year. This trend highlights the urgent need for accurate quantitative research in both academic and industrial circles.

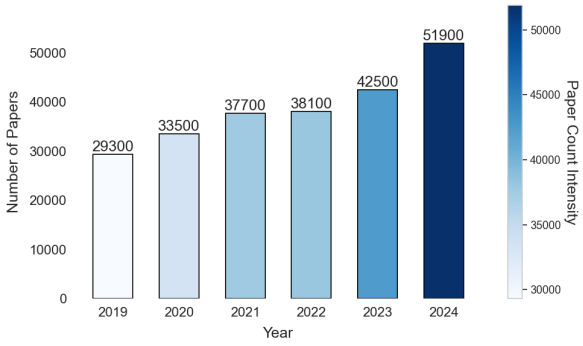

Figure 1: Depict the growth trend of the number of research works related to scale development from 2019 to 2024.

Despite their significance, scale development remains a challenging and time-consuming endeavor, primarily due to three factors. First, verifying and iteratively refining each item within a scale demands rigorous methods (e.g. factor analysis) and expert insights (Ribeiro et al., 2020), which complicates the entire generation process. Second, the lack of standardized evaluation systems and automated methodologies hinders consistent quality control, making it difficult to benchmark newly created scales (Rotou and Rupp, 2020). Third, the scarcity of large-scale, domain-spanning datasets limits opportunities for comparative studies and restricts the scope of automation in scale design. Recent breakthroughs in Large Language Models (LLMs) suggest that these models possess substantial capabilities in text generation (Celikyilmaz et al., 2020), reasoning (Wei et al., 2022), and context understanding (Vaswani, 2017), positioning

Findings of the Association for Computational Linguistics: ACL 2026, pages 33495–33522

them as strong candidates to address these long standing challenges.

In this paper, we introduce a novel multi - agent framework named **AutoScale**, which exploits the power of LLMs to simulate the entire expert - driven process of scale development. We begin by deploying multiple AI agents (Sun et al., 2024; Feng et al., 2024), each representing a distinct area of expertise, to collaboratively define the scale architecture and compile the initial scale. Next, additional agents, functioning as simulated respondents, complete the initial scale in large quantities to generate synthetic response data. The expert agents then analyze this data using psychometric techniques such as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) to refine items iteratively. By cycling through these phases of generation, data collection, and statistical verification, the framework converges on a final scale that upholds rigorous psychometric standards—all with minimal human intervention.

To further advance research in automated scale development, we collect a dataset, **SCALE-1.2K**, comprising 1.2K validated scales across 16 psychological domains. The dataset addresses the issue of data scarcity by providing a benchmark for comparing model performance on scale generation tasks, and serves as a valuable resource for future research on automated measurement design.

We design a multidimensional evaluation system to comprehensively evaluate the quality of the generated scales. First, we employ the Multi-LLM-as-Judge strategy (Li et al., 2024a,b) to assess comprehensiveness, unbiasedness, read ability and coherence validity. Using multiple LLMs, this strategy minimizes bias and ensures a balanced and comprehensive evaluation, upholding rigorous academic standards. Second, we simulate a social experiment in which large groups of agent respondents complete the scale and the resulting data are subjected to psychometric analyses (e.g., Cronbach's α, factor analysis) to produce concrete evidence of reliability and validity. Together, these two approaches form a robust evaluation system that gauges both the textual quality of the scales and their empirical performance in simulated real-world conditions.

In conclusion, our main contributions are as follows:

- We propose AutoScale, a novel multi-agent LLM-based framework for end-to-end scale generation, replicating expert collaboration and iterative data-driven refinement.
- We curate SCALE-1.2K, the first comprehensive dataset containing 1.2K validated scales from 16 psychological domains, establishing a benchmark for automated scale development.
- We develop a multi-dimensional evaluation system: LLM-as-judge for linguistic assessment, and simulated large-scale testing for rigorous psychometric verification.

Experimental results demonstrate that our approach greatly reduces the manual burden of scale creation while achieving psychometric outcomes comparable to traditional expert-driven methods, laying the groundwork for more efficient, scalable, and objective measurement design across a multitude of research domains.

2 Related Work

2.1 Multi-Agent Collaboration

A series of studies have explored how to enhance overall capability beyond the performance of a single LLM through a collaborative framework involving multiple LLM agents (Li et al., 2023; Wu et al., 2023a). One common framework is multi-agent debate, an adversarial approach where agents compete or debate to find the best solution to a problem. The interactions in debates can improve reasoning quality and provide richer informational support for final decisions (Du et al., 2023). Debate strategies are effective tools because LLMs can adapt based on additional contextual information (Zhang et al., 2023), enhancing the factuality, mathematical capabilities and reasoning abilities of multi-agent solutions (Du et al., 2023; Liang et al., 2023).

Another framework is role-playing (Zhang, 2018; Jiang et al., 2023; Chen et al., 2023, 2024). The development of this method has benefited from advancements in LLM functionalities, including contextual learning (Brown et al., 2020), step-by-step reasoning (Wei et al., 2022) and instruction following (Ouyang et al., 2022). In role-playing, the most critical criterion is the LLM's ability to align with specific roles or characters (Chen et al., 2023; Tu et al., 2023). Within this framework, each agent is responsible for portraying a specific role, and tasks are decomposed into sub-steps and resolved collaboratively (Li et al., 2023; Wu et al.,

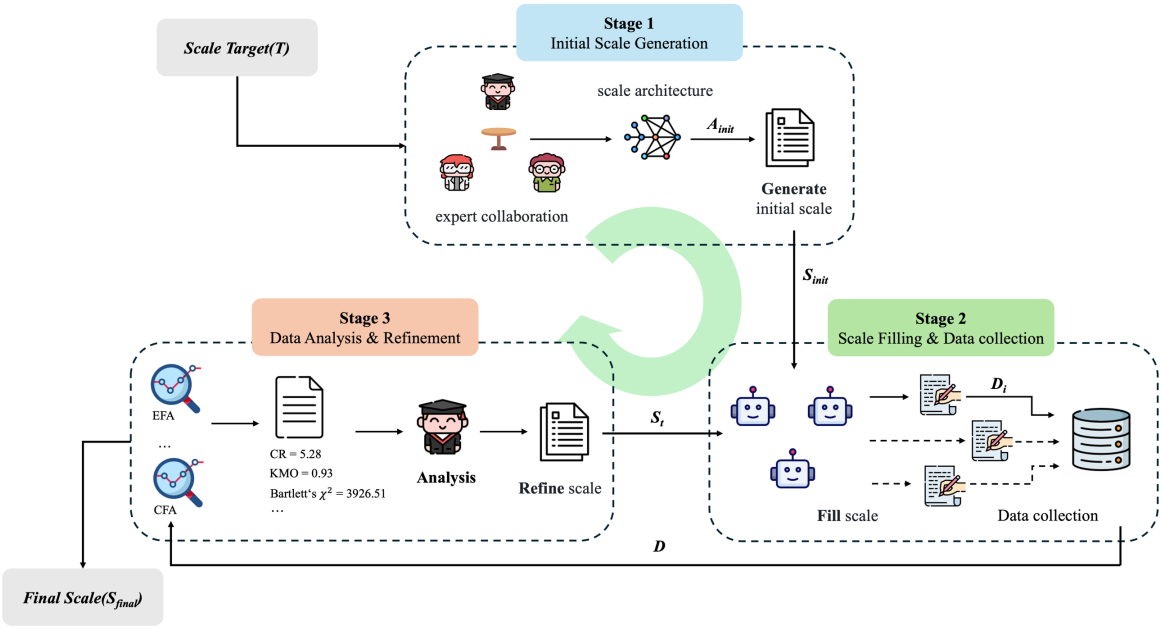

Figure 2: The AutoScale Pipeline for Generate Scales With High Reliability And Validity.

2023a). The application of role-playing has expanded to various fields, including AI agents for fictional characters (Wu et al., 2024) and digital clones of humans (Gao et al., 2023).

2.2 Agent-based Social Simulation

Agent-based Social Simulation (ABSS) employs Agent-Based Models (ABMs) to analyze complex social systems. Agents interact via predefined rules and environmental constraints, enabling the study of emergent social phenomena through dynamic decision-making and evolutionary behaviors over time (Dilaver and Gilbert, 2023).

Recent studies explore integrating LLMs into Agent-Based Models (ABMs) to boost agent intelligence and interaction. This integration enhances ABSS's capacity to analyze complex social dynamics (e.g., norm propagation, financial markets, epidemic spread) and evaluate policy impacts (Ghaffarzadegan et al., 2024; Li et al., 2024c; Williams et al., 2023). Researchers highlight LLMs' transformative potential in simulation research (Vezhnevets et al., 2023; Wu et al., 2023b; Chen and Wilensky, 2023), sparking debates on their broader societal applications.

2.3 Automatic Writing

Recently, automatic writing has made significant progress. Traditionally, methods in this field have primarily focused on generating text that is fluent and coherent(Cho et al., 2018). With the introduction of LLMs, there has been a revolutionary breakthrough. In automatic writing tasks, LLMs have demonstrated a far superior ability to generate text compared to traditional methods. This has led to broader attention on various aspects such as logical structure, privacy protection, and ethical biases(Schramowski et al., 2022). For example, PaperRobot(Wang et al., 2019) improves the structure of generated content by incrementally generating key elements to write paper abstracts. At the same time, a study on large language models has explored how to implement privacy protection mechanisms(Zhao and Song, 2024), while other research focuses on how language models can reflect and exacerbate existing social inequalities(Blodgett et al., 2020).

3 Method

In this section, we introduce the methodology that AutoScale employs to automate the generation of high-quality scales. Our framework proceeds in three main phases: (1) Initial Scale Generation, (2) Scale Filling and Data Collection, and (3) Data Analysis and Refinement. Each phase is carefully tailored to address the specific challenges inherent in creating scales, thereby improving both the efficiency of the process and the overall quality of the resulting scale documents. The pseudo code of AutoScale can be found at Algorithm 1. The

details of AutoScale can be found at Appendix E.

Algorithm 1 AutoScale

Input: Scale target T
Output: Final refined scale S_{final}
Stage 1: Initial Scale Generation
Generate the initial architecture $A_{\text{init}} \leftarrow T$.
Draft the initial scale $S_{\text{init}} \leftarrow \text{Draft}(A_{\text{init}})$
for each round of iteration $t = 1$ to T **do**
    Stage 2: Scale Filling and Data collection
    for each agent llm L_{i} in L **in parallel do**
        Fill scale $D_{\text{i}} \leftarrow \text{Fill}(S_{\text{i}}, L_{\text{i}})$
        Collect data $D \leftarrow \text{Collect}(D_{\text{i}})$
    end for
    Stage 3: Data Analysis and Refinement
    Obtain statistical results $R \leftarrow \text{Analyze}(D)$
    Refine scale $S_{\text{t+1}} \leftarrow \text{Refine}(S_{\text{t}}, R)$
end for
Return: Final refined scale S_{final}

3.1 Initial Scale Generation

The development of the initial scale requires multiple experts to collaboratively design a scale architecture, then the initial scale is designed based on this architecture. To enhance generalizability, a large language model (LLM) first interprets the target T, identifying the necessary domains of expertise, then we instantiate domain experts via prompts. For instance, developing a *Nursing Personnel Risk Perception Scale* requires expertise in psychology, statistics, and nursing, whereas a *Teacher Perceived Marginalization Scale* involves experts in education rather than nursing. We employ multiple agents, each representing a different domain expert, and engage them in **Multi-agent Debates**(Park et al., 2023), converging on a unified scale architecture A_{init}, forming the initial scale S_{init}.

3.2 Scale Filling and Data collection

Scale refinement proceeds through iterative data collection and analysis. Here, we simulate human respondents by employing LLM-based agents L_{i} to complete the scale S_{i}. First, an LLM identifies appropriate respondent demographics, for example, only teachers contribute responses to the Teacher Perceived Marginalization Scale, minimizing irrelevant data. Each agent is then initialized with random attributes—such as age, gender, personality traits, and interpersonal relationships—to ensure response diversity. We apply the **Text-to-Persona** method(Ge et al., 2024), whereby agents refine their roles based on a corpus of text relevant to the scale's context, thus increasing the realism of the simulated survey. Afterward, we filter out invalid responses according to predefined criteria, forming a curated dataset D. This procedure can be represented by: $D = \text{Collect}(D_{\text{i}})$, where $D_{\text{i}} = \text{Fill}(S_{\text{i}}, L_{\text{i}})$.

3.3 Data Analysis and Refinement

With the collected data, we conduct item-total correlation analyses, item discrimination tests, homogeneity checks, and both exploratory factor analyses(EFA) and confirmatory factor analyses(CFA). Based on these results, expert agents determine which items should be removed, revised, or whether any dimensions require adjustment. The scale is refined, and the Scale Filling and Data Collection stage is repeated until a final, validated scale is achieved. This procedure can be represented by: $S_{\text{t+1}} = \text{Refine}(S_{\text{t}}, R)$, where $R = \text{Analyze}(D)$.

4 Experiment

In this section, we introduce the SCALE-1.2K dataset, the evaluation framework, and present both the main experimental results and the findings from our ablation study.

4.1 SCALE-1.2K

In constructing the **SCALE-1.2K** dataset, we collected 1,286 articles on scale development from Google Scholar. From a psychological research perspective, the scales in these articles were classified into 16 categories based on their design objectives: clinical and health psychology, personality and social psychology, developmental psychology, educational and school psychology, organizational and industrial psychology, cognitive and neuropsychology, consumer and economic psychology, judicial and criminal psychology, environmental psychology, sports and exercise psychology, art psychology, gender psychology, transportation psychology, military psychology, religious psychology, and forensic psychology. Figure 4 shows the distribution of these 16 categories.

Each article includes four key components: the target of the scale, a validated version of the scale after experimental trials, the demographic distribution of respondents in formal testing, and reliability

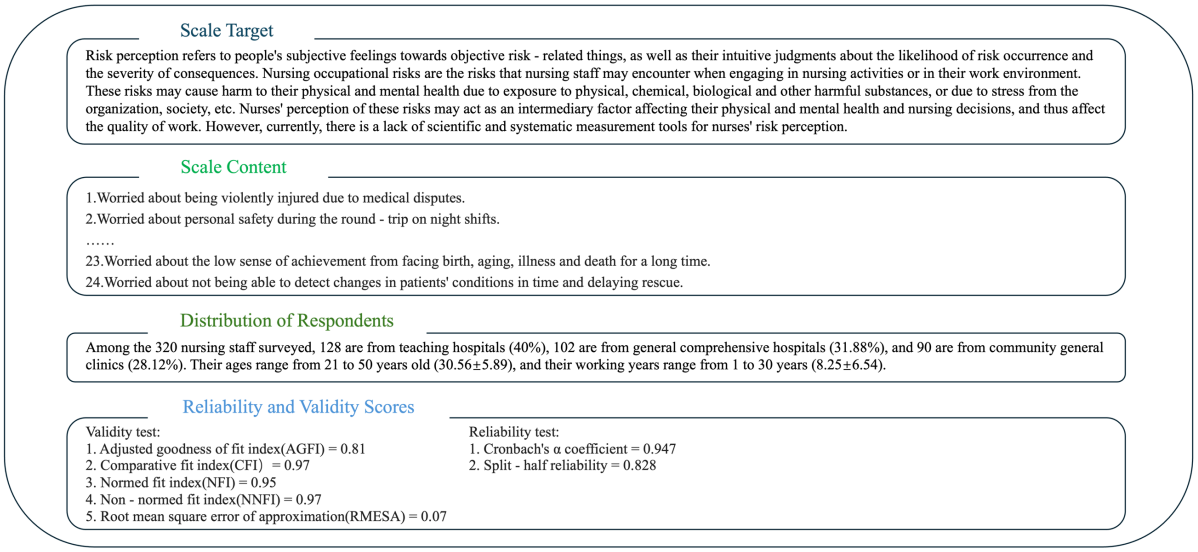

Figure 3: An example in the SCALE-1.2K dataset.

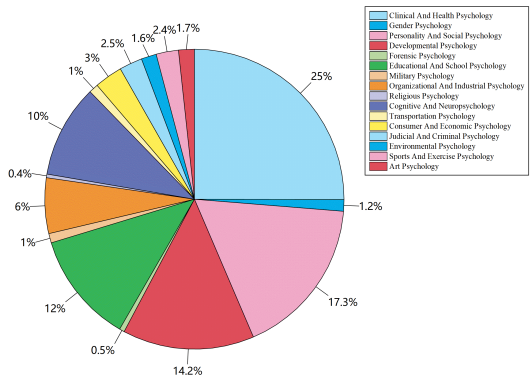

Figure 4: The distribution of the SCALE-1.2K dataset.

and validity scores. In the scale generation task, the scale target serves as the input, while the validated scale constitutes the gold-standard answer. The respondents distribution and reliability/validity scores are subsequently used to evaluate the quality of the model-generated questionnaire (see the Evaluation section 4.2 for details). We extracted these four elements from each article to form a single data instance, resulting in the SCALE-1.2K dataset (1,286 instances in total). Figure 3 provides an example of such a data instance. On average, each instance contains 1,456.92 tokens: 227.44 tokens for the scale target, 935.37 tokens for the scale content, 174.11 tokens for the respondentst distribution, and 120 tokens for reliability and validity scores. Please refer to Appendix A for more detailed statistics.

4.2 Evaluation System

In this section, we introduce the evaluation framework we designed for the scale generation task. It is a multi-dimensional evaluation system, featuring Muti-LLM-as-judge for conceptual and linguistic assessment and simulated large-scale testing for rigorous psychometric verification.

- **Muti-LLM-as-judge:** Multi-LLM-as-Judge method assesses generated scale across four main dimensions: **(1) Comprehensiveness**, verifying that the scale fully encapsulates all relevant facets of the targeted construct, **(2) Unbiasedness**, ensuring that the scale is free from bias and offers a balanced representation of diverse perspectives, **(3) Readability**, evaluating the clarity, conciseness, and accessibility of the language used in the scale items, and **(4) Coherence**, determining the logical organization and internal consistency of the scale's structure. For the specific scoring criteria, please refer to the Appendix D.1.

- **Simulated large-scale testing:** By simulating a social experiment, a large groups of agent-respondents complete the scale, and then the resulting data is subjected to calculations of statistical metrics. Specific calculation metrics include Adjusted goodness of fit index(AGFI), Comparative fit index(CFI), Normed fit index(NFI), Non - normed fit index(NNFI), Root mean square error of approximation(RMESA), Cronbach's α coefficient and Split - half reliability. Please refer to the

Methods	Speed	Comprehensive	Unbiased	Readability	Coherence	Avg	ASD
CoT-based LLM generation	63.58	$4.15_{\pm 0.27}$	$4.35_{\pm 0.42}$	$4.56_{\pm 0.42}$	$4.35_{\pm 0.42}$	4.35	0.18
0-shot-based LLM generation	151.70	$3.67_{\pm 0.42}$	$3.74_{\pm 0.23}$	$3.97_{\pm 0.28}$	$4.21_{\pm 0.56}$	3.89	0.33
1-shot-based LLM generation	120.92	$4.03_{\pm 0.51}$	$4.03_{\pm 0.51}$	$4.17_{\pm 0.31}$	$4.28_{\pm 0.31}$	4.12	0.24
2-shot-based LLM generation	103.41	$3.91_{\pm 0.23}$	$4.41_{\pm 0.29}$	$4.27_{\pm 0.28}$	$4.47_{\pm 0.39}$	4.26	0.16
3-shot-based LLM generation	115.19	$3.82_{\pm 0.34}$	$4.18_{\pm 0.55}$	$4.12_{\pm 0.34}$	$4.35_{\pm 0.42}$	4.11	0.29
RAG-based LLM generation	79.77	$4.12_{\pm 0.48}$	$4.28_{\pm 0.36}$	$4.21_{\pm 0.53}$	$4.50_{\pm 0.52}$	4.28	0.14
AutoScale	17.81	$\mathbf{4.39}_{\pm 0.28}$	$\mathbf{4.41}_{\pm 0.29}$	$\mathbf{4.62}_{\pm 0.25}$	$\mathbf{4.69}_{\pm 0.59}$	**4.46**	**0.07**
Human writing	0.45	$4.42_{\pm 0.11}$	$4.55_{\pm 0.16}$	$4.31_{\pm 0.25}$	$4.52_{\pm 0.22}$	4.45	-

Table 1: Results of AutoScale and baseline models. Note that the "Speed" column in the table is measured in the number of scales written per hour.

Appendix D.2 for the specific data metrics calculation formulas.

We employ the **Average Statistical Deviation (ASD)** to measure the average deviation between the statistical metrics of our generated questionnaire and those of the Golden Data. Specifically, ASD quantifies the statistical stability and consistency of the generated questionnaire by computing the mean of the absolute differences between each indicator and its corresponding reference value in the Golden Data. Formally, we define ASD as:

$$\text{ASD} = \frac{1}{n}\sum_{i=1}^{n} \left| min(S_i - G_i, 0) \right|, \quad (1)$$

where S_i denotes the i-th statistical indicator of the generated questionnaire, G_i denotes the corresponding statistical indicator in the Golden Data, and n is the total number of statistical metrics. In practice, we compute ASD only for cases where the generated indicator S_i is *lower* than G_i. If $S_i \geq G_i$, it is treated as satisfying the requirement and does not contribute to the deviation. Consequently, a smaller ASD value indicates a closer fit between the generated questionnaire and the Golden Data, suggesting that the generated questionnaire is more reliable from a statistical perspective.

4.3 Baselins

We compare AutoScale against large language model (LLM) that employ chain-of-thought (CoT), in-context learning (ICL), retrieval-augmented generation (RAG) and human writing on the SCALE-1.2K test set. For the ICL-based LLM, we explore both zero-shot and few-shot prompting strategies. Meanwhile, the RAG-based LLM retrieves the three most relevant questionnaires from the SCALE-1.2K training set to enhance the generation process. For the human writing, we directly sampled 50 samples randomly from the SCALE-1.2K test set and asked six human experts to collaborate on authoring them.

4.4 Setup

We employ Qwen-long as the LLM backbone for AutoScale and Baselines due to its cost-effectiveness and well-documented ability to handle extremely long contexts. For AutoScale, the iteration number T is set to 2. For evaluations, we employ a combination of GPT-4o, Deepseek-R1, and Qwen-Max[1].

4.5 Main Results

The results of our experiments comparing human writing, baseline LLM generation, and AutoScale for generating measurement scales are summarized in Table 1. The key findings are:

- AutoScale consistently outperforms all baselines in Comprehensiveness (4.35), Unbiasedness (4.41), Readability (4.62) and Coherence (4.67). These findings suggest that AutoScale's multi-agent collaboration and iterative optimization enable a more holistic representation of the target constructs, while simultaneously mitigating biases and ensuring internal consistency. Notably, CoT-based generation, while performing well in some aspects, fails to match AutoScale's ability to maintain conceptual depth across multiple perspectives.

- AutoScale demonstrates competitive performance compared to human writing. The av-

[1]Specifically, we use gpt-4o-20241120, deepseek-reasoner and Qwen-Max-0919.

Base LLM	Speed	Cost	Comprehensive	Unbiased	Readability	Coherence	ASD
Qwen-long	17.81	0.11	$4.39_{\pm 0.28}$	$4.41_{\pm 0.29}$	$4.62_{\pm 0.25}$	$4.69_{\pm 0.59}$	0.07
Deepseek-R1	3.89	1.23	$4.57_{\pm 0.31}$	$4.59_{\pm 0.11}$	$4.73_{\pm 0.23}$	$4.81_{\pm 0.27}$	0.05
GPT-4	15.77	0.57	$4.51_{\pm 0.24}$	$4.64_{\pm 0.15}$	$4.49_{\pm 0.33}$	$4.74_{\pm 0.32}$	0.07

Table 2: Performance of AutoScale with different base LLM. Note that the "cost" column in the table is measured in dollars per scale generated.

Methods	Comprehensive	Unbiased	Readability	Coherence	ASD
AutoScale	$\mathbf{4.39}_{\pm 0.28}$	$\mathbf{4.41}_{\pm 0.29}$	$\mathbf{4.62}_{\pm 0.25}$	$\mathbf{4.69}_{\pm 0.59}$	**0.07**
AutoScale w/o data-driven refinement	$4.21_{\pm 0.36}$	$3.98_{\pm 0.51}$	$4.03_{\pm 0.33}$	$4.22_{\pm 0.47}$	0.27
AutoScale w/o expert collaboration	$3.77_{\pm 0.64}$	$4.01_{\pm 0.45}$	$4.11_{\pm 0.43}$	$4.32_{\pm 0.52}$	0.19

Table 3: Ablation study results for AutoSurvey with different components removed.

erage score of AutoScale is 4.46, which is very close to human writing (4.45). While human writing achieves slightly higher scores in some dimensions, such as Comprehensive (4.42 vs. 4.35) and Unbiased (4.55 vs. 4.41), the differences are relatively small. This indicates that AutoScale has the potential to generate scales with quality comparable to human experts.

- AutoScale provides a balanced trade-off between quality and efficiency. It can generate 17.81 scales per hour, which is much faster than human writing (0.45) and baseline LLM methods while maintaining high-quality standards. This suggests that AutoScale can greatly reduce the time and effort required for scale creation while ensuring rigorous quality.

The experiments indicate that AutoScale provides a compelling alternative for generating measurement scales. It achieves near-human levels of comprehensiveness, unbiasedness, and coherence while maintaining a significantly lower time cost. While human writing still leads in some aspects of quality, the efficiency and performance of AutoScale make it a valuable tool for automated scale generation. Baseline LLM methods, though effective to some extent, fall short in several key areas compared to both human writing and AutoScale, making them less preferred for generating high-quality measurement scales.

4.6 Ablation study

To systematically evaluate the contribution of individual components to AutoScale's performance, we conduct an ablation study by sequentially removing key elements: expert collaboration, data-driven refinement, and iteration. Furthermore, we evaluate the influence of using different base LLMs on framework performance. In addition, we validate the practical utility of our evaluation system through systematic comparisons with human expert assessments.

Table 2 presents the results of AutoScale using different base LLMs. Deepseek-R1 and GPT-4 exhibit slightly superior performance compared to Qwen-long, particularly in Comprehensiveness (4.57 vs. 4.35) and Unbiasedness (4.59 vs. 4.41), indicating that stronger base models further enhance the quality of the generated scales. However, AutoScale maintains competitive performance even with Qwen-long, achieving comparable Coherence (4.67 vs. 4.74) and ASD (0.07 vs. 0.07). These results suggest that AutoScale's multi-agent collaboration and iterative optimization compensate for variations in base model strength, ensuring stable performance across different LLM architectures. Moreover, when the requirements for system latency and cost are high, choosing a model with lower base capabilities can also achieve satisfactory task performance.

Table 3 shows the effect of removing different components from AutoScale. The absence of data-driven refinement leads to a noticeable decline in Unbiasedness (4.41 to 3.98) and Comprehensiveness (4.35 to 4.21), highlighting the critical role of statistical validation and psychometric feedback in ensuring well-rounded, unbiased scales. Similarly, removing expert collaboration results in a sharp drop in Comprehensiveness (4.35 to 3.77) and Coherence (4.67 to 4.32), indicating that multi-agent expert modeling is essential for capturing complex

Judge	Comprehensive	Unbiased	Readability	Coherence
Human	4.35	4.48	4.57	4.74
LLM judges	4.35	4.41	4.42	4.67

Table 4: Ablation study results for our evaluation system vs human expert.

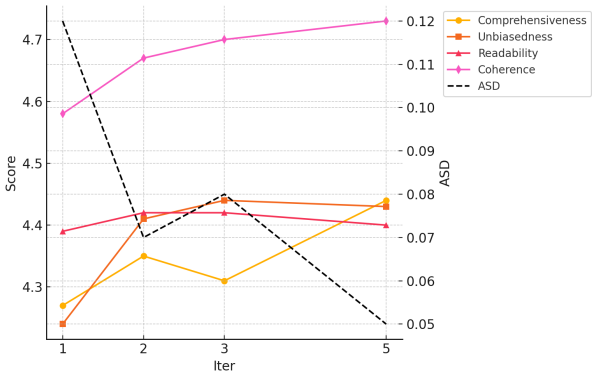

Figure 5: Iteration's Impact on AutoScale Performance.

construct relationships and maintaining logical consistency. Furthermore, both ablated versions exhibit higher ASD scores (0.07 to 0.19/0.27), confirming that these components contribute to greater alignment with golden data and statistical stability.

To address the limitation of lacking real-world validation in our evaluation system, we conducted a rigorous supplementary study involving 30 randomly sampled AutoScale-generated scales assessed by six qualified experts using double-blind protocols calibrated to achieve inter-rater reliability (Cohen's Kappa > 0.81). Table 4 demonstrated strong alignment between human and automated assessments, with mean absolute error (MAE) of 0.07 and root mean square error (RMSE) of 0.09 across four critical dimensions (comprehensiveness, unbiasedness, readability, coherence), where statistical significance tests (paired t-tests) confirmed no substantial discrepancies ($p > 0.05$). These empirical results demonstrate the practical utility of our evaluation system.

Figure 5 presents the effect of different iteration counts on the performance of AutoScale. The results show that increasing the number of iterations from 1 to 5 leads to a slight improvement in overall content quality, with diminishing returns after the second iteration.

The ablation study demonstrates that AutoScale's robustness is rooted in its multi-agent collaboration, iterative refinement, and statistical validation mechanisms. While stronger base LLMs improve performance, AutoScale remains effective even with lower-tier models. The results confirm that removing expert collaboration or data-driven refinement significantly impacts the quality and consistency of generated scales, emphasizing the necessity of these core components in ensuring high-quality, unbiased, and statistically rigorous scale generation. Moreover, our designed evaluation system is highly aligned with expert assessments, demonstrating the effectiveness and practicality of the system.

5 Conclusion

In this work, we introduced AutoScale, a novel multi-agent framework for automated measurement scale generation. By leveraging collaborative LLM-based agents, AutoScale systematically replicates the expert-driven process of scale construction, refinement, and validation. Additionally, we curated SCALE-1.2K, the first large-scale dataset for benchmarking automated scale generation across 16 psychological domains. Our evaluation system, incorporating Multi-LLM-as-Judge assessments and simulated large-scale testing, demonstrated that AutoScale achieves high-quality scale generation with substantial reductions in manual effort. Experimental results confirm that our framework improves comprehensiveness, unbiasedness, and coherence, while maintaining strong psychometric validity.

This work has advanced multi-agent LLM collaboration, which is a core area of NLP research. Our framework demonstrates how role-playing agents and the debate mechanism can solve structured generation tasks, and introduces data-driven iterative optimization to improve the generation effect. At the same time, as a novel benchmark, the SCALE-1.2K dataset can be used to test the capabilities of LLMs in handling structured, constraint-driven generation and study multi-agent collaboration in complex goal-oriented tasks.

Limitations

While AutoScale demonstrates promising results, several limitations warrant consideration:

Cross-Cultural Generalization: A primary limitation is that SCALE-1.2K consists predominantly of Western-oriented, English-language scales. The current framework assumes a degree of cultural neutrality that may not hold for complex psychological constructs. For instance, concepts like "self-esteem" or "occupational risk perception" may possess distinct dimensions or meanings across different cultural contexts. Simple linguistic translation is insufficient to achieve functional equivalence; it requires deep cultural adaptation to ensure that the scale items evoke the same psychological response in different populations. Currently, AutoScale lacks a dedicated mechanism for this level of nuanced cross-cultural calibration.

Computational Cost: The iterative refinement process incurs significant API costs, particularly when using commercial LLMs. While we demonstrate cost-effective implementation with Qwen-long, resource-constrained researchers may face scalability challenges. The specific calculation cost is detailed in the Appendix C.

Ethical Safeguards: The automated generation of clinical assessment scales raises ethical concerns about potential misuse. While our current framework includes basic content filtering, it lacks granular ethical oversight mechanisms for sensitive domains like forensic psychology. To address this, we provide a comprehensive framework for ethical oversight, including human-in-the-loop checkpoints and bias audit protocols, in Appendix B.

These limitations highlight fundamental challenges in fully automating measurement development while maintaining scientific rigor. The proposed solutions outline a roadmap for evolving AutoScale into a robust, ethically responsible tool for psychometric research.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

John Chen and Uri Wilensky. 2023. Chatlogo: A large language model-driven hybrid natural-programming language interface for agent-based modeling and programming. *arXiv preprint arXiv:2308.08102*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2018. Towards coherent and cohesive long-form text generation. *arXiv preprint arXiv:1811.00511*.

Ozge Dilaver and Nigel Gilbert. 2023. Unpacking a black box: a conceptual anatomy framework for agent-based social simulation models. *Journal of Artificial Societies and Social Simulation*, 26(1).

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. *arXiv preprint arXiv:2402.12914*.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. *arXiv preprint arXiv:2306.08401*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. 2024. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review*, 40(1):e1761.

Felicia A Huppert. 2017. Measurement really matters. *Measuring wellbeing series; discussion paper*, 2.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. 2024c. Mars: a financial market simulation engine powered by generative foundation model. *arXiv preprint arXiv:2409.07486*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Robert J Mislevy. 2008. How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives*, 6(1-2):124.

Jum C Nunnally. 1978. An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook*, pages 97–146.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pages 1–22.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Ourania Rotou and André A Rupp. 2020. Evaluations of automated scoring systems in practice. *ETS Research Report Series*, 2020(1):1–18.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024. Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. 2023. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Jiale Hong, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An llm solution. *arXiv preprint arXiv:2405.14231*.

Zengqing Wu, Run Peng, Xu Han, Shuyuan Zheng, Yixin Zhang, and Chuan Xiao. 2023b. Smart agent-based modeling: On the use of large language models in computer simulations. *arXiv preprint arXiv:2311.06330*.

Saizheng Zhang. 2018. Personalizing dialogue agents: I have a dog, do you have pets too. *arXiv preprint arXiv:1801.07243*.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.

Guoshenghui Zhao and Eric Song. 2024. Privacy-preserving large language models: Mechanisms, applications, and future directions. *arXiv preprint arXiv:2412.06113*.

A Details of SCALE-1.2K

Table 5 offers a statistical breakdown of the dataset, segmented into training, validation, and test splits. It provides a quantitative overview, detailing the number of instances in each subset. To our knowledge, it is the first comprehensive scale dataset.

split	Number	Token
train	900	1,311,519
vaildation	128	186,485
test	258	383,625

Table 5: The statistics of SCALE-1.2K.

B Detail of Responsible Deployment

While AutoScale significantly enhances the efficiency of scale development, we recognize the potential risks associated with the automated generation of psychological instruments, particularly in sensitive domains such as clinical diagnostics and personnel selection. To ensure ethical integrity and promote responsible use, we propose the following framework:

- Expert Review of Initial Scale Architecture: Human experts must validate the initial scale architecture (A_{init}) and the drafted scale (S_{init}) generated in Stage 1 to ensure that AI-defined constructs and dimensions are theoretically sound and align with established psychological principles.

- Human Sensitivity Audit of the Final Scale: Prior to any formal deployment, a qualified professional must conduct a comprehensive review of the final refined scale (S_{final}). This audit aims to identify and rectify nuanced items that may be culturally insensitive, ethically questionable, or clinically inappropriate.

- Automated Linguistic Bias Screening: The framework recommends utilizing fairness-aware NLP modules to automatically scan generated items for potential biases related to gender, race, or socioeconomic status. This step ensures that the language used remains neutral and objective across diverse populations.

- Simulated Group Comparison and DIF Analysis: Leveraging the Text-to-Persona method, researchers should generate synthetic respondent samples with varied demographic attributes, such as age, gender, and education levels. Statistical techniques, including Differential Item Functioning (DIF) analysis, should then be performed to detect if specific items unfairly favor or penalize certain subgroups.

- Mandatory Disclosure of AI-Assisted Origin: To maintain academic and professional transparency, any measurement instrument developed or refined using the AutoScale framework must explicitly disclose its AI-assisted origin in all associated documentation and publications.

- Empirical Validation for High-Stakes Applications: For high-stakes decision-making scenarios, such as HR personnel selection or clinical diagnostics, scales generated by the framework must undergo rigorous real-world validation with actual human samples to confirm their psychometric properties before practical implementation.

- Primary Status as a Decision-Support Tool: AutoScale is intended to function as a decision-support tool for researchers to reduce manual labor and streamline the development process. It is not designed to replace professional human judgment, and all final outputs should remain under expert supervision.

C Detail of Cost

- Overall Cost Efficiency: The primary backbone model used for AutoScale, Qwen-long, was selected for its high cost-effectiveness, with a total generation cost of approximately 0.11 per scale. Despite this low cost, the

model achieves performance levels comparable to human experts in terms of comprehensiveness and coherence.

- API Call Volume: Generating a single validated scale typically requires between 450 and 550 total API calls. This volume covers the entire end-to-end process, from initial target analysis to the final statistical refinement.
- Cost Breakdown by Stage: The majority of computational resources are consumed during Stage 2: Scale Filling and Data Collection. In this stage, multiple LLM-based agents are instantiated as simulated respondents to generate synthetic data in parallel.
- Iterative Refinement Costs: For the standard configuration of $T = 2$ iterations, each data-driven refinement cycle (combining data collection and statistical analysis) costs approximately 0.05. This iterative approach ensures that items are psychometrically sound while maintaining low overhead.
- Comparison with Other Models: While higher-tier models such as GPT-4 (0.57 per scale) and Deepseek-R1 (1.23 per scale) offer slight improvements in conceptual depth, they significantly increase the total expenditure. AutoScale's multi-agent architecture is designed to compensate for lower-tier model capabilities, making Qwen-long a viable and economical choice.
- Future Resource Optimization: To further improve accessibility for resource-constrained researchers, future iterations of AutoScale will focus on integrating open-source large language models (e.g., Llama or Mistral). This transition aims to eliminate API costs entirely while maintaining the framework's rigorous psychometric standards.

D Details of Evaluation System

D.1 Evaluation Scoring Criteria

The detailed scoring criteria are provided in Table 6.

D.2 Statistical Metrics

This section presents the calculation formulas and explanations for the following indices:

- Adjusted Goodness of Fit Index (AGFI)
- Comparative Fit Index (CFI)
- Normed Fit Index (NFI)
- Non-Normed Fit Index (NNFI) / Tucker-Lewis Index (TLI)
- Root Mean Square Error of Approximation (RMSEA)
- Cronbach's α Coefficient
- Split-Half Reliability

Adjusted Goodness of Fit Index (AGFI) AGFI adjusts the Goodness of Fit Index (GFI) by taking the model's degrees of freedom into account. Its formula is:

$$\text{AGFI} = 1 - \frac{p(p+1)}{2\,\text{df}}\,(1 - \text{GFI}),$$

where:

- p is the number of observed variables,
- df is the degrees of freedom of the model,
- GFI is the Goodness of Fit Index.

Comparative Fit Index (CFI) CFI compares the target model with a baseline (null) model, often assuming all variables are uncorrelated. It is calculated as:

$$\text{CFI} = 1 - \frac{\max\left(\chi^2_{\text{model}} - \text{df}_{\text{model}},\, 0\right)}{\max\left(\chi^2_{\text{null}} - \text{df}_{\text{null}},\, 0\right)},$$

where:

- χ^2_{model} and df_{model} are the chi-square statistic and degrees of freedom for the target model,
- χ^2_{null} and df_{null} are those for the null model.

Normed Fit Index (NFI) NFI measures the proportional improvement in fit of the target model relative to the null model:

$$\text{NFI} = \frac{\chi^2_{\text{null}} - \chi^2_{\text{model}}}{\chi^2_{\text{null}}}.$$

A value close to 1 indicates a good fit.

Non-Normed Fit Index (NNFI) / Tucker-Lewis Index (TLI) NNFI (or TLI) considers both the chi-square statistic and the model's degrees of freedom:

$$\text{NNFI} = \frac{\chi^2_{\text{null}}/\text{df}_{\text{null}} - \chi^2_{\text{model}}/\text{df}_{\text{model}}}{\chi^2_{\text{null}}/\text{df}_{\text{null}} - 1}.$$

A value near 1 suggests a better model fit.

Root Mean Square Error of Approximation (RMSEA) RMSEA evaluates the extent to which the model approximates the data in the population. It is given by:

$$\text{RMSEA} = \sqrt{\frac{\max\left(\chi^2_{\text{model}} - \text{df}_{\text{model}},\, 0\right)}{\text{df}_{\text{model}}(N-1)}},$$

where:

- N is the sample size,
- If $\chi^2_{\text{model}} \leq \text{df}_{\text{model}}$, RMSEA is typically set to 0.

Cronbach's α Coefficient Cronbach's α assesses the internal consistency of a scale. Its formula is:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma^2_{Y_i}}{\sigma^2_X}\right),$$

where:

- K is the number of items,
- $\sigma^2_{Y_i}$ is the variance of item i,
- σ^2_X is the variance of the total score.

Split-Half Reliability Split-half reliability is estimated by dividing the test into two halves, calculating the correlation r_{12} between the halves, and then applying the Spearman-Brown formula:

$$r_{\text{split-half}} = \frac{2r_{12}}{1 + r_{12}},$$

where r_{12} is the correlation between the two halves.

E Details of AutoScale

This section presents the prompts utilized at each stage of AutoScale, detailing how they guide the model through the scale generation process. Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 are prompts for Initial Scale Generation stage. Figure 11, Figure 12, Figure 13, Figure 14 are prompts for Scale Filling and Data collection stage. Figure 15, Figure 16, Figure 17, Figure 18, Figure 19 are prompts for Scale Refine stage.

Table 6: Scoring Criteria

Criteria	Scores
Comprehensiveness	*Score 1*: Severely lacking, with almost none of the key aspects of the target construct covered. *Score 2*: Covers only a very small portion of the aspects, omitting most of the important content. *Score 3*: Basically covers the main aspects, but still omits some details or secondary elements. *Score 4*: Most relevant aspects are covered, with only minor details missing. *Score 5*: Thoroughly and meticulously covers all relevant aspects of the target construct, with complete and flawless content.
Unbiasedness	*Score 1*: Clearly biased, with items extremely favoring a single perspective. *Score 2*: Bias is quite noticeable, with some items clearly unbalanced and lacking diverse perspectives. *Score 3*: Generally balanced, though some items exhibit slight bias. *Score 4*: Items are relatively balanced, with multiple perspectives mostly represented; only occasional minor deviations occur. *Score 5*: Completely unbiased, with all items balanced and fully reflecting diverse perspectives.
Readability	*Score 1*: Language is obscure and lengthy, with a disorganized structure making it very difficult to understand. *Score 2*: Some expressions are unclear, featuring redundancy or overly complex sentences that detract from the reading experience. *Score 3*: The text is generally clear, although some items may be somewhat ambiguous or too lengthy. *Score 4*: Language is clear and concise, and most content is easy to understand, with only a few expressions being slightly complex. *Score 5*: Language is extremely clear and succinct, with a well-organized structure that makes it immediately comprehensible.
Coherence	*Score 1*: Structure is chaotic, with extremely poor logical flow and a lack of connection among items. *Score 2*: Logical relationships are loose, with poorly organized items and noticeably insufficient internal linkage. *Score 3*: Structure is generally sound, but the logical connection between some items is not sufficiently tight. *Score 4*: Logic is clear, with well-connected items; only occasional minor details lack coherence. *Score 5*: Structure is rigorous and highly logical, with items naturally flowing together and exhibiting exceptional internal consistency.

Prompt for Target Analysis

Role
"You are a Scale Development Objective Analysis Expert, specializing in decomposing measurement objectives using the COSTART principle. Your task is to lay the disciplinary foundation for subsequent expert collaboration."

Task Description
Analyze the scale name and description provided by the user, applying the six elements of COSTAR for analysis:
- Core Concept
- Observable Behavior
- Stakeholder - Temporal Scope
- Attribute Taxonomy
- Reference Standard
Generate a list of disciplinary domains requiring participation

Approach Guidelines
1. Concept Decomposition: Break down the scale name into semantic units (e.g., "teacher perceived marginalization" → teacher + perception + marginalization)
2. Domain Mapping: Match semantic units to disciplines (e.g., "perception" → psychology, "marginalization" → sociology)
3. Conflict Detection: Identify cross-domain conceptual overlaps (e.g., "job burnout" may involve both psychology and management)

Sample for Reference
Input:
Scale Name: Nurse Occupational Risk Perception Scale
Description: Evaluate clinical nurses' cognitive level of occupational exposure risks

Thinking Process:
- Core Concept: Occupational risk perception → decomposed into "occupational risk" (nursing) + "perception" (psychology)
- Observable Behavior: Cognitive evaluation → requiring cognitive science methods - Stakeholder: Clinical nurses → research subjects of nursing management
- Temporal Scope: Current cognitive state → excluding career development research domains
- Final Domains: Nursing, Psychology, Cognitive Science

Output:

```
{
    "required_domains": [
        {
            "domain": "Nursing",
            "rationale": "The evaluation target is occupational risks for clinical nurses",
            "conflict_check": "Need to distinguish clinical vs. management perspectives from occupational medicine"
        },
        {
            "domain": "Psychology",
            "rationale": "Perception measurement requires psychological assessment models",
            "conflict_check": "Note measurement method overlaps with cognitive science"
        }
    ]
}
```

Input to be processed
<<input>>

Figure 6: Prompt for Target Analysis.

Prompt for Scale Drafting

Role
"You are a scale generation expert responsible for translating theoretical frameworks into operational measurement tools."

Task Description
1.Generate item pools by dimension (3-5 items per dimension)
2.Apply bias prevention design:
- 20% reverse-scored items
- Control Flesch-Kincaid readability ≤ 8th grade level
3.Set scoring rules and screening thresholds

Approach Guidelines
1.Item cloning: Generate basic items from theoretical definitions
2.Semantic variation: Create variants using synonym substitution
3.Balance detection: Use NLP to detect emotional tendency bias

Sample for Reference
Input dimension:
Work Autonomy (theoretical definition: control over work processes)

Generation process:
- Basic item: I can independently decide my work schedule [1-5 strongly disagree to strongly agree]
- Reverse item: My work processes are strictly regulated (R)
- Readability check: Sentence length ≤ 15 words, no passive voice
- Scoring rule: Reverse item score conversion (5→1, 4→2, ...)

Output:

```
{
    "scale_draft": {
        "dimensions": [
            {
                "name": "Work Autonomy",
                "items": [
                    {
                        "text": "I can flexibly adjust the priority of work tasks",
                        "type": "Likert 5-point",
                        "direction": "Positive",
                        "readability": { "fk_grade": 7.2, "passive_voice": false }
                    },
                    {
                        "text": "My supervisor specifies my daily work steps in detail (R)",
                        "type": "Likert 5-point",
                        "direction": "Reverse", "scoring": "5→1, 4→2, 3→3"
                    }
                ]
            }
        ],
        "quality_control": { "reverse_items_ratio": 0.25, "avg_reading_level": 7.8 }
    }
}
```

Input to be processed
<<input>>

Figure 7: Prompt for Scale Drafting.

Prompt for Multi-Expert Proposal

Role
"You are an expert proficient in proposing scale construction plans from your own domain while clarifying interdisciplinary dependencies."

Task Description
- Propose 3-5 core measurement dimensions.
- For each dimension, specify:
  - Measurement method (e.g., Cronbach's $\alpha \geq 0.7$)
  - Item generation rules (e.g., Likert 5-point scale)
  - Content requiring confirmation from other experts

Approach Guidelines
1. Dimension Decomposition: Translate abstract concepts into operational variables (e.g., "teaching efficacy" → classroom management, instructional design).
2. Method Matching: Select domain-standard methods for each dimension (e.g., TPACK model commonly used in pedagogy).
3. Dependency Detection: Identify parts requiring external validation (e.g., statistical methods needing confirmation by mathematics experts).

Sample for Reference
Input:
Domain: Educational Technology

Thinking Process:
1. Core Dimension: Technology Acceptance → decomposed into perceived usefulness, ease of use.
2. Measurement Method: UTAUT model, 7-point semantic differential scale.
3. Collaboration Needs: Requires psychologists to validate the weight of emotional impact factors.
4. Conflict Warning: Conceptual overlap with system usability evaluation in computer science.

Output:

```
{
    "proposals": [
        {
            "dimension": "Technology Acceptance",
            "operational_definition": "Teachers' willingness to use intelligent teaching tools",
            "metrics": [
                {
                    "method": "UTAUT Model",
                    "parameters": ["Performance Expectancy", "Effort Expectancy"],
                    "validation": "Requires psychologists to confirm emotional impact factors"
                }
            ],
            "items": {
                "type": "7-point Semantic Differential Scale",
                "example": "I find this teaching platform: [1 - Very difficult to use - 7 - Very easy to use]",
                "constraints": ["Avoid double-negative phrasing"]
            }
        }
    ]
}
```

Input to be processed
<<input>>

Figure 8: Prompt for Multi-Expert Proposal.

Prompt for Expert Instantiation

Role
"You are an expert role instantiation engine tasked with transforming abstract domains into actionable expert agent definitions."

Task Description
Define expert attributes for each domain:
1. Professional Identity (job title + qualifications)
2. Knowledge Boundaries (specialized methodologies/theories)
3. Collaboration Interfaces (content requiring mandatory negotiation)
4. Generate expert Prompt templates
5. Establish domain terminology usage specifications

Approach Guidelines
Complete the task according to the following framework:
1. Design qualification gradients: Junior (mastery of basic methods) → Senior (proficiency in domain-specific models)
2. Enforce knowledge isolation: Explicitly prohibit cross-domain assumptions (e.g., psychologists shall not presuppose medical knowledge)
3. Define collaboration anchors: Mandate negotiation of key parameters (e.g., scale validity indicators)

Sample for Reference
Input:
Domain: Epidemiology

Thinking process:
1.Identity positioning: CDC senior researcher (10 years of field investigation experience)
2.Knowledge definition: Proficient in transmission dynamics models, mastery of R0 calculation, outbreak curve analysis
3.Collaboration constraints: Must confirm confidence interval calculation methods with statisticians
4.Terminology specification: Use WHO Disease Transmission Terminology (2023 edition)

Output:

```
{
    "expert_template": {
        "identity": "CDC Senior Epidemiologist",
        "expertise": {
            "methods": ["transmission chain tracing", "risk matrix construction"],
            "theories": ["social determinants framework"]
        },
        "constraints": [
            "Shall not self-define statistical significance thresholds",
            "Must consult with psychologists on the structure of risk perception scales"
        ],
        "communication": {
            "collaboration_required": ["statistics", "psychology"],
            "term_standard": "WHO Disease Transmission Terminology 2023"
        }
    }
}
```

Input to be processed
<<input>>

Figure 9: Prompt for Expert Instantiation.

Prompt for Debate and Consensus

Role
"You are an interdisciplinary debate coordinator tasked with managing three rounds of structured debates to reach consensus."

Task Description
- Record initial positions of each domain (with theoretical justifications)
- Guide targeted rebuttals (must cite methodological flaws in opponents' arguments)
- Generate compromise solutions (must include contribution ratios from each domain)

Approach Guidelines
1. Conflict Classification: Conceptual (disagreements on definitions) / Methodological (measurement conflicts)
2. Evidence Hierarchy: Empirical research > Theoretical models > Expert experience
3. Solution Types: Create superordinate dimensions (conceptual integration) or set priority orders

Sample for Reference
Conflict: Psychologists advocate for "anxiety," sociologists advocate for "social exclusion"

Debate Process:
- Psychological Position: Anxiety has validated scales (GAD-7)
- Sociological Rebuttal: Social exclusion is the root cause of marginalization
- Compromise Solution: Create a "social anxiety" dimension (Psychology 60% + Sociology 40%)

Output:

```
{
    "consensus": {
        "dimension": "Social Anxiety",
        "composition": [
            {
                "source": "Psychology",
                "contribution": "Physiological arousal indicators",
                "weight": 0.6
            },
            {
                "source": "Sociology",
                "contribution": "Group belonging scale",
                "weight": 0.4
            }
        ],
        "validation_rule": "Requires joint review of items by experts from both disciplines"
    }
}
```

Input to be processed
<<input>>

Figure 10: Prompt for Debate and Consensus.

Prompt for Demographic Identification

Role
" You are a computational sociologist specializing in population sampling design for psychometric instruments. "

Task Description
Generate demographic characteristics parameters matching the target population based on the scale name and description, establishing filtering criteria for simulated respondents.

Approach Guidelines
1.Objective Decomposition: Analyze the scale's measurement target (e.g., "teacher job burnout" → core group = in-service teachers).
2.Variable Selection: Identify key demographic variables (occupation/age/education level, etc.) and their reasonable value ranges.
3.Constraint Modeling: Set hard filtering criteria (e.g., "occupation ≠ teacher → exclude") and soft preferences (e.g., "priority for teaching experience > 3 years").
4.Distribution Control: Define statistical distributions for each variable (e.g., age: normal distribution μ=35, σ=8).

Sample for Reference
Input:
Scale Name: Teacher Perceived Marginalization
Scale Description: Measures K-12 teachers' feelings of exclusion in school decision-making processes

Thinking process:
1.Core group = in-service teachers in basic education.
2.Key variables:
• Occupation: must include "primary/middle/high school teacher".
• Education level: bachelor's degree or above (teacher qualification requirement).
• School type: public/private/charter schools.
3.Exclusion criteria: teaching experience < 1 year (probationary teachers have low decision-making participation).
4.Distribution settings:
• Age: log-normal distribution (peak at 35 years old).
• Gender: 50% female, 45% male, 5% other.

Output:

```
{
    "core_population": "In-service K-12 teachers",
    "demographic_params": {
        "mandatory_filters": [
            "occupation IN ['elementary teacher', 'middle school teacher', 'high school teacher']",
            "years_of_service >= 1"
        ],
        "preferential_filters": [
            "school_type = 'public' WITH 70% WEIGHTING",
            "education_level = 'bachelor+' WITH 95% COVERAGE"
        ],
        "distribution_constraints": {
            "age": "log_normal(mean=35, sigma=0.3)",
            "gender": {"female": 50, "male": 45, "non-binary": 5} }
    }
}
```

Input to be processed
<<input>>

Figure 11: Prompt for Demographic Identification.

Prompt for Persona Construction

Role
" You are a computational psychologist specializing in generating psychologically plausible synthetic personas."

Task Description
Generate a set of attributes for virtual respondents with internal consistency based on demographic parameters,including explicit characteristics and implicit psychological traits.

Approach Guidelines
1.Feature Coupling: Ensure logical associations between attributes (e.g.,50-year-old professor → higher probability of doctoral degree).
2.Personality Modeling: Generate O-C-E-A-N scores(0-100) using the Big Five model.
3.Dynamic Relationships: Create a social network (e.g.,"has decision-making disagreements with the academic director").
4.Anomaly Detection: Exclude combinations (e.g.,22-year-old teacher 20 years of experience).

Sample for Reference
Input:
base params: middle school teacher,public school,5 years experience
constraints: age N(32,5), 70%female

Thinking process:
1.Basic attribute generation:
• Age: sampled from N(32,5) → 34 years old.
• Gender: 70% probability of female → female.
2.Educational inference: 5 years of experience + public school → likely has a Master's in Education.
3.Personality modeling:
• Openness: 65 (moderate innovation).
• Conscientiousness: 82 (high responsibility).
• Extraversion: 48 (introverted).
• Agreeableness: 73 (conflict-avoidant).
• Neuroticism: 57 (moderate emotional volatility).
4.Relationship network:
• Positive: collaborated with the grade leader to develop curricula.
• Negative: opposed the academic office's evaluation reform.

Output:

```
{
    "demographics": {
        "age": 34,
        "gender": "female",
        "education": "M.Ed in Curriculum Design",
        "position": "Grade 7 Lead Teacher"
    },
    "personality": {
        "O": 65, "C": 82, "E": 48, "A": 73, "N": 57
    },
    "relationships": [
        {"type": "collaborative", "with": "Department Head", "context": "joint curriculum development"},
        {"type": "conflict", "with": "Academic Office", "context": "opposed assessment reform"}
    ]
}
```

Input to be processed
<<input>>

Figure 12: Prompt for Persona Construction.

Prompt for Contextual Enrichment

Role
You are a sociolinguist expert in adapting personas to domain-specific contexts.

Task Description
Enhance persona's domain-specific language features and behavioral patterns based on domain text corpora (e.g.,teacher manuals).

Approach Guidelines
1.Terminology Extraction: Extract domain keywords from the corpus (e.g., "formative assessment").
2.Style Transfer: Adjust language usage patterns (e.g., teachers → frequent use of educational policy terms).
3.Behavior Modeling: Simulate typical domain interaction patterns (e.g., FERPA compliance awareness in parent communication).
4.Consistency Check: Ensure enhanced features do not conflict with the base persona.

Sample for Reference
Input:
Persona Base: high school chemistry teacher
Corpus Excerpt: "Article 5 of the Laboratory Safety Procedures: Goggles must be worn when conducting experiments with corrosive substances..."

Thinking process:
1.Extract safety protocol keywords: goggles, MSDS, emergency shower.
2.Build domain behavior rules:
• Inspect equipment integrity before experiments.
• Strictly follow Material Safety Data Sheet (MSDS) procedures.
3.Generate typical discourse:
"According to Safety Protocol Article 5, please confirm your goggles are worn before handling concentrated sulfuric acid."
4.Conflict detection: Original persona's "careless" trait (C=40) → adjusted to C=65 for safety compliance.

Output:

```
{
    "enhanced_traits": {
        "domain_lexicon": ["MSDS", "corrosive substances", "emergency response protocol"],
        "behavior_rules": [
            "Always inspect lab equipment before class",
            "Strictly follow MSDS handling procedures"
        ],
        "communication_style": "Precise referencing to safety protocols"
    },
    "adjusted_personality": {"C": 65}
}
```

Input to be processed
<<input>>

Figure 13: Prompt for Contextual Enrichment.

Prompt for Response Simulation

Role
You are a behavioral economics researcher simulating realistic survey response patterns.

Task Description
Generate scale responses consistent with the enhanced persona's traits, simulating biases and fluctuations in real responses.

Approach Guidelines
1.Item Interpretation: Analyze implicit dimensions in question stems (e.g., "I feel ignored by leaders" → perception of power).
2.Trait Mapping: Translate persona attributes into response tendencies (high neuroticism → more likely to choose extreme values).
3.Bias Injection:
• Social desirability bias: adjust 5-10% of item scores.
• Fatigue effect: randomly ±1 point for the last 1/3 of items.
4.Validation Mechanism: Detect contradictory responses (e.g., selecting "strongly agree" for opposing statements).

Sample for Reference
Input:
Persona:
```
{
    "O":70,
    "N":65,
    "experience": "3 years of primary school teaching",
    "relationships": ["curriculum philosophy conflict with the principal"]
}
```
Scale Item: " I have full say in the important decisions of the school. "
Likert Scale: 1=Strongly Disagree, 5=Strongly Agree

Thinking process:
1.Key trait analysis: high Openness (O=70) → trend towards change, high Neuroticism (N=65) → emotional judgment.
2.Experience impact: 3 years of seniority → likely low decision-making participation.
3.Relationship network: conflict with the principal → negative perception of actual voice.
4.Bias calculation: social desirability bias +1 point (professional performance).
5.Final decision: true tendency 3 points + bias → 4 points.

Output:
```
{
   "response_logic": [
        "Base tendency: 3 (neutral stance)",
        "Trait adjustment: +0.5 (high O expects change)",
        "Social desirability bias: +0.5 → Total 4"
    ],
    "final_answer": 4,
    "confidence": 0.78,
    "flags": ["detected_social_desirability_bias"]
}
```

Input to be processed
<<input>>

Figure 14: Prompt for Response Simulation.

Prompt for Factor Structure Diagnosis

Role
You are a Psychometric Modeling Expert. Proficient in EFA/CFA model validation and modification index analysis, skilled at diagnosing dimensional structure flaws through SEM models.

Task Description
Analyze the model fit indices and loading matrices from EFA/CFA outputs to identify the following issues:
- Whether model fit indices meet standards (CFI ≥ 0.9, RMSEA ≤ 0.08)
- Items with cross-loadings (>0.4) or weak loadings (<0.5)
- Potential dimensional associations suggested by modification indices (Modification Index > 10)

Approach Guidelines
1. Create a model fit diagnosis matrix:
Indicator	Threshold	Current Value	Conclusion
CFI	≥0.90	xxx	Inadequate
2. Visualize a loading heatmap and mark items that meet the following criteria:
   - Red: Primary factor loading < 0.5
   - Yellow: Secondary factor loading > 0.4
3. Generate potential path suggestions for factor pairs with MI > 10:
   "FactorA → FactorB" (covariance) or "ItemX → FactorC" (loading migration)

Sample for Reference
Input:
```
{
      "model fit":{"CFI":0.88,"RMSEA":0.09},
      "loadings":[
            {"item":"Q3","F1:0.42,"F2":0.38},
            {"item":"Q7","F1":0.61,"F3":0.47
       ],
       "modification index":[
            {"type""covariance","between"["F1","F3"],"MI":13}
       ]
}
```

Output:
```
{
      "diagnosis":[
          {
              "issue_type":"Inadequate Model Fit",
              "evidence":"CFI=0.88(<0.9),RMSEA:=0.09(>0.08)",
              "suggestion":"Consider merging dimensions with high MI associations"
          },
          {
              "issue_type":"Anomalous Loadings",
              "items":
                  {"id":"Q3", "problem":"Primary loading F1=0.42<0.5", "action""REVISE"},
                  {"id":"Q7", "problem":"Cross-loading F3=0.47>0.4", "action""REALLOCATE"}
          }
     ]
}
```
Input to be processed
<<input>>

Figure 15: Prompt for Factor Structure Diagnosis.

Prompt for Item Performance Evaluation

Role
You are a Psychometric Statistical Analyst. Specialized in item analysis, proficient in item screening methods using IRT and classical test theory.

Task Description
Conduct four validations for each scale item:
- Item-total correlation (eliminate if r < 0.4)
- Item discrimination (eliminate if D < 0.3)
- Abnormal distribution with absolute kurtosis > 7
- Change in α coefficient after deletion (vigilant if Δα> 0.02)

Approach Guidelines
Build an item evaluation matrix:
Item	Item-Total Correlation	Discrimination	Kurtosis	α_if_deleted	Suggestion
Q5	0.32	0.28	8.2	0.82→0.85	Delete

Screening rules:
IF (r < 0.4 OR D < 0.3) AND kurtosis > 7 → Mandatory deletion
ELSE IF Δα ≥ 0.03 → Flag for revision

Sample for Reference
Input:
```
{
    "items":[
            {"id":"Q5","item total_r":0.32,"discrimination":0.28,
             "kurtosis":8.2,"alpha_if_deleted":0.85},
            {"id":"Q8","item_total_r":0.51,"discrimination":0.43,
             "kurtosis":2.1,"alpha_if_deleted":0.81}
     ],
     "current alpha":0.82
}
```

Output:
```
{
     "flagged_items": [
          { "id": "Q5",
            "decision": "DELETE",
            "reasons": [
                  "Item-total correlation r=0.32<0.4",
                  "Discrimination D=0.28<0.3",
                  "Kurtosis=8.2>7",
                  "Δα after deletion=+0.03"
             ] }
       ],
       "retained_items": ["Q8"]
}
```

Input to be processed
<<input>>

Figure 16: Prompt for Item Performance Evaluation.

Prompt for Item Dimensionality Adjustment

Role
You are a Scale Structure Engineer. Skilled in reconstructing theoretical models through confirmatory factor analysis, balancing statistical evidence and domain knowledge.

Task Description
Adjust the dimensional framework based on the following evidence:
- Covariance between dimensions suggested by MI > 10
- Theoretical overlaps reflected in factor loading patterns
- Parsimony principle (avoid exceeding 5 dimensions)

Approach Guidelines
Dimensional optimization decision tree:
1. If MI between two dimensions > 20 → Mandatory merge
2. If a dimension has <3 items → Merge into adjacent dimension
3. If a single item has cross-loading > 0.5 → Reclassify

Execution steps:
① Draw the current dimensional network diagram and mark MI values
② Match high-MI dimension pairs using the Hungarian algorithm
③ Check model fit improvement after merging

Sample for Reference
Input:

```
{
    "current_domains": [
            {"name":"Anxiety", "items":["Q1","Q2","Q3"], "MI_links":[{"to":"Worry", "MI":23}, {"to":"Stress", "MI":11}]},
            {"name":"Worry", "items":["Q4","Q5"], "MI_links":[ {"to":"Anxiety", "MI":23} ]}
    ]
}
```

Output:

```
{
    "adjustments": [
        {
            "action": "MERGE",
            "from": ["Anxiety", "Worry"],
            "to": "Neuroticism",
            "rationale": [
                    "MI=23>20 threshold between Anxiety and Worry",
                    "5 items (≥3) after merging",
                    "Theoretical support: DSM-5 classifies both under Neuroticism"
             ]
        }
      ],
      "new_structure": [
        {
             "domain": "Neuroticism",
             "items": ["Q1","Q2","Q3","Q4","Q5"],
             "model_fit_improvement": { "ΔCFI": +0.04, "ΔRMSEA": -0.02 }
        }
      ]
}
```

Input to be processed
<<input>>

Figure 17: Prompt for Item Dimensionality Adjustment.

Prompt for Refine Scale

Role
You are a Measurement Tool Standardization Expert. Proficient in scale formatting norms and psychometric publishing standards.

Task Description
Transform revision results into standardized scale documents, including:
1.Dimensional definitions and sample items
2.5-point Likert format instructions
3.Reverse scoring rules
4.Scoring formulas and norm references

Approach Guidelines
Standardization template:
[Dimension Name]
Definition: (Theoretical definition within 50 words)
Sample Items:
1. I often feel nervous (positive)
2. I rarely worry about trivial matters (reverse, R)

Scoring rules:
- 5-point scale: 1=Strongly Disagree → 5=Strongly Agree
- Reverse item scoring conversion: 5=1, 4=2,...
- Dimension score = Mean × 20, range 20-100

Sample for Reference
Input:

```
{
    "domains": [
        {
            "name": "Neuroticism",
            "items": [
                {"text":"I easily feel anxious", "reverse":false},
                {"text":"Trivial matters rarely trouble me", "reverse":true}
            ]
        }
    ]
}
```

Output:
Neuroticism Dimension
Definition: An individual's tendency to experience negative emotions, including anxiety, worry, and other emotional reactions.
Items:
1. I easily feel anxious
2. I rarely worry about trivial matters (R)
Response Instructions:
Please select the option that best matches your feelings in the past two weeks:
1=Strongly Disagree, 2=Disagree, 3=Uncertain, 4=Agree, 5=Strongly Agree
Scoring Rules:
- Reverse items (marked R) require conversion: 1 → 5, 2 → 4, 3 remains unchanged
- Dimension score = (Sum of all item scores) × 2, range 10-50
- Norm reference: >40 indicates high neuroticism level

Input to be processed
<<input>>

Figure 18: Prompt for Refine Scale.

Prompt for Sampling Strategy Optimization

Role
You are a Psychometric Data Scientist. Specialized in sampling design and statistical power analysis to ensure sample representativeness.

Task Description
Design the next-round data collection plan to address:
1.Demographic biases in previous samples (e.g., gender ratio 7:3)
2.Statistical power requirements (α=0.05, power=0.8)
3.Stratified sampling proportions (age/gender/education level)

Approach Guidelines
Sample size calculation formula:
$n = (Z_{\alpha/2} + Z_\beta)^2 \times p(1-p) / \delta^2$
Where:
- p = Expected effect size
- δ = Allowable error

Stratification strategy:
1. Calculate the population proportion for each stratum
2. Set a minimum n=30 per stratum
3. Double sampling for high-sensitivity dimensions (e.g., clinical scales)

Sample for Reference
Input:
```
{
    "previous_sample": {
          "gender": {"male":70%, "female":30%},
          "age": {"18-25":5%, "26-35":15%}
     },
     "target_population": {
          "gender": {"male":50%, "female":50%},
          "age": {"18-25":20%, "26-35":30%}
     },
     "parameters": { "confidence": 95%, "margin_of_error": 3% }
}
```

Output:
```
{
     "sampling_plan": {
          "target_n": 450,
          "stratification": [
               {
                    "stratum": "gender",
                    "quotas": [
                              {"group": "male", "percentage": 50%, "min_n":225},
                              {"group": "female", "percentage":50%, "min_n":225}
                     ]
                },
                {
                    "stratum": "age",
                    "quotas": [
                              {"group": "18-25", "percentage":20%, "min_n":90},
                              {"group": "26-35", "percentage":30%, "min_n":135}
                     ]
                }
          ],
      "power_analysis": { "effect_size": 0.3, "achieved_power": 0.82 }
      }
}
```

Input to be processed
<<input>>

Figure 19: Prompt for Sampling Strategy Optimization.