ConceptPsy: A Benchmark Suite with Conceptual Comprehensiveness in Psychology

Anonymous ACL submission

Abstract

The effective incorporation of Large Language Models (LLMs) into the field of psychology necessitates a comprehensive domain benchmark to guide their development and adaptation. Existing Chinese benchmarks in the style of MMLU, such as CMMLU, do include psychology subjects, but their concept coverage is far from exhaustive. The number of questions in each domain is just in the hundreds, and an uneven question sampling process can lead to a "concept bias" issue. This bias, stemming from using a question set with a low concept coverage rate to represent a subject, can potentially lead to skewed results. To address this, we present ConceptPsy, a Chinese conceptual benchmark specifically designed for evaluating LLMs' complex reasoning and knowledge in psychology. ConceptPsy encompasses 12 core subjects and 1,383 concepts from official exams. To avoid copyright issues, we prompt `GPT-4` to generate questions for each of the concepts, which are then validated by psychology professionals to ensure high quality. Besides the overall scores, we annotate each question with a chapter label to provide fine-grained results. We evaluate a range of LLMs on ConceptPsy and the results show significant performance differences across psychology concepts, even among models from the same series. We anticipate the comprehensive concept coverage and the fine-grained strengths and weaknesses identified by ConceptPsy can facilitate the development and growth of the Chinese psychology domain.

1 Introduction

Domain-specific benchmarks are essential for the advancement and adaptation of LLMs. Recently, there has been a significant advancement in LLMs, exemplified by models like `GPT-4` (Achiam et al., 2023), `Claude-3` (Anthropic, 2023), `Qwen2` (Bai et al., 2023a), `DeepSeek-V2` (DeepSeek-AI, 2024). These models have shown remarkable abilities in a wide array of standard tasks. However, a critical gap remains in the systematic understanding of these models' performances within the important domain of psychology. This gap primarily arises from the absence of a comprehensive psychological benchmark.

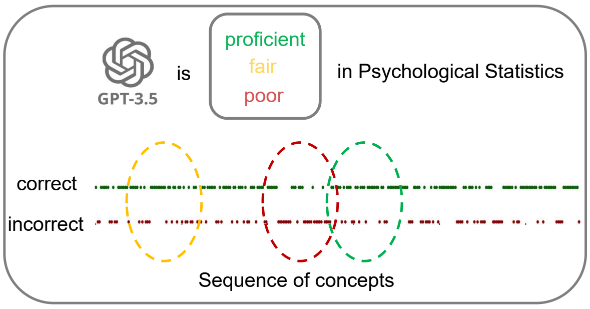


Figure 1: `GPT-3.5-Turbo`'s concept-wise performance on Psychological Statistics. The x-axis represents the sequence of concepts, arranged in the order they appear in the textbook. The dashed circles represent sampled question sets. Different samplings can mislead people's understanding of a model.

While some Chinese MMLU-style benchmarks such as CMMLU (Li et al., 2023) have offered a wide range of subjects for assessment, even including psychology. They tend to focus more on covering many subjects rather than diving deep into each one. This is reflected by the limited number of questions per subject and the lack of emphasis on concept coverage during the sampling process. As a result, these questions do not thoroughly cover the necessary concepts within each subject in psychology. We term this problem as "concept bias" and discuss more in §2.1. Relying on question sets with limited conceptual breadth to represent psychology can lead to misleading results. Figure 1 illustrates the performance variability of models across concepts, with each question set distinguished by a unique color. These variations highlight that models can exhibit a range of proficiency—from poor

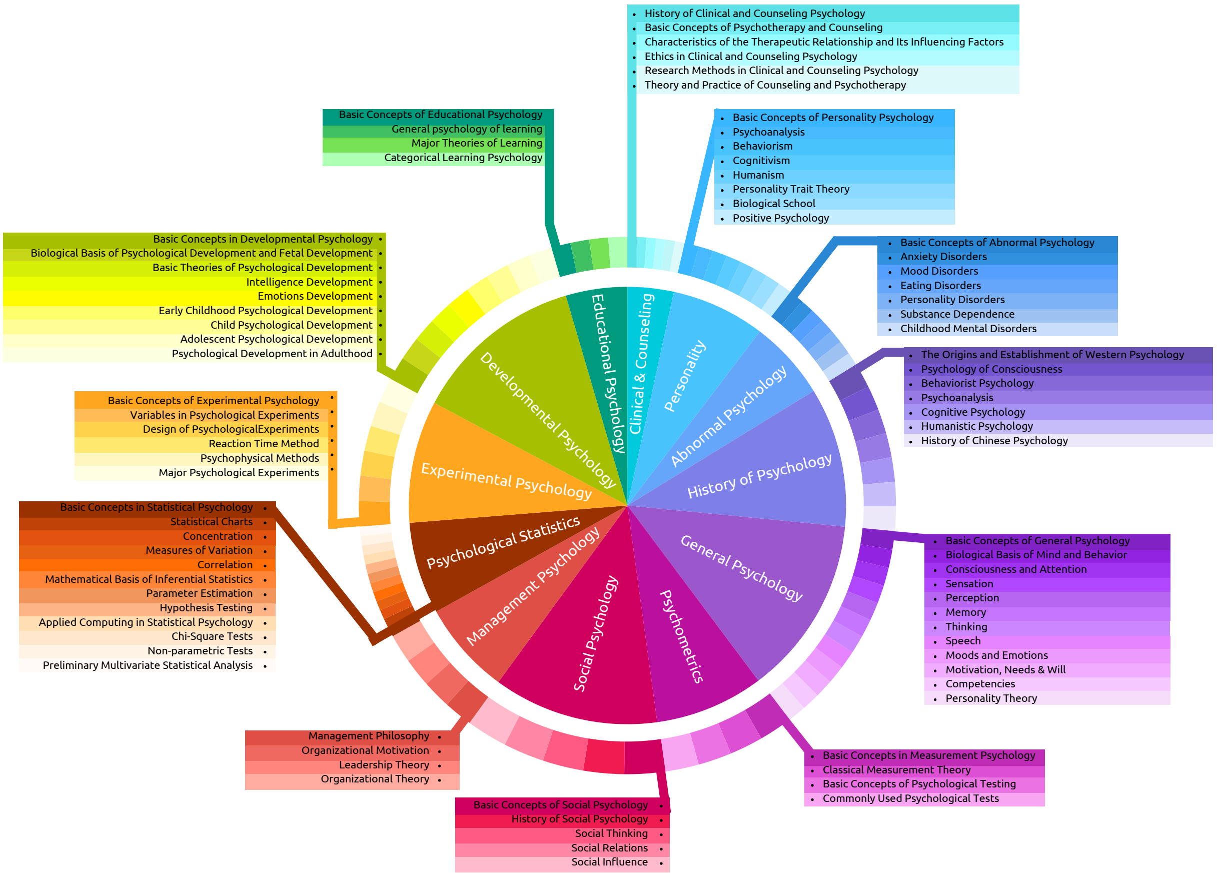

History of Clinical and Counseling Psychology
Basic Concepts of Psychotherapy and Counseling
Characteristics of the Therapeutic Relationship and Its Influencing Factors
Ethics in Clinical and Counseling Psychology
Research Methods in Clinical and Counseling Psychology
Theory and Practice of Counseling and Psychotherapy
Basic Concepts of Educational Psychology
General psychology of learning
Major Theories of Learning
Categorical Learning Psychology
Basic Concepts of Personality Psychology
Psychoanalysis
Behaviorism
Cognitivism
Humanism
Personality Trait Theory
Biological School
Positive Psychology
Basic Concepts in Developmental Psychology
Biological Basis of Psychological Development and Fetal Development
Basic Theories of Psychological Development
Intelligence Development
Emotions Development
Early Childhood Psychological Development
Child Psychological Development
Adolescent Psychological Development
Psychological Development in Adulthood
Basic Concepts of Abnormal Psychology
Anxiety Disorders
Mood Disorders
Eating Disorders
Personality Disorders
Substance Dependence
Childhood Mental Disorders
The Origins and Establishment of Western Psychology
Psychology of Consciousness
Behaviorist Psychology
Psychoanalysis
Cognitive Psychology
Humanistic Psychology
History of Chinese Psychology
Basic Concepts of Experimental Psychology
Variables in Psychological Experiments
Design of PsychologicalExperiments
Reaction Time Method
Psychophysical Methods
Major Psychological Experiments
Educational Psychology
Clinical & Counseling
Personality
Abnormal Psychology
History of Psychology
Developmental Psychology
Experimental Psychology
Psychological Statistics
General Psychology
Management Psychology
Social Psychology
Psychometrics
Basic Concepts in Statistical Psychology
Statistical Charts
Concentration
Measures of Variation
Correlation
Mathematical Basis of Inferential Statistics
Parameter Estimation
Hypothesis Testing
Applied Computing in Statistical Psychology
Chi-Square Tests
Non-parametric Tests
Preliminary Multivariate Statistical Analysis
Basic Concepts of General Psychology
Biological Basis of Mind and Behavior
Consciousness and Attention
Sensation
Perception
Memory
Thinking
Speech
Moods and Emotions
Motivation, Needs & Will
Competencies
Personality Theory
Management Philosophy
Organizational Motivation
Leadership Theory
Organizational Theory
Basic Concepts in Measurement Psychology
Classical Measurement Theory
Basic Concepts of Psychological Testing
Commonly Used Psychological Tests
Basic Concepts of Social Psychology
History of Social Psychology
Social Thinking
Social Relations
Social Influence

Figure 2: Diagram overview of concepts in ConceptPsy. We sample questions based on the requirement of the National Post-graduate Entrance Examination in China. Each question is tagged with a modified chapter name, serving as the chapter-level concept, to further provide chapter-level accuracy.

to fair to proficient—based on the specific questions used for evaluation. As a result, it might mislead developers about how well a model can handle psychology-related topics.

Domain benchmarks are key to advancing LLMs and have been introduced in fields like (Fei et al., 2023), Finance (Zhang et al., 2023), Mathematics (Wu et al., 2024), Medicine (Chen et al., 2024). However, a comprehensive Chinese benchmark for evaluating advanced knowledge and reasoning abilities in psychology is missing. To narrow the gap, we introduce ConceptPsy, the first comprehensive Chinese evaluation suite designed to thoroughly assess LLMs' psychology knowledge and reasoning abilities. To ensure the comprehensiveness of concepts, we manually gathered 1383 college-level concepts based on the official requirements of National Post-graduate Entrance Examination. To avoid copyright issues, we design prompts inspired by the Chinese Academy of Sciences Psychological qualification exam. We prompt GPT-4 to generate four multiple-choice questions per concept, which are reviewed by professional psychologists. We also assigned a chapter-level label to each question, providing fine-grained performance analysis to facilitate model improvements.

We conducted experiments evaluating a wide array of LLMs. Results show GPT-4 surpassing human performance with an average accuracy of 84%. Chinese models like `Yi-34B-Chat` and `Qwen1.5-72B-Chat` also scored highly, though disparities in performance across chapters reached up to 50%. Even within the same series, large-scale models displayed significant chapter-specific performance variations. We believe ConceptPsy will help developers better comprehend their models' psychology abilities and foster the development of foundational psychology models for Chinese users.

2 Case Study of Concept Bias

We conduct a case study to analyze the concept bias in popular Chinese MMLU benchmarks C-EVAL (Huang et al., 2023) and CMMLU (Li et al., 2023). We define a concept as a fundamental unit

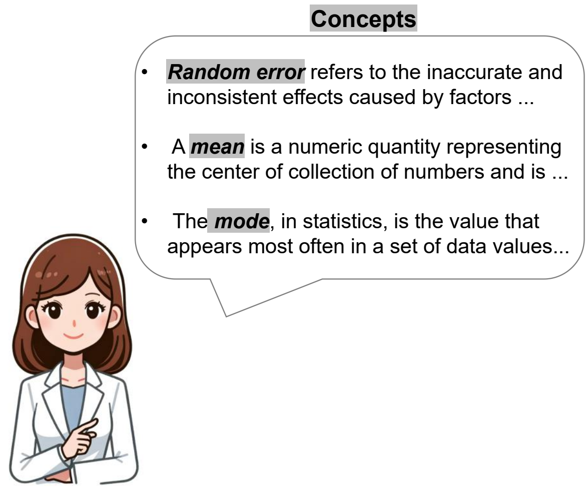

Concepts
• Random error refers to the inaccurate and inconsistent effects caused by factors ...
• A mean is a numeric quantity representing the center of collection of numbers and is ...
• The mode, in statistics, is the value that appears most often in a set of data values...

Figure 3: Examples of concepts. We define a "concept" as fundamental units of understanding that encapsulate specific knowledge within a broader field of study.

Subject	Coverage Rate
STEM	
Computer Architecture	0.55
Computer Network	0.31
High School Biology	0.52
Social Science	
High School Geography	0.48
Marxism	0.70
Humanity	
High School History	0.76
Logic	0.47
Avg	0.54

Table 1: The concept coverage rate of randomly sampled subjects, with chapter-level concepts generated by GPT-4, as evaluated on C-EVAL.

Subject	Benchmark	#Questions	#Concepts	Coverage_rate
Professional Psychology	C-EVAL	-	-	-
	CMMLU	232	84 (chapter-level)	0.59
Advanced Math	C-EVAL	173	94	0.54
	CMMLU	104	36	0.35

Table 2: The concept coverage rate of subjects with manually collected required concepts. We prompt GPT-4 to classify each question into one or more concepts and subsequently calculate the coverage rate.

in the human learning process. For example, in Figure 3, the "random error" is a concept when studying "random variables". We study concept bias from two perspectives: 1. concept coverage rate (§2.1): the proportion of concepts tested in the sampled questions to the total concepts required. 2. performance variance (§2.2): the difference in model performance across various concepts.

Subject	Benchmark	min	max	mean	std
Professional Psychology	C-EVAL	-	-	-	-
	CMMLU	0.0	1.0	0.58	0.34
Advanced Math	C-EVAL	0.0	0.80	0.35	0.24
	CMMLU	0.0	0.75	0.41	0.18

Table 3: Performance of GPT-3.5-Turbo on subjects with manually collected required concepts across different chapter-level concepts.

2.1 Analysis on the Concept Coverage

In our research, we introduce the term "concept coverage rate" to quantify the extent to which the tested concepts in a given question set align with the required concepts in a given subject. Specifically, we calculate it as the ratio of the number of tested concepts, denoted as $C_{\text{collection}}$, to the number of required concepts, denoted as $C_{\text{requirement}}$: $(Concept_{coverage\_rate} = \frac{C_{collection}}{C_{requirement}}$. C-EVAL excludes subjects related to psychology. Within CMMLU, a subject termed "professional psychology" contains 232 questions, and the average score is utilized to gauge a model's proficiency in psychology. Our analysis delves into the concept coverage rate specific to psychology within CMMLU. Additionally, we conduct sampling across diverse disciplines to investigate the potential presence of concept bias in other subjects.

Setup We first assess the concept coverage of "professional psychology" in C-MMLU. Utilizing National Post-graduate Entrance Examination, we manually gather all necessary college-level concepts (1383 in total). We then employ GPT-4 to assign each question to one or more concepts. To investigate concept bias in other subjects, we sample subjects from various disciplines (STEM, Social Science, Humanities) in C-EVAL. Due to the resource limitation, we only manually collect the required concepts for advanced math. For other subjects, we use GPT-4 to generate chapter names representing the required concepts. Each question is also categorized into chapters using GPT-4. More experimental details can be found in Appendix E.

The concept coverage rate is Low. Although psychology is a vital domain for AI to learn human behavior, it is absent in C-EVAL, and its chapter-level concept coverage rate in CMMLU is only 59%. Questions with such an extremely low concept coverage rate could potentially mislead developers. This low coverage rate, although partly at-

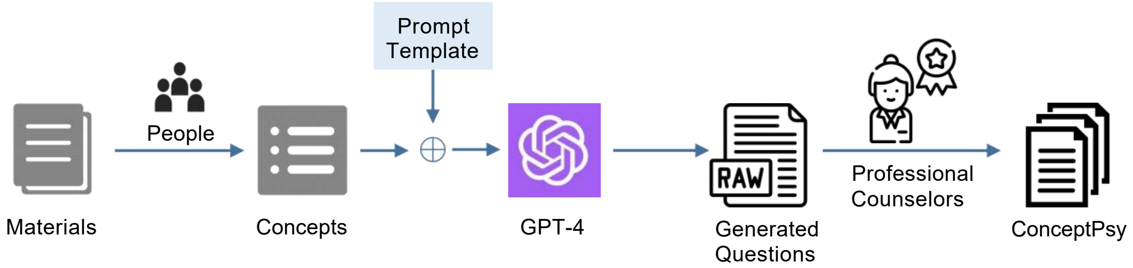


Figure 4: Overview of Our Concept-Driven Framework. We collect relevant concepts based on the requirements of corresponding examinations. To diversify the types of questions, we summarize three question patterns from these exams and design specific prompts for each type. Questions are then generated using `GPT-4`. Subsequently, we hire professional psychological counselors to review the questions for accuracy and relevance.

Subject	Baichuan2-13B	Qwen1.5-MoE-A2.7B
Computer Architecture	0.50±0.17	0.72±0.14
Computer Network	0.59±0.19	0.80±0.17
High School Biology	0.52±0.25	0.68±0.32
High School Geography	0.73±0.16	0.84±0.19
Marxism	0.88±0.10	0.96±0.03
High School History	0.59±0.33	0.80±0.29
Logic	0.46±0.36	0.60±0.34

Table 4: The mean and standard deviation of model performance on randomly sampled C-EVAL subjects across different chapters.

tributable to the broad spectrum of subjects within psychology, is not exclusive to this field. Even in more focused domains like advanced mathematics, the coverage rate is disappointingly low. Other subjects in C-EVAL (as shown in Table 1), such as High School History and Marxism, fare slightly better with coverage rates around 70%. However, the average coverage rate across all subjects is only 54%. This figure might fall short of user expectations, considering these questions are used as a representation of a subject.

2.2 Variations in Performance Across Different Concepts

A high variance in scores across concepts can make the average accuracy more misleading. To explore this, we measure the variance of different concepts within a subject.

Setup Following the method in section §2.1, we categorize questions into chapter-level concepts and calculate the chapter-level performance variance.

The performance across different chapters varies greatly. As demonstrated in Table 3 and 4, powerful models like GPT-3.5-Turbo (OpenAI, 2022) exhibit a standard deviation exceeding 10% across different chapters. We also evaluate open-source models such as Baichuan2-13B (Yang et al., 2023a) and Qwen1.5-MoE-A2.7B (Bai et al., 2023b). These models persist in showing a standard deviation of over 10% on subjects sampled from diverse disciplines. This high variance indicates that the final accuracy of these models is greatly influenced by the question set sampled.

3 ConceptPsy

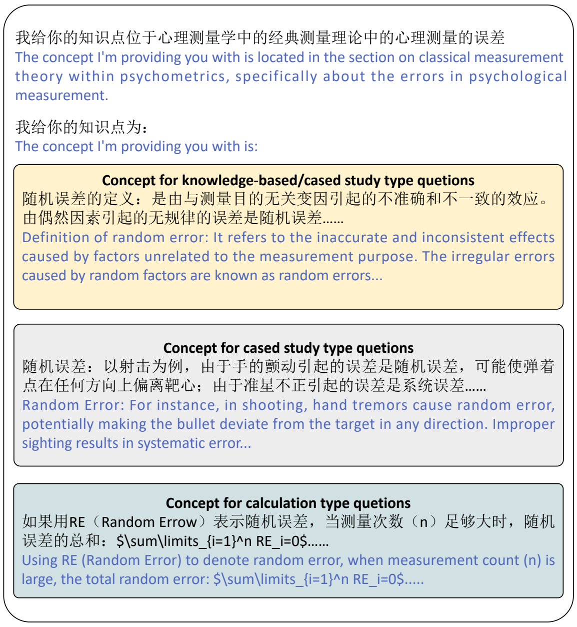
Our proposed ConceptPsy seeks to fill the gap by offering a complete benchmark for evaluating the knowledge and reasoning ability of models in psychology. To the best of our knowledge, we are the first to explain the issue of concept bias in Chinese MMLU-style benchmarks and mitigate this problem by constructing question sets that cover necessary concepts.

3.1 Overview

In ConceptPsy, we manually summarize 1383 concepts from the National Post-graduate Entrance Examination, then prompting `GPT-4` to generate questions for each to ensure comprehensive coverage of all key concepts. This method sets ConceptPsy apart from prior Chinese MMLU benchmarks like CMMLU and C-EVAL, which relied on questions from online sources or textbooks. Our choice to use `GPT-4` for question generation was driven by two main reasons: 1) To avoid copyright issues associated with using existing questions, facilitating broader industrial application; 2) Many concepts lacked associated multiple-choice questions, posing a challenge for creating new ones. To maintain question quality, we develop specific

Category	# C	# Q	L_Q	L_A
In terms of subject				
Clinical & Counseling Psychology	56	156	38.9	12.0
Psychology of Personality	91	318	36.7	11.1
Abnormal Psychology	89	268	35.8	12.4
History of Psychology	126	472	27.4	10.5
General Psychology	183	605	35.3	9.4
Psychometrics	115	368	43.2	10.1
Social Psychology	169	559	26.2	13.7
Management Psychology	88	315	37.2	10.4
Psychological Statistics	99	311	57.0	8.4
Experimental Psychology	141	413	59.3	9.3
Developmental Psychology	159	580	30.7	11.3
Educational Psychology	67	208	42.9	11.5
In terms of split				
Dev	-	60	-	-
Valid	-	428	-	-
Test	-	4085	-	-
Total	1383	4573	-	-

Table 5: Statistics of ConceptPsy. The column "#C" indicates the number of concepts we have annotated for each subject, with each concepts generating 4 questions and filtered by professional psychological annotators. The number of questions obtained after the review process is displayed in the column "#Q". L_Q and L_A is the average length of a question and answer separately.



我给你的知识点位于心理测量学中的经典测量理论中的心理测量的误差
The concept I'm providing you with is located in the section on classical measurement theory within psychometrics, specifically about the errors in psychological measurement.

我给你的知识点为：
The concept I'm providing you with is:

Concept for knowledge-based/cased study type quetions
随机误差的定义：是由与测量目的无关变因引起的不准确和不一致的效应。由偶然因素引起的无规律的误差是随机误差......
Definition of random error: It refers to the inaccurate and inconsistent effects caused by factors unrelated to the measurement purpose. The irregular errors caused by random factors are known as random errors...

Concept for cased study type quetions
随机误差：以射击为例，由于手的颤动引起的误差是随机误差，可能使弹着点在任何方向上偏离靶心；由于准星不正引起的误差是系统误差......
Random Error: For instance, in shooting, hand tremors cause random error, potentially making the bullet deviate from the target in any direction. Improper sighting results in systematic error...

Concept for calculation type quetions
如果用RE（Random Errow）表示随机误差，当测量次数（n）足够大时，随机误差的总和：$\sum\limits_{i=1}^n RE_i=0$......
Using RE (Random Error) to denote random error, when measurement count (n) is large, the total random error: $\sum\limits_{i=1}^n RE_i=0$.....

Figure 5: An example of an annotator assigning a suitable prompt to a concept. For the concept "random error", we collect multiple descriptions. The appropriate prompt is assigned based on the type of description provided.

prompts based on the Professional Counselor Examination formats and had them reviewed by three professional psychologists. ConceptPsy differs from CMMLU and C-EVAL in the following ways:

- **Low Concept Bias:** ConceptPsy covers all required concepts to provide a low-biased results for users.
- **Fine-grained Score:** Instead of only providing an average score, we also provide chapter-level scores to reflect fine-grained strengths and weaknesses for developers.
- **An in-depth benchmark:**Instead of aiming to cover as many subjects as possible, we focus on providing the comprehensive and high in-depth coverage within the field of psychology, encompassing all college-level subjects.

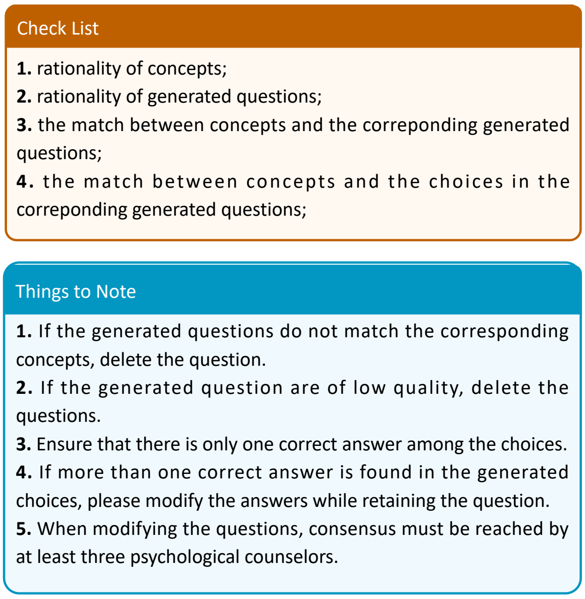
3.2 Data Collection

Subject selection. As shown in Figure 2 and Table 5, We select 12 core standard subject within the discipline of psychology meticulously considered in accordance with both higher education standards and professional qualification requirements. We select 11 courses from the Peking University Core Curriculum Handbook for Undergraduate Program [1]. Additionally, we included Psychological Counselling as a fundamental subject within our data set to adhere to professional qualification criteria.

Concepts collection. We manually collect concepts based on the requirements of the 12 subjects outlined in National Post-graduate Entrance Examination. This task is undertaken by eight graduate students over the course of one month. Each concept is summarized from the tutorials of the exams. An example of the collected concepts is shown in Figure 5. For each concept, we summarize its description and manually determine which type of prompt (Calculation, Theory Understanding, Case Study) is suitable for generating questions about that concept (Prompts can be found in Appendix I). To enhance GPT-4's understanding of the provided concepts, we supply not only the concepts themselves but also the name of the subject, and the primary and secondary headings they fall under.

Prompts design. In the process of designing prompts to generate high-quality questions, we conduct a thorough analysis of question types found in the Professional Counselor Examination, categorizing them into three distinct groups: (1) calculation;

[1] http://www.dean.pku.edu.cn/web/student_info.php?type=1&id=2



Check List

1. rationality of concepts;
2. rationality of generated questions;
3. the match between concepts and the correponding generated questions;
4. the match between concepts and the choices in the correponding generated questions;

Things to Note

1. If the generated questions do not match the corresponding concepts, delete the question.
2. If the generated question are of low quality, delete the questions.
3. Ensure that there is only one correct answer among the choices.
4. If more than one correct answer is found in the generated choices, please modify the answers while retaining the question.
5. When modifying the questions, consensus must be reached by at least three psychological counselors.

Figure 6: The review rules the question review process are as follows. Professional psychological annotators will filter, modify, and review the generated questions based on these requirements.

(2) theory understanding; (3) case study (Appendix I). As depicted in Figure 5, while collecting concepts, we manually assign one or more prompt types to each concept. This assignment strategy ensures the relevance and diversity of the generated questions. To maintain a relatively even number of questions for each concept, we prompt GPT-4 to generate four questions per concept.

3.3 Questions Review

To ensure the quality of our dataset, we employ three professional psychological annotators to meticulously review each question. Our main objective is to verify the accuracy of the questions and to remove any that are found to be unreasonable. Each question is subjected to a thorough evaluation by at least one professional psychologist, who identifies any aspects that are not reasonable. When inaccuracies are discovered, all professional psychologists collaborate to discuss and agree upon a suitable and precise correction strategy. The criteria for reviewing questions are detailed in Figure 6. Throughout this process, the annotators check how relevant and accurate the knowledge points, the questions, and their connections are.

4 Experiments

4.1 Dataset Statistics

We carefully select 12 psychology subjects as shown in Table 5. For each concept, we generate 4 questions, leading to a total of 4573 high-quality questions after a thorough review and filtering process. Furthermore, as illustrated in Figure 2, we collect chapter names based on the syllabus of these subjects and assign chapter-level labels to each generated question to provide detailed insights about the strengths and weaknesses.

4.2 Setup

We evaluate a diverse set of strong models to ensure a comprehensive evaluation. More details about the models can be found in Appendix H. We conduct the evaluations using a 5-shot method. We set the temperature as 0, *top_p* as 1.0.

4.3 Results

The main results of different LLMs on ConceptPsy are shown in Table 6. We also provide fine-grained results at Figure 7, 8 and 9.

Strong Chinese models achieve comparable overall results with GPT-4: In Table 6, surprisingly, some Chinese models achieve comparable performance close to that of GPT-4 on ConceptPsy. Note that GPT-4 have the concepts as input during the question generation process but not in the evaluation stage. Overall, GPT-4 performed significantly better in subjects requiring more reasoning ability, such as Psychological Statistics (79% compared to 69%). However, in subjects requiring a deep understanding of the Chinese context and knowledge, such as Clinical & Counseling, Yi-34B-Chat performed better. This reflects the differences in abilities between the two models, indicating that GPT-4 might have insufficient training data in some areas of Chinese psychology, but its reasoning ability is much stronger.

Chinese models struggle with math and reasoning: Analysis from Figure 7, 8, and 9 reveals that Chinese models' performance drops on concepts requiring math and reasoning, like "Mathematical Basis of Inferential Statistics" and "Non-Parametric Tests". While these models excel in knowledge-based areas, they falter in math and reasoning segments. Mixtral-8x7B-Ins stands out by underperforming in knowledge questions but outshining others in math and reasoning. This suggests that despite high accuracy from training on

	Clinical & Counseling	Psy of Personality	Abnormal Psy	History of Psy	General Psy	Psy-chometrics	Social Psy	Management Psy	Psychological Statistics	Experimental Psy	Developmental Psy	Educational Psy	Avg
GPT-4	0.83	**0.86**	0.88	**0.81**	0.84	**0.80**	**0.89**	0.87	**0.79**	**0.85**	0.82	**0.86**	0.84
GPT-3.5-Turbo	0.78	0.74	0.80	0.67	0.73	0.70	0.82	0.80	0.64	0.74	0.77	0.81	0.75
Chi-Alpaca2-7B	0.59	0.58	0.59	0.52	0.45	0.51	0.63	0.62	0.46	0.58	0.59	0.67	0.57
Llama-2-13B-Chat	0.66	0.62	0.65	0.56	0.54	0.57	0.66	0.63	0.48	0.57	0.58	0.66	0.60
Chatglm2-6B	0.71	0.63	0.68	0.63	0.62	0.62	0.69	0.64	0.54	0.64	0.66	0.72	0.65
Llama-2-70B-Chat	0.74	0.66	0.73	0.62	0.61	0.59	0.75	0.70	0.52	0.63	0.69	0.69	0.66
Mistral-7B-Ins	0.73	0.65	0.66	0.59	0.62	0.63	0.71	0.72	0.52	0.65	0.66	0.74	0.66
Baichuan2-7B-Chat	0.72	0.69	0.77	0.65	0.67	0.66	0.75	0.72	0.53	0.66	0.70	0.76	0.69
Baichuan2-13B-Chat	0.78	0.73	0.80	0.68	0.73	0.69	0.82	0.75	0.60	0.71	0.74	0.83	0.74
Mixtral-8x7B-Ins	0.82	0.76	0.78	0.65	0.73	0.70	0.83	0.78	0.71	0.73	0.74	0.79	0.75
Qwen1.5-7B-Chat	0.83	0.75	0.84	0.66	0.75	0.67	0.79	0.81	0.61	0.72	0.73	0.83	0.75
Internlm2-7B-Chat	0.86	0.75	0.84	0.68	0.78	0.7	0.82	0.82	0.62	0.75	0.78	0.81	0.77
Yi-6B-Chat	0.85	0.78	0.89	0.71	0.84	0.70	0.85	0.86	0.62	0.72	0.78	0.83	0.79
Qwen1.5-72B-Chat	0.87	0.74	0.89	0.80	0.81	0.77	0.88	0.87	0.38	0.77	0.83	0.85	0.79
Qwen1.5-14B-Chat	0.86	0.80	0.85	0.72	0.83	0.72	0.83	0.85	0.67	0.81	0.79	0.85	0.80
Yi-34B-Chat	**0.92**	0.84	**0.91**	**0.81**	**0.86**	0.78	0.88	**0.90**	0.69	0.84	**0.86**	**0.86**	**0.85**

Table 6: Performances on ConceptPsy with different LLMs.

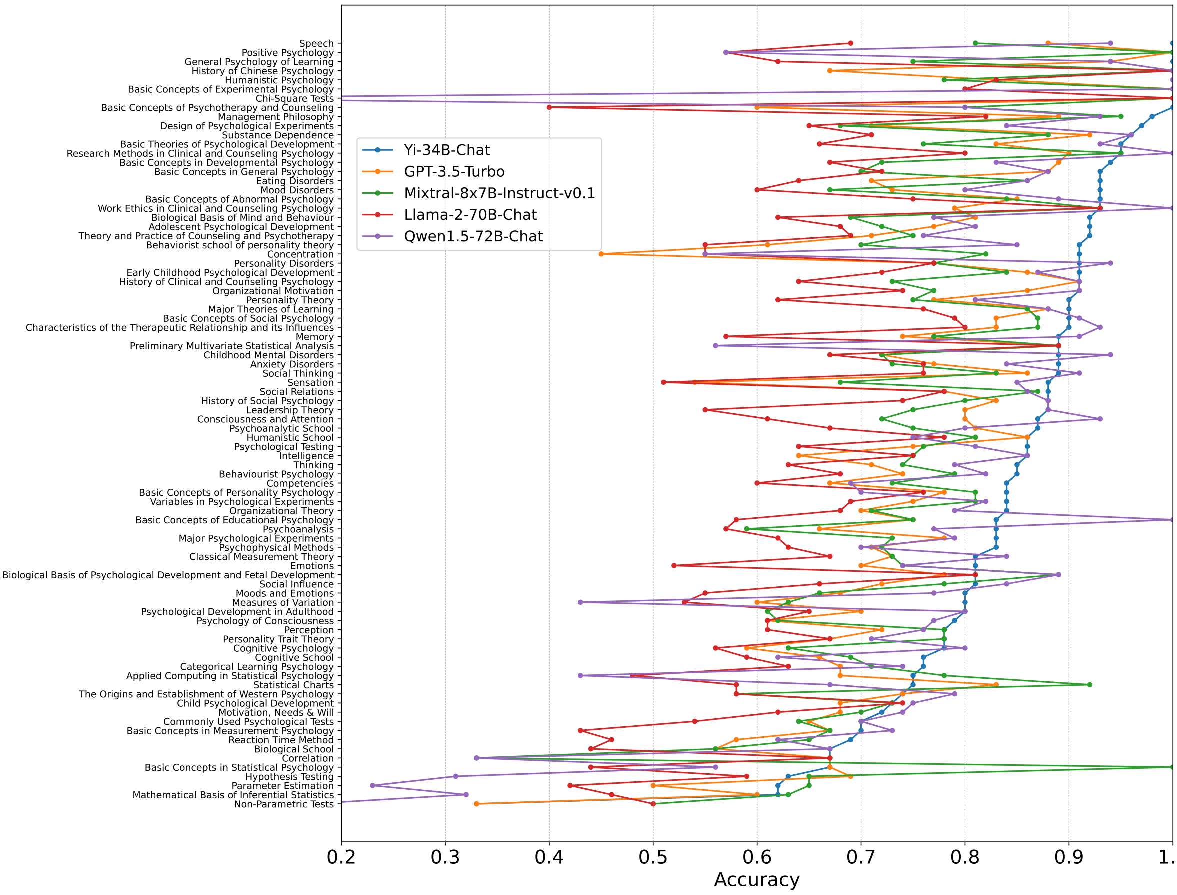


Figure 7: Concept-level results for models more than 34B, in addition to GPT-3.5-Turbo.

extensive Chinese texts, Chinese models' math and reasoning skills, especially in psychology, need enhancement.

Models show varied significant chapter performance: There's notable variation in how models perform across chapters. Yi-34B-Chat, for example, despite leading in average accuracy, exhibits over 50% performance differences between chapters—excelling in basic concepts but lagging in math reasoning, even behind Mixtral-8x7B-Ins with its 10% lower average accuracy. This underscores ConceptPsy's role in offering detailed insights for model enhancement.

Variability in conceptual mastery across models from the same series: Although Qwen1.5-14B-Chat and Qwen1.5-72B-Chat achieve similar overall results, this does not mean their performance is similar across all chapters. As shown in Figures 7 and 8, Qwen1.5-72B-Chat performs much worse in chapters that focus on

Subjects	Pro	Non-Pro
Clinical & Counseling Psychology	0.85	0.28
Psychology of Personality	0.82	0.24
Abnormal Psychology	0.89	0.32
History of Psychology	0.73	0.22
General Psychology	0.9	0.26
Psychometrics	0.75	0.3
Social Psychology	0.82	0.16
Management Psychology	0.8	0.36
Psychological Statistic	0.78	0.2
Experimental Psychology	0.82	0.32
Developmental Psychology	0.83	0.16
Educational Psychology	0.82	0.16
Avg	0.82	0.25

Table 7: Human evaluation results of ConceptPsy. "Pro" and "Non-pro" represent Professional Counselor and Non-Professional Student respectively.

math and reasoning, like "Non-Parametric Tests" and "Parameter Estimation." While the overall accuracy suggests that both models perform similarly, the detailed results from ConceptPsy show that `Qwen1.5-72B-Chat` is generally better than `Qwen1.5-14B-Chat` in most chapters, except for a few math-related ones where it falls behind. This demonstrates how ConceptPsy can highlight specific strengths and weaknesses, helping developers make targeted improvements.

4.4 Ablation Study

Human baseline. To assess ConceptPsy's ability to distinguish between human expertise levels, we ask professional counselors and non-professional grad students to evaluate a subset of tasks. Table 7 shows significant performance differences (Appendix A).

The effect of linguist style. To investigate the effect of different linguistic styles of generated questions, we paraphrase each question using `Qwen-70b-Chat` evaluate LLMs on it. The results are shown in Table 8. There is a slight performance drop on paraphrased questions, potentially due to the lack of manual review for these paraphrased versions. Nonetheless, the consistency in model rankings and overall performance suggests that for multiple-choice questions, stylistic variations of the questions have limited effect.

5 Related Works

Chinese MMLU Benchmarks: While English language benchmarks continue to evolve (Hendrycks et al., 2020; Huang et al., 2023; Li et al., 2023; Zhong et al., 2023), Chinese MMLU benchmarks remain underdeveloped. The CLUE benchmark (Xu et al., 2020) is a widely used large-scale NLU benchmark for Chinese. AGIEval (Zhong et al., 2023) expands this with questions from various Chinese exams, and MMCU (Zeng, 2023) includes questions from diverse domains. The C-EVAL benchmark (Huang et al., 2023) gathers questions from different educational levels and professional exams, using non-paper-based and simulation questions to prevent dataset leakage. ConceptMath (Wu et al., 2024) evaluates LLMs in a concept-wise manner but focuses solely on mathematics, is not a comprehensive MMLU benchmark, and was released more than three months after our work.

Benchmarks for LLMs in Psychology. Although psychology is a crucial domain for achieving artificial intelligence, there are currently few benchmarks in this field. CMMLU has collected hundreds of questions as "professional psychology" to evaluate a model's knowledge understanding and reasoning capabilities. Other works (Yang et al., 2023b; Amin et al., 2023; Lamichhane, 2023) approach this from a mental health perspective, utilizing various classification tasks, including various emotions, suicidal tendencies, and more. PsyEval (Jin et al., 2023) further expands in the mental health field by adding tasks such as "Diagnosis Prediction", but lacking comprehensive subjects in college-level psychology.

6 Discussion

Domain benchmarks are crucial for adapting LLMs. However, a comprehensive benchmark for assessing models' knowledge and reasoning abilities in psychology is lacking. To address this gap, we introduce ConceptPsy. In designing ConceptPsy, we reveal concept bias issues in previous Chinese MMLU benchmarks and study concept coverage rates and performance variances. To tackle these issues, we propose a concept-wise question generation framework and provide chapter-level fine-grained results. Evaluation across various models demonstrates ConceptPsy's effectiveness in highlighting model strengths and weaknesses.

Limitations

We focused exclusively on concept bias within the Chinese MMLU dataset series, not exploring biases in other dimensions due to the labor-intensive nature of collecting the required concepts in a discipline, which necessitates trained annotators. Our experiments were thus confined to the field of psychology. Although our question generation method could be easily applied to other domains, due to copyright issues and the absence of corresponding multi-choice questions for many concepts, we opted to use GPT-4 for question generation instead of relying on human-designed questions.

Ethics Statement

We have thoroughly examined our data to ensure that there are no ethical issues. The data is generated by GPT-4, using concepts prescribed by the National Entrance Examination for Postgraduates. Furthermore, the generated data is subjected to rigorous scrutiny by professional psychologists to ensure its ethical soundness.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Mostafa M Amin, Erik Cambria, and B Schuller. 2023. Will affective computing emerge from foundation models and general ai. *A first evaluation on ChatGPT. ArXiv, abs/2303.03186*.

Anthropic. 2023. Introducing claude.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023b. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt.

2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Haoan Jin, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2023. Psyeval: A comprehensive large language model evaluation benchmark for mental health. *arXiv preprint arXiv:2311.09189*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

OpenAI. 2022. Introducing chatgpt.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yanan Wu, Jie Liu, Xingyuan Bu, Jiaheng Liu, Zhanhui Zhou, Yuanxing Zhang, Chenchen Zhang, Zhiqi Bai, Haibin Chen, Tiezheng Ge, et al. 2024. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models. *arXiv preprint arXiv:2402.14660*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023b. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *arXiv preprint arXiv:2304.12986*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Human Evaluation

We also provide human performance on the subset of ConceptPsy. Specifically, we randomly sample 50 questions from each of 12 subjects, totaling 600 questions. A professional counselor independently completes all 600 questions, while ten non-experts are each assigned different subjects for evaluation. The results are shown in Table 7. The professional counselor significantly outperforms GPT-3.5, achieving an average score of 82%. However, the performance of non-expert undergraduates and graduate students is poor, averaging 25%, which is close to random. This underscores our benchmark's demand for specialized knowledge in psychology and its effectiveness in distinguishing whether a model has mastered these professional concepts.

B The effect of Linguistic Style of generated questions

To investigate the effect of linguistic style, we paraphrase each question using `Qwen-70b-Chat`. We then evaluated the performance difference between its style and `GPT-4`'s style. The results are summarized in the table below:

In comparison to Table 6, models perform slightly worse on paraphrased questions generated by `Qwen-70b-Chat`. This may be because the paraphrased questions are not manually reviewed. However, despite the style changes, the model rankings and overall performance remained consistent. This indicates that for multiple-choice questions, the style of the questions has minimal impact on evaluation.

C Competition with Human-designed Questions

We randomly selected 15-20 questions for each subject and an equivalent number of questions from real exams available online. We shuffle the order of the questions, presenting one generated and one real question side by side, and ask two psychological annotators to judge which was better or if there was a tie. The averaged results are shown in Table 9. The "Win Rate" indicates the proportion of instances where our generated questions were rated better than the real ones. As can be seen, the majority of the generated questions matched or even surpassed the performance of actual exam questions. This success can be attributed to two factors: 1) our prompts are a carefully designed summary of the question types found in Chinese psychological professional examinations; 2) GPT-4's robust capabilities in synthetic data.

D Chapter-level Statistic of ConceptPsy

In Table 10 and Table 11, we present the chapters of each subject along with the corresponding number of concepts and questions.

E Experiments Details of Calculating Concept Coverage Rate

To calculate the concept coverage rate, the challenge lies in obtaining the required concepts in a subject. For advanced math, we first collect its chapters and the concepts under different chapters based on relevant exam requirements. We initially prompt GPT-4 to classify each question into different chapters, then prompt GPT-4 to further classify the question into specific instance-level concepts. We calculate the concept coverage rate and performance variance based on the concepts covered by these questions. This hierarchical classification allows for more accurate categorization.

For psychology in CMMLU, due to the vast number of concepts, prompting GPT-4 with these concepts in the input prompt each time is too expensive, so we use chapter-level concepts. We collect 84 chapter names according to the requirements, and classify each question into one or more chapters.

For other subjects, manually collecting concepts is too costly, and it's impossible to collect concepts for each subject. We prompt GPT-4 to generate a 3-level syllabus for the subject as the required concepts. We then prompt GPT-4 to categorize each question into one or more first-level headings in the syllabi. We further classify them into the second- and third-level headings under the selected first-level headings.

F More Results on Fine-grained Performance

We provide more chapter-level results in this Figure 8, 9.

G Qualifications of Our psychology Annotators

Our psychological annotators is professional. We enlist three psychological annotators with diverse expertise. The first is a postdoctoral researcher

	Clinical & Counseling	Psy of Personality	Abnormal Psy	History of Psy	General Psy	Psy- chometrics	Social Psy	Management Psy	Psychological Statistics	Experimental Psy	Developmental Psy	Educational Psy	Avg
Mistral-7B-Instruct	0.71	0.63	0.64	0.58	0.57	0.59	0.69	0.71	0.53	0.64	0.62	0.69	0.63
gpt-3.5-Turbo	0.80	0.72	0.79	0.67	0.70	0.68	0.77	0.78	0.60	0.73	0.71	0.76	0.73
Yi-6B-Chat	0.80	0.75	0.87	0.67	0.80	0.70	0.83	0.83	0.58	0.69	0.76	0.86	0.76
Yi-34B-Chat	0.86	0.80	0.90	0.78	0.83	0.78	0.87	0.88	0.70	0.82	0.82	0.85	0.82

Table 8: Performances of models on questions paraphrased by `Qwen-70b-Chat`

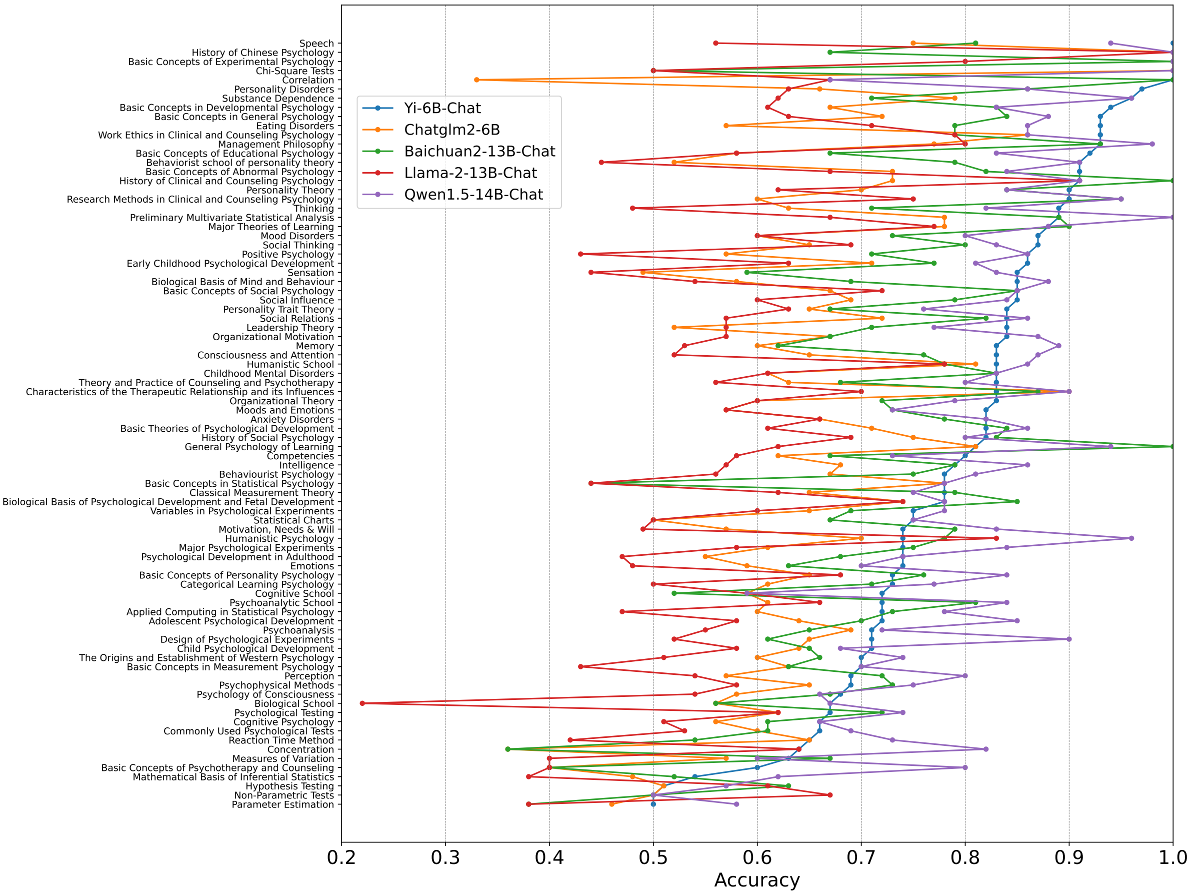


Figure 8: Concept-level results for models wi 14 Billion parameters

specializing in experimental and counseling psychology. The second is a registered psychology researcher with the British Psychological Society with five years of research experience in interdisciplinary psychology laboratories. The third, a counseling psychologist, brings over three years of practical counseling experience.

H Details About Evaluated Models

The evaluated models include `Chinese-Alpaca-2-7B`(Cui et al., 2023), `Chatglm-6B`(Du et al., 2022), `Chatglm2-6B`(Du et al., 2022), `Llama-2-13B-Chat`(Touvron et al., 2023), `Llama-2-70B-Chat`(Touvron et al., 2023), `Baichuan2-7B-Chat`(Yang et al., 2023a), `Baichuan2-13B-Chat`(Yang et al., 2023a), `Mistral-7B-Instruct-v0.2`(Jiang et al., 2023), `Mixtral-8x7B-Instruct-v0.1`(Jiang et al., 2024), `Qwen1.5-7B-Chat`(Bai et al., 2023b), `Qwen1.5-72B-Chat`(Bai et al., 2023b), `Qwen1.5-14B-Chat`(Bai et al., 2023b), `Internlm2-7B-Chat`(Cai et al., 2024), `Yi-6B-Chat`(AI et al., 2024), and `Yi-34B-Chat`(AI et al., 2024). The model codes can be found in Table 12 . We evaluate these models with vLLM (Kwon et al., 2023).

I Prompts for Questions Generation

We totally design four distinct prompts to steer GPT-4 in generating questions based on provided knowledge points. In designing these prompts, we have taken into account several guidelines. Firstly, the generated questions should exhibit a high level of difficulty and complexity. Secondly, while gen-

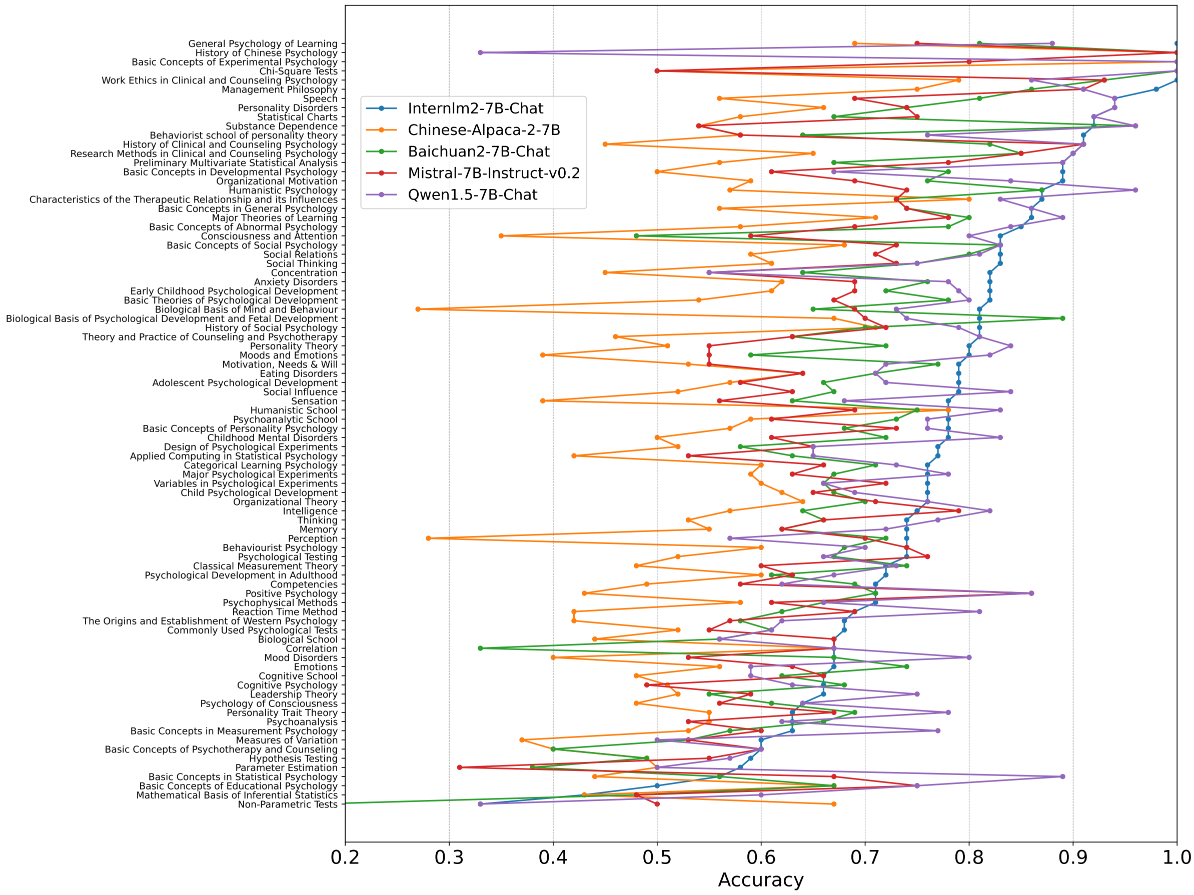


Figure 9: Concept-level results for some other models

Subjects	Win Rate	Tie Rate
Clinical & Counseling Psychology	0	0.8
Psychology of Personality	0	0.9
Abnormal Psychology	0	0.8
History of Psychology	0.1	0.8
General Psychology	0.1	0.9
Psychometrics	0.1	0.7
Social Psychology	0.82	0.16
Management Psychology	0.1	0.8
Psychological Statistic	0.4	0.5
Experimental Psychology	0.2	0.7
Developmental Psychology	0.4	0.4
Educational Psychology	0.3	0.7
Avg	0.27	0.62

Table 9: We hire two professional counselor to compete the quality of generated questions and human-designed questions.

erating questions, GPT-4 should primarily rely on the given knowledge points, but it can also incorporate its inherent psychological knowledge. Thirdly, each knowledge point unit may yield multiple questions, but the content, type, or perspective of the questions should be distinct.

Figures 10, 11, 12, and 13 show the specific prompts that we have inputted into GPT-4. These prompts have been meticulously designed, with each one being tailored to control the generation of a different kind of question. The first three prompts correspond to generating Theory understanding, Case study, and Calculation type questions, respectively, while the last one encompasses all of the aforementioned types. We choose the appropriate prompt for each knowledge point based on its content.

你是一个中国关于[科目]的考试出题人，请根据给定的心理学知识点生成四道综合性和较高难度的单项选择题，题目应根据给定知识点，同时结合您在[科目]的知识，要求考生对知识点有较深入的理解。知识点可能会包含多个内容，四道题目应从不同的内容考察，以全面评估考生的理解程度。题目应具有挑战性，以考核考生是否具备合格的心理咨询师资质。题目应要求考生对知识点进行整合和思考，而非简单地回忆知识点内容。选择题的四个选项只有一个是正确的。

我给你的知识点属于［知识点所在章节］
我给你的知识点为：［知识点］

那么你出的四道综合性且高难度的，正确合理的的选择题是什么？请给出答案和解析。

You are a test question designer for a Chinese **[Subject]** examination. Please generate four multiple-choice questions that are integrative and of high difficulty based on the given psychological concepts. The questions should be based on the given concepts, and at the same time, integrate your knowledge in **[Subject]**. They should require the test takers to have a deep understanding of the concepts. The concepts might encompass multiple contents. The four questions should assess different aspects to fully evaluate the test taker's level of understanding. The questions should be challenging, aimed at examining whether the test taker possesses the qualified credentials of a psychological counselor. The questions should require the test takers to integrate and think about the concepts rather than simply recalling the content. Only one of the four options in the question is the correct answer. Only one of the four options in the multiple-choice question is correct.

The concepts you have provided belong to **[the chapter of concepts]**.
The concepts are as follows: **[concepts]**.

Please provide the correct answers and explanations for the four comprehensive and challenging multiple-choice questions you have formulated.

Figure 10: Theory understanding

你是一个中国关于[科目]的考试出题人，请根据给定的心理学知识点生成四道综合性和较高难度的案例分析的单项选择题。知识点可能会包含多个内容，四道题目应从不同的内容考察。你需要首先为每个单项选择题生成一个知识点相关的真实案例，再根据案例出单项选择题。题目应根据给定知识点，同时结合您在[科目]领域的知识，要求考生对知识点有较深入的理解。题目应具有挑战性，以考核考生是否具备合格的心理咨询师资质。题目应要求考生对知识点进行整合和思考，而非简单地回忆知识点内容。选择题的四个选项只有一个是正确的。

我给你的知识点属于［知识点所在章节］
我给你的知识点为：［知识点］

那么你出的四道综合性且高难度的，正确合理的的选择题是什么？请给出答案和解析。

You are a test question designer for a Chinese **[Subject]** examination. Please generate four comprehensive and high-difficulty single-choice questions for case analysis based on the given psychological concepts. The concepts may contain multiple contents, and the four questions should examine different contents. You need to first generate a real case related to the knowledge point for each single-choice question, then formulate a single-choice question based on the case. The questions should be based on the given concepts, and also integrate your knowledge in the **[Subject]** field, requiring test takers to have a deep understanding of the concepts. The questions should be challenging, aimed at determining whether the test taker has the qualified qualifications of a psychological counselor. The questions should require the test takers to integrate and think about the concepts, rather than merely recalling the content of the concepts. Only one of the four options in the multiple-choice question is correct.

The concepts you have provided belong to **[the chapter of concepts]**.
The concepts are as follows: **[concepts]**.

Please provide the correct answers and explanations for the four comprehensive and challenging multiple-choice questions you have formulated.

Figure 11: Case study

Subject	Chapter	# C	# Q
Clinical & Counseling Psychology	History of Clinical and Counseling Psychology	4	13
	Basic concepts of psychotherapy and counseling	3	7
	Characteristics of the therapeutic relationship and its influences	10	32
	Work Ethics in Clinical and Counseling Psychology	5	16
	Research Methods in Clinical and Counseling Psychology	8	25
	Theory and Practice of Counseling and Psychotherapy	25	63
Psychology of Personality	Basic Concepts of Personality Psychology	12	43
	Psychoanalytic School	23	84
	Behaviorist school of personality theory	10	37
	Cognitive School	9	33
	Humanistic School	11	40
	Personality Trait Theory	18	57
	Biological School	4	13
	Positive Psychology	4	11
Abnormal Psychology	Basic Concepts of Abnormal Psychology	17	60
	Anxiety Disorders	26	81
	Mood Disorders	7	19
	Eating Disorders	6	18
	Personality Disorders	14	40
	Substance Dependence	10	28
	Childhood Mental Disorders	8	22
History of Psychology	The Origins and Establishment of Western Psychology	17	62
	Psychology of Consciousness	40	154
	Behaviourist Psychology	21	82
	Psychoanalysis	25	93
	Cognitive Psychology	13	48
	Humanistic Psychology	8	30
	History of Chinese Psychology	1	3
General Psychology	Basic Concepts in General Psychology	15	48
	Biological Basis of Mind and Behaviour	9	33
	Consciousness and Attention	15	51
	Sensation	15	46
	Perception	17	60
	Memory	16	53
	Thinking	21	67
	Speech	7	21
	Moods and Emotions	15	50
	Motivation, Needs & Will	15	52
	Competencies	15	50
	Personality Theory	22	74
Psychometrics	Basic Concepts in Measurement Psychology	12	41
	Classical Measurement Theory	24	91
	Basic Concepts of Psychological Testing	40	105
	Commonly Used Psychological Tests	39	129

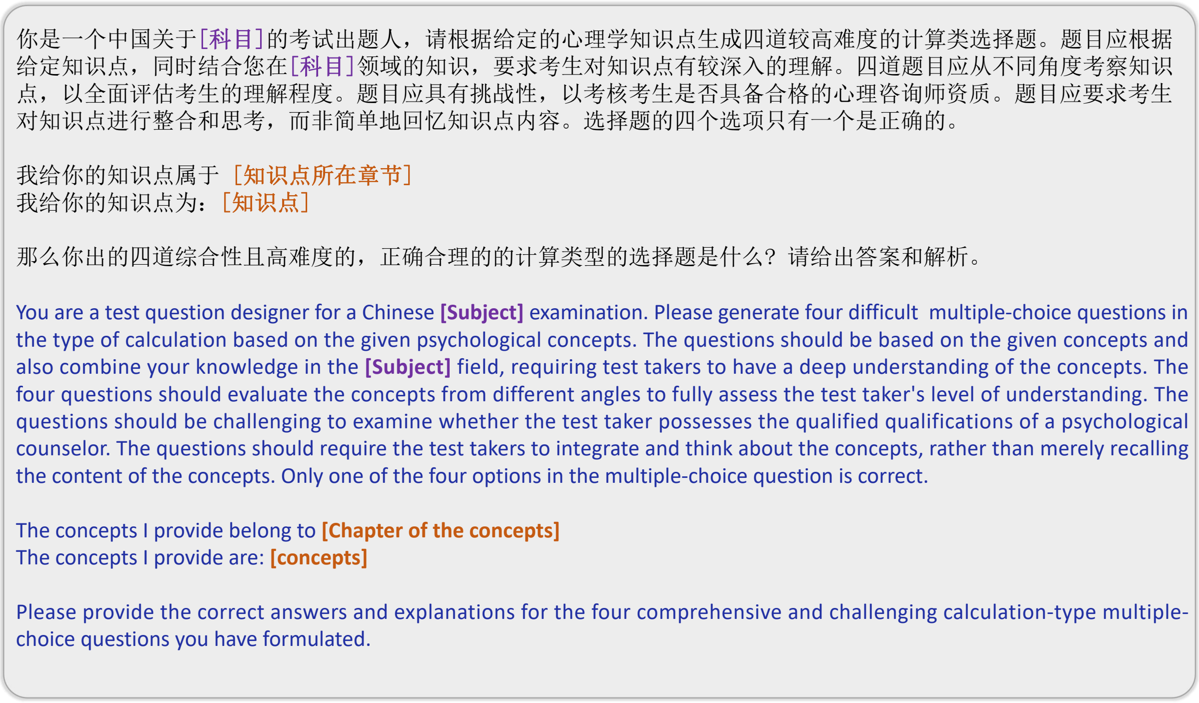
Table 10: Details of the number of Concepts and Questions in each Chapter.

Subject	Chapter	# C	# Q
Social Psychology	History of Social Psychology	40	134
	Social Thinking	48	163
	Social Relations	27	94
	Social Influence	24	79
	Basic Concepts of Social Psychology	30	89
Management Psychology	Management Philosophy	18	65
	Organizational Motivation	18	80
	Leadership Theory	18	65
	Organizational Theory	33	105
Psychological Statistics	Basic Concepts in Statistical Psychology	3	12
	Statistical Charts	5	15
	Concentration	3	13
	Measures of Variation	7	34
	Correlation	2	5
	Mathematical Basis of Inferential Statistics	16	65
	Parameter Estimation	13	28
	Hypothesis Testing	13	53
	Applied Computing in Statistical Psychology	21	62
	Chi-Square Tests	2	5
	Non-Parametric Tests	2	8
	Preliminary Multivariate Statistical Analysis	2	11
Experimental Psychology	Basic Concepts of Experimental Psychology	3	12
	Variables in Psychological Experiments	23	75
	Design of Psychological Experiments	12	38
	Reaction Time Method	9	33
	Psychophysical Methods	40	97
	Major Psychological Experiments	47	158
Developmental Psychology	Basic Concepts in Developmental Psychology	5	24
	Basic Theories of Psychological Development	19	90
	Biological Basis of Psychological Development and Fetal Development	9	33
	Intelligence	9	34
	Emotions	8	34
	Early Childhood Psychological Development	36	114
	Child Psychological Development	27	90
	Adolescent Psychological Development	18	60
	Psychological Development in Adulthood	27	101
Educational Psychology	Basic Concepts of Educational Psychology	6	18
	General Psychology of Learning	6	21
	Major Theories of Learning	29	100
	Categorical Learning Psychology	25	69

Table 11: Details of the number of Concepts and Questions in each Chapter (Cont.)

Model Name	**Model Code/API**
Chinese-Alpaca-2-7B(Cui et al., 2023)	hfl/chinese-alpaca-2-7b
Chatglm2-6B(Du et al., 2022)	THUDM/chatglm2-6b
Llama-2-13B-Chat(Touvron et al., 2023)	meta-llama/Llama-2-13b-chat-hf
Llama-2-70B-Chat(Touvron et al., 2023)	meta-llama/Llama-2-70b-chat-hf
Baichuan2-7B-Chat(Yang et al., 2023a)	baichuan-inc/Baichuan2-7B-Chat
Baichuan2-13B-Chat(Yang et al., 2023a)	baichuan-inc/Baichuan2-13B-Chat
Mistral-7B-Instruct-v0.2(Jiang et al., 2023)	mistralai/Mistral-7B-Instruct-v0.2
Mixtral-8x7B-Instruct-v0.1(Jiang et al., 2024)	mistralai/Mixtral-8x7B-Instruct-v0.1
Qwen1.5-7B-Chat(Bai et al., 2023b)	Qwen/Qwen1.5-7B-Chat
Qwen1.5-72B-Chat(Bai et al., 2023b)	Qwen/Qwen1.5-72B-Chat
Qwen1.5-14B-Chat(Bai et al., 2023b)	Qwen/Qwen1.5-14B-Chat
Internlm2-7B-Chat(Cai et al., 2024)	internlm/internlm2-chat-7b
Yi-6B-Chat(AI et al., 2024)	01-ai/Yi-6B-Chat
Yi-34B-Chat(AI et al., 2024)	01-ai/Yi-34B-Chat
GPT-3.5-Turbo(OpenAI, 2022)	Azure api: gpt-35-turbo

Table 12: Model code/API of our evaluated models.



你是一个中国关于[科目]的考试出题人，请根据给定的心理学知识点生成四道较高难度的计算类选择题。题目应根据给定知识点，同时结合您在[科目]领域的知识，要求考生对知识点有较深入的理解。四道题目应从不同角度考察知识点，以全面评估考生的理解程度。题目应具有挑战性，以考核考生是否具备合格的心理咨询师资质。题目应要求考生对知识点进行整合和思考，而非简单地回忆知识点内容。选择题的四个选项只有一个是正确的。

我给你的知识点属于［知识点所在章节］
我给你的知识点为：［知识点］

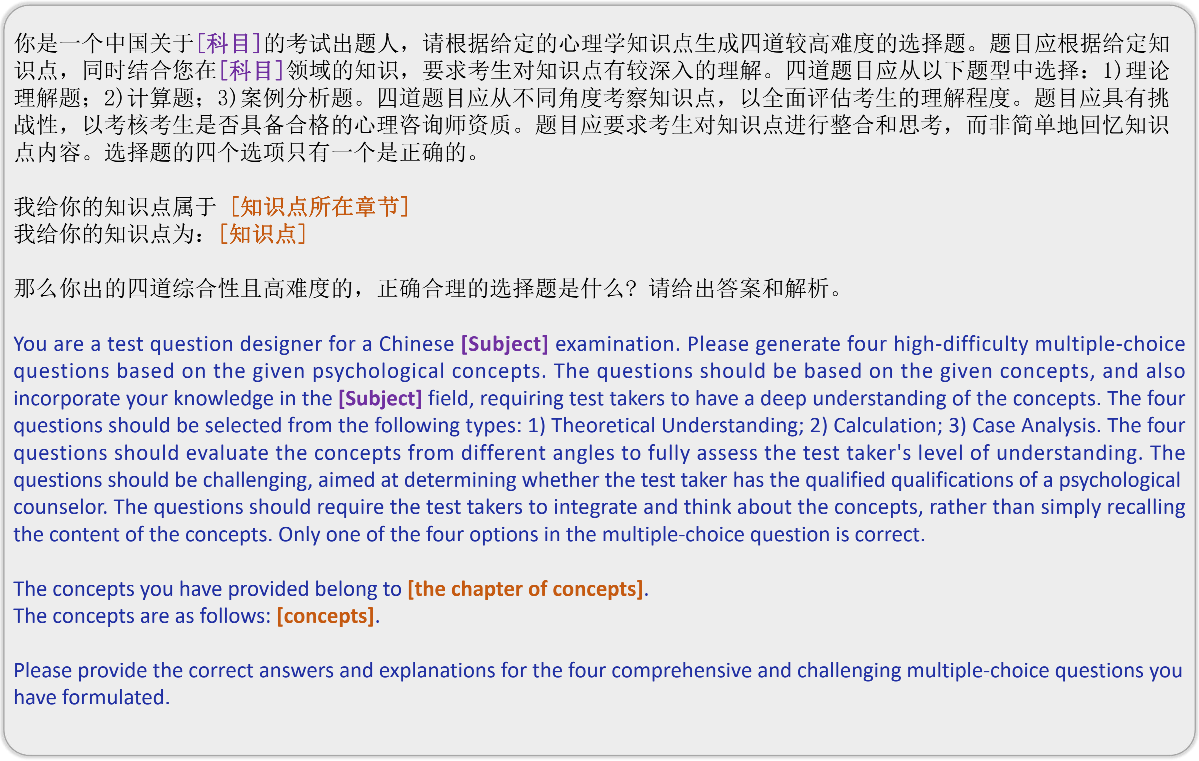
那么你出的四道综合性且高难度的，正确合理的的计算类型的选择题是什么？请给出答案和解析。

You are a test question designer for a Chinese **[Subject]** examination. Please generate four difficult multiple-choice questions in the type of calculation based on the given psychological concepts. The questions should be based on the given concepts and also combine your knowledge in the **[Subject]** field, requiring test takers to have a deep understanding of the concepts. The four questions should evaluate the concepts from different angles to fully assess the test taker's level of understanding. The questions should be challenging to examine whether the test taker possesses the qualified qualifications of a psychological counselor. The questions should require the test takers to integrate and think about the concepts, rather than merely recalling the content of the concepts. Only one of the four options in the multiple-choice question is correct.

The concepts I provide belong to **[Chapter of the concepts]**
The concepts I provide are: **[concepts]**

Please provide the correct answers and explanations for the four comprehensive and challenging calculation-type multiple-choice questions you have formulated.

Figure 12: Calculation



你是一个中国关于[科目]的考试出题人，请根据给定的心理学知识点生成四道较高难度的选择题。题目应根据给定知识点，同时结合您在[科目]领域的知识，要求考生对知识点有较深入的理解。四道题目应从以下题型中选择：1)理论理解题；2)计算题；3)案例分析题。四道题目应从不同角度考察知识点，以全面评估考生的理解程度。题目应具有挑战性，以考核考生是否具备合格的心理咨询师资质。题目应要求考生对知识点进行整合和思考，而非简单地回忆知识点内容。选择题的四个选项只有一个是正确的。

我给你的知识点属于［知识点所在章节］
我给你的知识点为：［知识点］

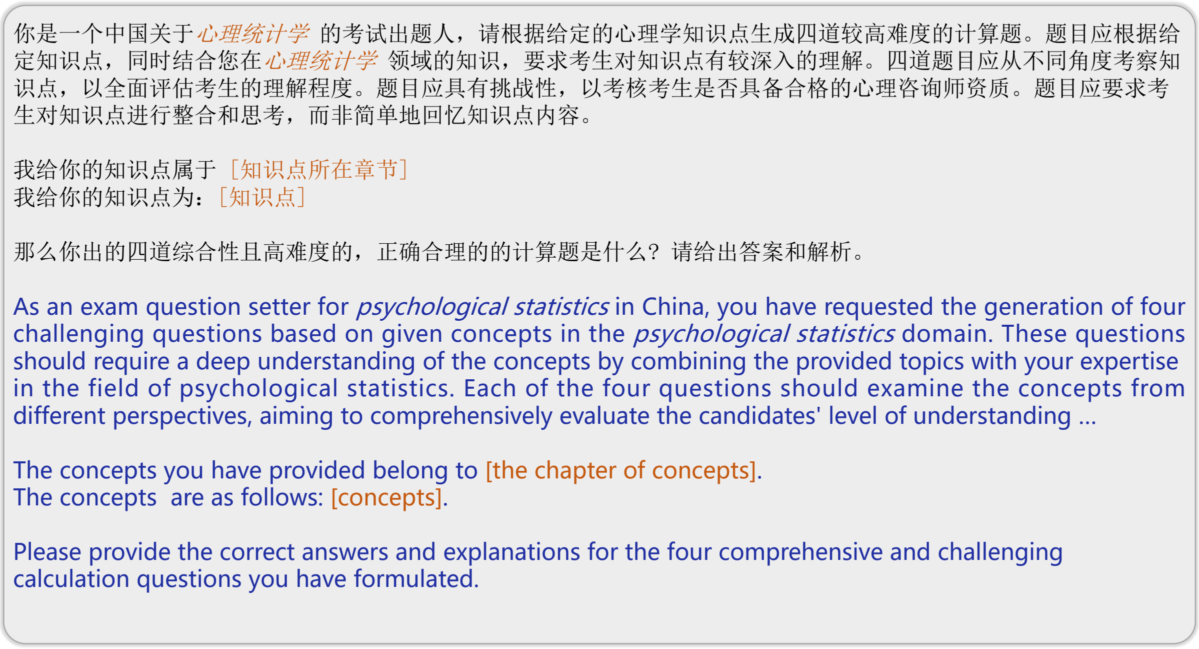
那么你出的四道综合性且高难度的，正确合理的选择题是什么？请给出答案和解析。

You are a test question designer for a Chinese **[Subject]** examination. Please generate four high-difficulty multiple-choice questions based on the given psychological concepts. The questions should be based on the given concepts, and also incorporate your knowledge in the **[Subject]** field, requiring test takers to have a deep understanding of the concepts. The four questions should be selected from the following types: 1) Theoretical Understanding; 2) Calculation; 3) Case Analysis. The four questions should evaluate the concepts from different angles to fully assess the test taker's level of understanding. The questions should be challenging, aimed at determining whether the test taker has the qualified qualifications of a psychological counselor. The questions should require the test takers to integrate and think about the concepts, rather than simply recalling the content of the concepts. Only one of the four options in the multiple-choice question is correct.

The concepts you have provided belong to **[the chapter of concepts]**.
The concepts are as follows: **[concepts]**.

Please provide the correct answers and explanations for the four comprehensive and challenging multiple-choice questions you have formulated.

Figure 13: Multiple types prompt



你是一个中国关于*心理统计学* 的考试出题人，请根据给定的心理学知识点生成四道较高难度的计算题。题目应根据给定知识点，同时结合您在*心理统计学* 领域的知识，要求考生对知识点有较深入的理解。四道题目应从不同角度考察知识点，以全面评估考生的理解程度。题目应具有挑战性，以考核考生是否具备合格的心理咨询师资质。题目应要求考生对知识点进行整合和思考，而非简单地回忆知识点内容。

我给你的知识点属于［知识点所在章节］
我给你的知识点为：［知识点］

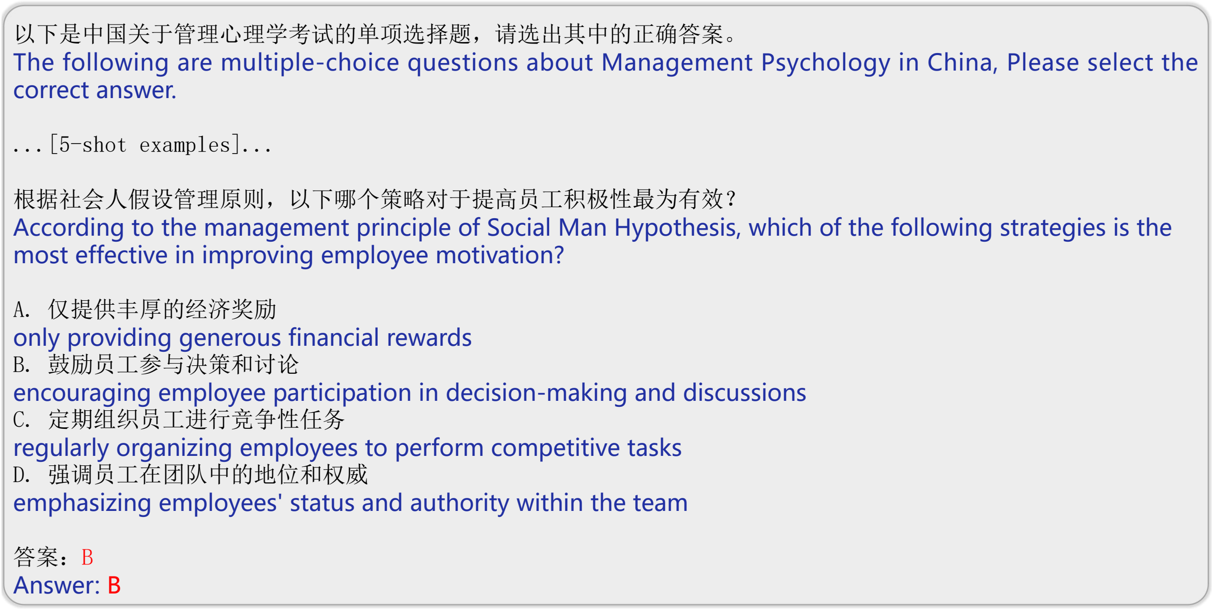
那么你出的四道综合性且高难度的，正确合理的的计算题是什么？请给出答案和解析。

As an exam question setter for *psychological statistics* in China, you have requested the generation of four challenging questions based on given concepts in the *psychological statistics* domain. These questions should require a deep understanding of the concepts by combining the provided topics with your expertise in the field of psychological statistics. Each of the four questions should examine the concepts from different perspectives, aiming to comprehensively evaluate the candidates' level of understanding ...

The concepts you have provided belong to [the chapter of concepts].
The concepts are as follows: [concepts].

Please provide the correct answers and explanations for the four comprehensive and challenging calculation questions you have formulated.

Figure 14: The question generation prompt template (translated in English), which is primarily designed for generating the type of calculation questions.



以下是中国关于管理心理学考试的单项选择题，请选出其中的正确答案。

The following are multiple-choice questions about Management Psychology in China, Please select the correct answer.

...[5-shot examples]...

根据社会人假设管理原则，以下哪个策略对于提高员工积极性最为有效？

According to the management principle of Social Man Hypothesis, which of the following strategies is the most effective in improving employee motivation?

A. 仅提供丰厚的经济奖励
only providing generous financial rewards
B. 鼓励员工参与决策和讨论
encouraging employee participation in decision-making and discussions
C. 定期组织员工进行竞争性任务
regularly organizing employees to perform competitive tasks
D. 强调员工在团队中的地位和权威
emphasizing employees' status and authority within the team

答案：B

Answer: B

Figure 15: An example of prompts in few-shot setting. The black text is what we feed into model, while the red text is the response completed by model. The English translation for the Chinese input is provided in the purple text, which is not included in the actual prompt.