ConceptPsy: A Benchmark Suite with Conceptual Comprehensiveness in Psychology

Anonymous ACL submission

Abstract

In psychology, assessing the knowledge understanding and reasoning ability of Large Language Models (LLMs) remains a substantial challenge. A key issue is the "concept bias" present in popular Chinese Massive Multitask Language Understanding (MMLU) benchmarks. This bias stems from the collected questions only cover a small set of necessary concepts. Previous Chinese MMLU benchmarks either lack the psychology discipline or only include a small subset of the required concepts. This low concept coverage rate can result in potentially misleading accuracy due to substantial performance variations across different concepts, which could further misdirect model development and refinement. To address this issue, we introduce ConceptPsy, a benchmark that comprehensively covers all college-level required concepts. In addition, we assign a chapter-level concept tag to each question, thereby enabling a more fine-grained evaluation. Our results indicates though some models achieving high average accuracy, they fail in specific concepts. In conclusion, as a valuable addition to the current MMLU benchmarks. We hope ConceptPsy can help developers to understand a their models' ability at a concept-to-concept level, subsequently guiding them to develop their models.

1 Introduction

Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020; Huang et al., 2023; Li et al., 2023; Zhong et al., 2023) benchmarks play a crucial role in the development of LLMs. These benchmarks provide average scores of each subject to give insights on a model's knowledge understanding and complex reasoning abilities, and further guide developers in refining and deploying their models. Therefore, the fairness and accuracy of these scores are vital importance.

However, in the domain of psychology, previous MMLU benchmarks either lack this important area or solely rely on a single subject to evaluate a model's ability in this broad domain, which encompasses a wide range of professional subjects. In this paper, we explore the issue of "concept bias" in Chinese MMLU benchmarks. This bias implies that the questions used for evaluation only cover a small subset of the required concepts (Figure 1). For instance, consider a subject comprising ten chapters, each contains dozons of concepts. If the questions sampled only cover the concepts from the first three chapters, the accuracy can only reflect the performance on a subset instead of the subject. We study the concept bias issue in the psychology domain of previous MMLU benchmarks and observe a low concept coverage. Consequently, the accuracy derived from biased questions can differ from people's understanding, potentially misleading developers. This bias can become more severe if an LLM's performance fluctuates significantly across different concepts.

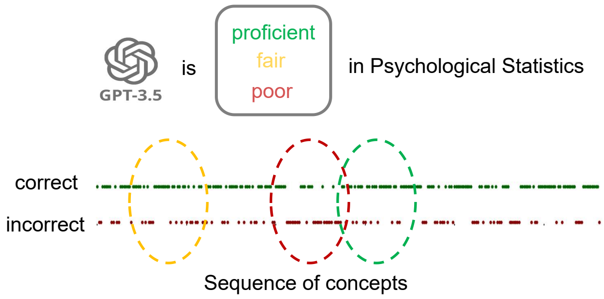

Figure 1: `GPT-3.5-Turbo`'s concept-wise performance on Psychological Statistics. The x-axis represents the sequence of concepts, arranged in the order they appear in the textbook. The dashed circles represent sampled questions. It can be observed that when the tested concept in a sampled question only covers a subset, different samplings can mislead people's understanding of a model. Although GPT-3.5-Turbo achieves an average score of 75% upon re-running, its performance across different concepts varies significantly.

To mitigate the lack of a comprehensive bench-

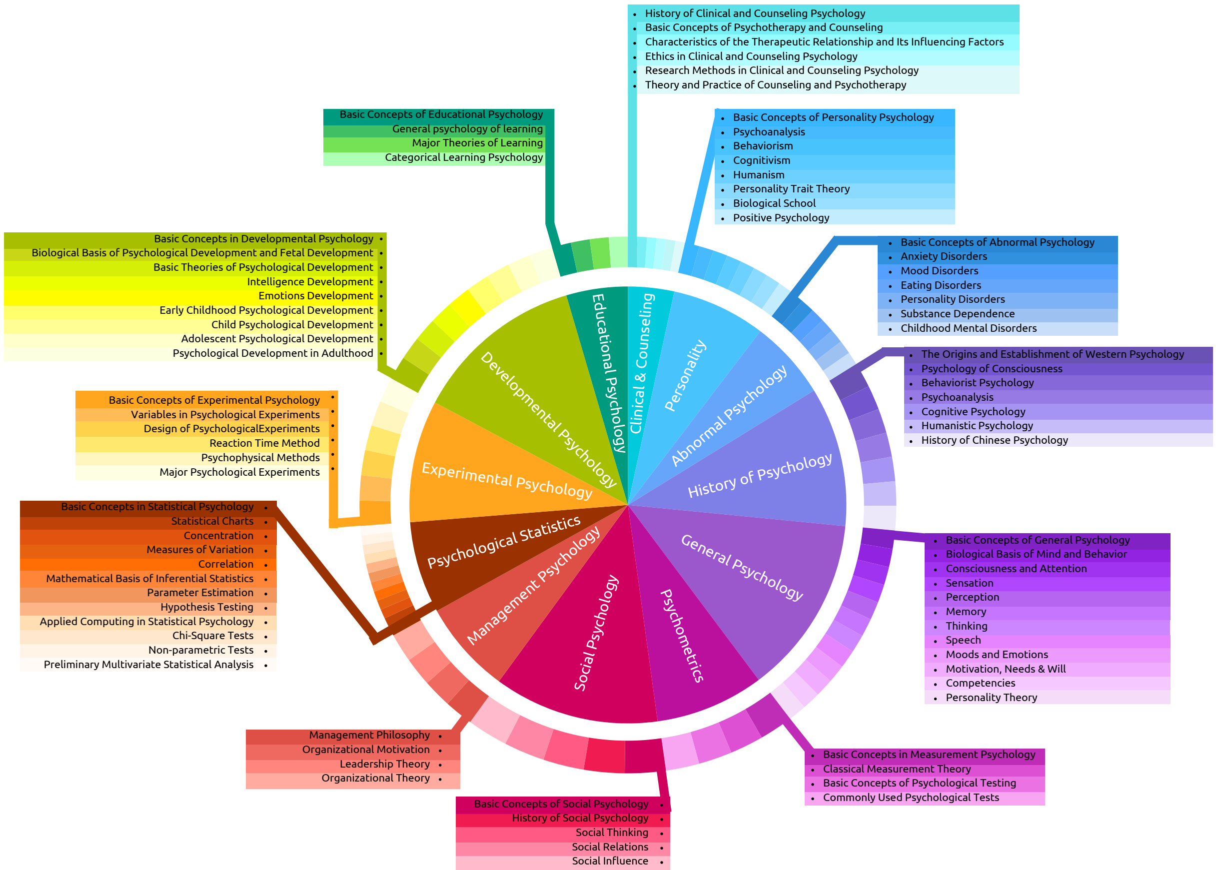

Figure 2: Diagram overview of concepts in ConceptPsy. We sample questions based on the requirement of the National Post-graduate Entrance Examination in China. Each question is tagged with a modified chapter name, serving as the chapter-level concept, to further provide chapter-level accuracy.

mark for concepts in psychology,, we present ConceptPsy, the first comprehensive, concept-wise MMLU-style benchmark in the psychology domain. ConceptPsy includes 12 college-level subjects according to the curriculum of Peking University [1]. For each subject, we manually gather all concepts (illustrated in Figure 3) based on the requirement of National Post-graduate Entrance Examination. Unlike the traditional collecting process of accumulating questions to meet a target number, collecting questions for each concept is challenging due to the volume of concepts and copyright restrictions. To overcome this, we propose a question generation framework aided by the power of GPT-4, followed by a review process conducted by professional psychological counselors. We label each question a chapter-level concepts based on the chapter of its tested concepts. Chapter-level concepts differ from the concepts used for question generation. The former uses chapter titles, while the latter refers to a knowledge unit (Figure 3). This concept-level accuracy helps developers to understanding their models' knowledge and reasoning capabilities in psychology, from both concept-to-concept and comprehensive ways.

We also conduct a case study of concept bias on popular Chinese MMLU benchmarks: C-EVAL (Huang et al., 2023) and CMMLU (Li et al., 2023). Specifically, we assess the concept coverage and performance variance of different concepts. The first metric reflects the coverage rate of the required concepts tested by the sampled questions for the subject. A higher performance variance can lead to a more biased final average accuracy. Calculating these two metrics is labor-intensive as it requires the collection of the necessary concepts within a subject. We manually gather these for two subjects: Psychology and Advanced Mathematics. To further investigate the presence of concept bias beyond psychology, we sample some subjects based on different disciplines and prompt GPT-4 to generate their chapter-level required concepts (subject outlines). We then prompte GPT-4 to catego-

[1]Training Programs of Peking University

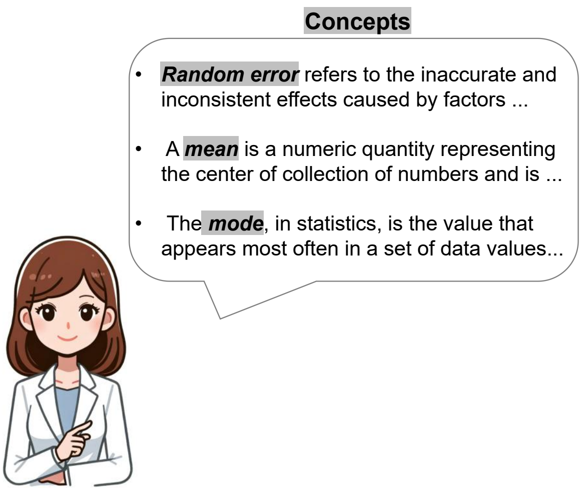

Figure 3: Examples of concepts. We define a "concept" as fundamental units of understanding that encapsulate specific knowledge within a broader field of study.

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.

rize each question under one or more chapter-level concepts. The experiments reveal that these popular Chinese benchmarks have a low concept coverage rate for each subject's sampled questions, averaging only 54%. Furthermore, LLMs like Qwen1.5-MoE-A2.7B (Bai et al., 2023) shows a standard deviation of over 10% across different concepts, which further exacerbates the impact of concept bias on average accuracy.

Finally, based on ConceptPsy, we conduct extensive experiments to evaluate a wide range of LLMs. Interestingly, while some models achieve high average scores, they perform poorly on specific chapters or concepts, highlighting the importance of comprehensive concept-wise evaluation.

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.

2 Case Study on Chinese MMLU benchmarks

We conduct a case study to analysis 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 in the human learning process , and a chapter-level concept as the modified chapter title where a tested concept is located. For example, as illustrated in Figure 3 , the "random error" can serve as 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 in a subject. 2. performance variance (§2.2): the difference in model performance across various concepts.

2.1 Analysis on the Concept Coverage

We define the "concept coverage rate" as the ratio between the number of concepts $C_{collection}$ tested by the collected questions and the number of concepts $C_{requirement}$ required by the subject:$Concept\ coverage\ rate = \frac{C_{collection}}{C_{requirement}}$. In MMLU benchmarks, a question serves as an effective tool to assess a model's knowledge comprehension and reasoning ability about the concepts being tested. Psychology is crucial for AI to comprehend human behavior. However, popular Chinese MMLU benchmarks, such as C-EVAL, do not include psychology. CMMLU includes a subject called "professional psychology" represented by 232 questions, and uses the average score to reflect a model's ability in psychology. We study the concept coverage rate in psychology for CMMLU. We also sample various subjects from each discipline to explore the potential existence of concept bias in other subjcts.

Setup Firstly, we calculate the concept coverage in psychology. We manually collect all college-level concepts (total 1383 concepts) based on the

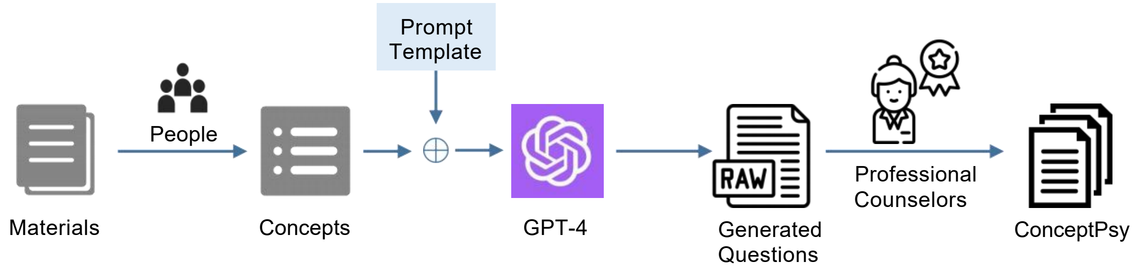

Figure 4: Overview of Our Concept-Driven Framework. We collect relevant concepts based on the requirements of corresponding examinations. To diversify the types of questions, we summarize three question patterns from these exams and design specific prompts for each type. Questions are then generated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

Subject	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.

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.

requirements of the National Post-graduate Entrance Examination in China. We deploy the robust `GPT-4` to categorize each question into one or more concepts. Due to the constraints of context length, rather than categorizing each question into specific concepts, we utilize chapter-level concept tags for each question. It's important to note that this approach may yield a higher score than the instance-level concept coverage rate. Given the scarcity of psychology-related subjects in the existing Chinese MMLU benchmark, we sampled some subjects from various disciplines (STEM, Social Science, Humanity) to further investigate the concept bias in current MMLU benchmarks. Manually collecting all required concepts for each subject is costly and impractical. Therefore, we randomly select advanced math and manually collect the concepts it requires. For the other sampled subjects, we carefully prompt `GPT-4` to generate the first/second/third-level chapter names that represent the required concepts for each subject. We also prompted `GPT-4` to categorize each question to chapters. We calculated the coverage rate of the lowest-level headings to represent the concept coverage rate. More experiments details can be found in Appendix D.

The concept coverage rate is Low. Although psychology is an vital domain for AI to learn human behavior, it is absent in C-EVAL, and its chapter-level (we use the chapter name instead of concept instance to calculate due to the budget limitation of `GPT-4`) concept coverage in CMMLU is only 59%. Questions with such an extremely low concept coverage rate could potentially mislead developers. Given the diverse range of subjects within psychology, it's not surprising that the representation of a single subject in CMMLU is low. However, even for fields like advanced mathematics, which naturally focus on a single subject, the coverage rate remains low.

Compared to CMMLU, which uses general task names (e.g., computer science), C-EVAL employs more specific subject names, which facilitates the collection of accurate required concepts. Therefore, we further calculate the coverage rate of different subjects within various disciplines in C-EVAL (Table 1). High School History and Marxism have relatively high coverage rates, but they only reach around 70%. The average coverage rate is only 54%, which might be lower than what users expect based on the average scores reflecting the number

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.

of concepts in these subjects.

2.2 Variations in Performance Across Different Concepts

Setup Based on the questions categorized into various chapter-level concepts as discussed in §2.1, we further calculate the variance in model performance across these different concepts.

The various of performance across different chapters is large. A high variance across different concepts can exacerbate the impact of a low concept rate on the average score, making it more misleading. As illustrated in Table 3 and 4, strong models like `GPT-3.5-Turbo`(OpenAI, 2022) exhibit a standard deviation exceeding 10% across different chapters. We also test open-source models such as Baichuan2-13B(Yang et al., 2023a) and Qwen1.5-MoE-A2.7B (Bai et al., 2023). They continue to demonstrate a standard deviation of over 10% on subjects sampled from different disciplines.

3 ConceptPsy

For popular Chinese MMLU benchmarks, C-EVAL lacks the critical field in human society: psychology. While CMMLU reflects the model's related ability in psychology with an average score derived from 232 questions, its chapter-level (84 chapters) concept coverage rate is only 59%, the instance-level (1383 instances). These findings highlight the urgent need for an MMLU-style psychology benchmark with low concept bias. Instead of only providing an average accuracy, we also label each question a chapter-level concept to provide more fine-grained results.

3.1 Overview

We introduce ConceptPsy to address the problem that the psychology is either absent in some MMLU benchmarks or represented with a concept bias question set. Inspired by the findings in Section 2, we design ConceptPsy with the objective of providing a conceptually comprehensive benchmark. Unlike the collection process of previous MMLU benchmarks, where questions are directly collected from public database, collecting questions for each concept can be challenging due to copyright issues. To address this, we propose generating high-quality questions using the powerful `GPT-4` model. As shown in Figure 4, we select 12 core subjects based on the renowned undergraduate psychology program at Peking University. For each subject, we manually collect required concepts based on the National Post-graduate Entrance Examination. To ensure the quality of generated questions, we carefully design prompts based on the question types of the Professional Counselor Examination[2]. We then hire professional psychological annotators to review these questions to ensure the quality. Additionally, we manually tag each question with a chapter-level concept (the name of the chapter where the concept is found). This provides a more fine-grained performance measure to assist developers in evaluating and refining their models. ConceptPsydiffers from CMMLU and C-EVAL in the following ways:

- **Low Concept Bias:** ConceptPsycovers all required concepts according to the graduate school entrance examination for the 12 core subjects in psychology.
- **Fine-grained Score:** Instead of only providing an average score for the whole subject, we also provide a chapter-level concept score.
- **Focus on Psychology:** ConceptPsy concentrates on the psychology domain, aiming to

[2]https://jcpx.psych.ac.cn/

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 5: 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.

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 knowledge point 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 knowledge point I'm providing you with is:

Knowledge points 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...

Knowledge points 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...

Knowledge points 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 6: 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.

tackle the absence and concept bias issues in psychology within previous MMLU benchmarks.

3.2 Data Collection

Subject Selection Our selection of 12 core standard subjects (shown in Table 5) within the discipline of psychology has been 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. Additionally, we included Psychological Counselling as a fundamental subject within our data set to adhere to professional qualification criteria.

Concepts Collection The Graduate Entrance Examination is an official test that evaluates the professional knowledge acquired by undergraduate students. We collect concepts based on the requirement of the 12 subjects in this exam. An example of our method for recording concept is depicted in Figure 6. To improve the quality of generated questions, we provide the GPT-4 model not only with the concepts but also with the first/second level chapter names.

To generate high-quality questions, we meticulously design prompts by analyzing professional exam question types and categorizing them into three groups: (1) Calculation; (2) Theory understanding; (3) Case Study. As shown in Figure 6, we assign an appropriate question type based on the category of the knowledge point during question generation. We utilize four types of prompts (details can be found in Appendix H). Three of these prompts are used to generate the three types of questions, and the fourth type is used to generate all three types of questions for the same concept simultaneously. Figure 14 provides an example of a calculation type multiple-choice question. To ensure a relatively even number of questions for each concept, we generate four questions per concept.

3.3 Questions Review

After using GPT-4 to generate the questions, three professional psychological annotators will review each one to ensure the quality of the dataset. We primarily review the correctness of the questions and filter out those that are unreasonable. The rules and precautions for question review are shown in Figure 5. The annotators will check the rationality of the knowledge points, the questions, and their relationship. If there are errors in a question, modifications require a consensus from at least three psychological counselors to reduce subjective bias and mistakes.

	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
`Chinese-Alpaca-2-7B`	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	0.6
`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	0.7	0.52	0.63	0.69	0.69	0.66
`Mistral-7B-Instruct`	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	0.7	0.76	0.69
`Baichuan2-13B-Chat`	0.78	0.73	0.8	0.68	0.73	0.69	0.82	0.75	0.6	0.71	0.74	0.83	0.74
`GPT-3.5-Turbo`	0.78	0.74	0.8	0.67	0.73	0.7	0.82	0.8	0.64	0.74	0.77	0.81	0.75
`Mixtral-8x7B-Instruct`	0.82	0.76	0.78	0.65	0.73	0.7	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	0.7	0.85	0.86	0.62	0.72	0.78	0.83	0.79
`Qwen1.5-72B-Chat`	0.87	0.74	0.89	0.8	0.81	0.77	**0.88**	0.87	0.38	0.77	0.83	0.85	0.79
`Qwen1.5-14B-Chat`	0.86	0.8	0.85	0.72	0.83	0.72	0.83	0.85	0.67	0.81	0.79	0.85	0.8
`Yi-34B-Chat`	**0.92**	**0.84**	**0.91**	**0.81**	**0.86**	**0.78**	**0.88**	**0.9**	0.69	**0.84**	**0.86**	**0.86**	**0.85**

Table 6: Performances on ConceptPsywith different LLMs.

4 Experiments

4.1 Dataset Statistics

We carefully select 12 psychology subjects, including: *clinical and counseling psychology, psychology of personality, abnormal psychology, history of psychology, general psychology, psychometrics, social psychology, management psychology, psychological statistics, experimental psychology, developmental psychology, educational psychology.* We manually collected 1383 concept instances from the National Post-graduate Entrance Examination-dataset. We generated 4 questions per instance, resulting in 4573 high-quality questions after a rigorous review and filtering process. Additionally, we collected 84 chapter-level concepts and assigned labels to each question to evaluate chapter-level performance.

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 G . We conduct the evaluations using a 5-shot method. We set the temperature as 0, *top_p* as 1.0.

4.3 Results

In Table 6, we showcase the performance of popular Chinese LLMs. The `Yi-34B-Chat` model outperforms, achieving an average score of 0.85, surpassing strong models such as `Qwen1.5-72B-Chat` and `GPT-3.5-Turbo`. Both `Qwen1.5-14B-Chat` and `Qwen1.5-72B-Chat` exhibit similar performance, which may suggest a lack of training data in the psychology domain, preventing the models from leveraging their larger parameter counts. Many smaller Chinese models, such as `Yi-6B-Chat` and `Qwen1.5-7B-Chat`, outperform larger and stronger models like `Mixtral-8x7B-Instruct`, potentially due to the former having more training data in Chinese. Most models demonstrate lower performance in Psychological Statistics and Psychometrics compared to other subjects, implying these models need to enhance their reasoning abilities in psychology.

4.4 Chapter-level Performance

ConceptPsyalso provides chapter-level performance insights, offering more detailed information. From Figure 7, we can understand why the strong `Qwen-72B-Chat` model's average score is lower than the smaller model `Yi-34b-Chat`. `Qwen-72B-Chat` underperforms `Yi-34b-Chat` in most chapters, particularly in reasoning chapters such as Applied Calculations in Statistical Psychology or Statistical Charts. In contrast, `Mistral-8x7B-Instruct` performs worse in theoretical chapters, possibly indicating a lack of corresponding knowledge in its training data. All models perform poorly in non-parametric tests, suggesting potential areas of improvement for developers. More chapter-level results can be found in Appendix E

5 Related Works

5.1 Chinese MMLU benchmarks

While the evolution of English language benchmarks continues to flourish (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.The AGIEval benchmark (Zhong et al., 2023) expands this with questions from various Chinese exams, while MMCU (Zeng, 2023) includes questions from diverse domains. MCU (Zeng, 2023) expands on this by collecting

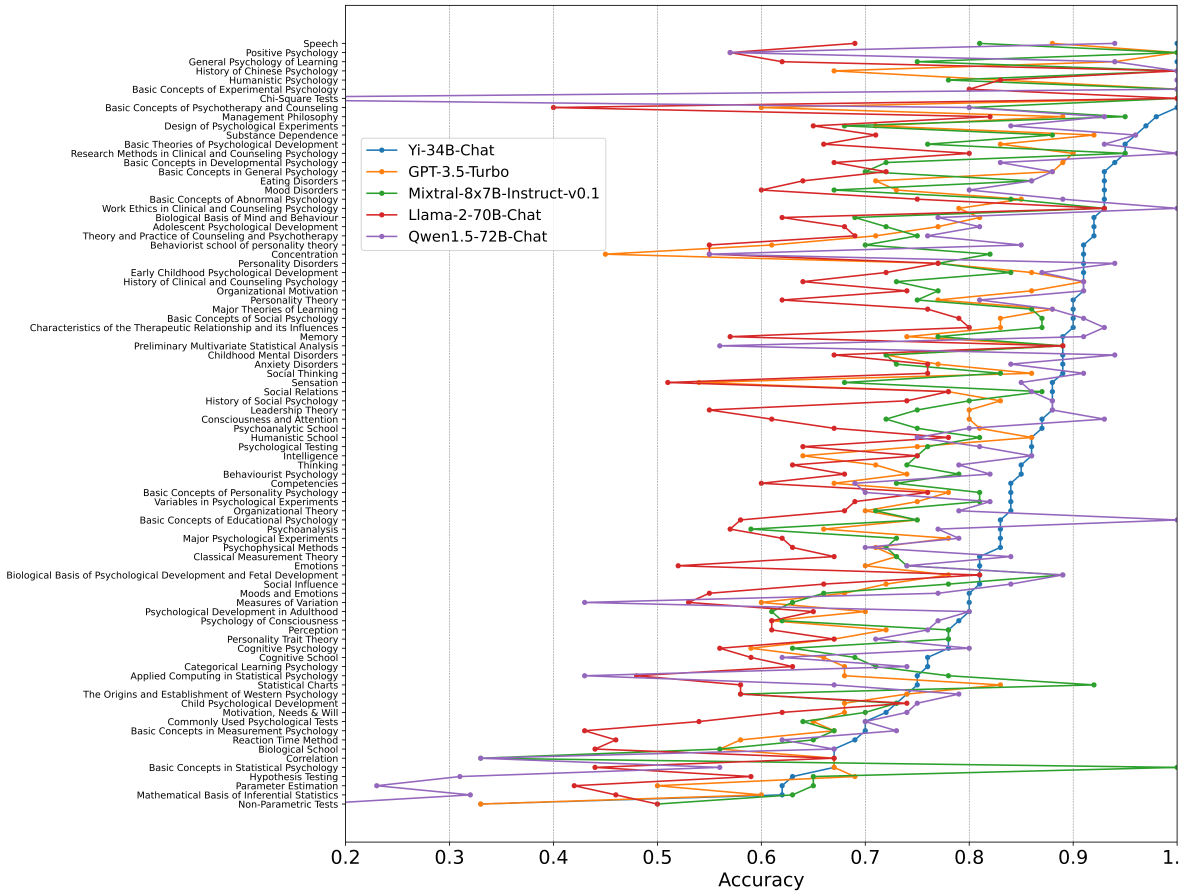

Figure 7: Concept-level results for models more than 34B, in addition to `GPT-3.5-Turbo`.

questions from a broader range of domains. The C-EVAL benchmark (Huang et al., 2023) gathers questions from different levels, including middle school, high school, and professional qualification exams, employing strategies to mitigate dataset leakage by using non-paper-based questions and simulation questions. ConceptMath (Wu et al., 2024) also evaluates LLM's ability in a concept-wise manner. However, it differs from our work in that it (1) focuses on mathematics, (2) is not a comprehensive MMLU benchmark, and (3) is released more than three months after our work.

5.2 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.

6 Discussion

This paper introduces ConceptPsy, the first comprehensive benchmark for manually collected psychology concepts in the psychology domain. This addresses the current issue where popular Chinese MMLU benchmarks either lack psychology subjects or have significant concept bias in psychology. In addition to providing an average score for each subject in ConceptPsy, we manually label each question with a chapter-level concept to offer more fine-grained results. This assists developers in better improving and deploying their models. Furthermore, we conduct a case study on the presence of concept bias in current Chinese MMLU benchmarks, further underscoring the urgent need for ConceptPsy.

Limitations

We have solely explored the concept-related bias within the Chinese MMLU series of datasets and have not investigated biases in other areas. Additionally, due to resource constraints, we have only collected and analyzed the concepts from limited subjects. Our work exclusively examines Chinese benchmarks and does not extend to benchmarks in other languages.

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

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*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. 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.

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

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.

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*.

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*.

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.

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

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 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 8. 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.

C Chapter-level Statistic of ConceptPsy

In Table 9 and Table 10, we present the chapters of each subject along with the corresponding number of concepts and questions.

D 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.

E More Results on Fine-grained Performance

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

F Qualifications of Our psychology Annotators

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

G 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., 2023), `Qwen1.5-72B-Chat`(Bai et al., 2023), `Qwen1.5-14B-Chat`(Bai et al., 2023), `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 11 . We evaluate these models with vLLM (Kwon et al., 2023).

H 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 generating 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.

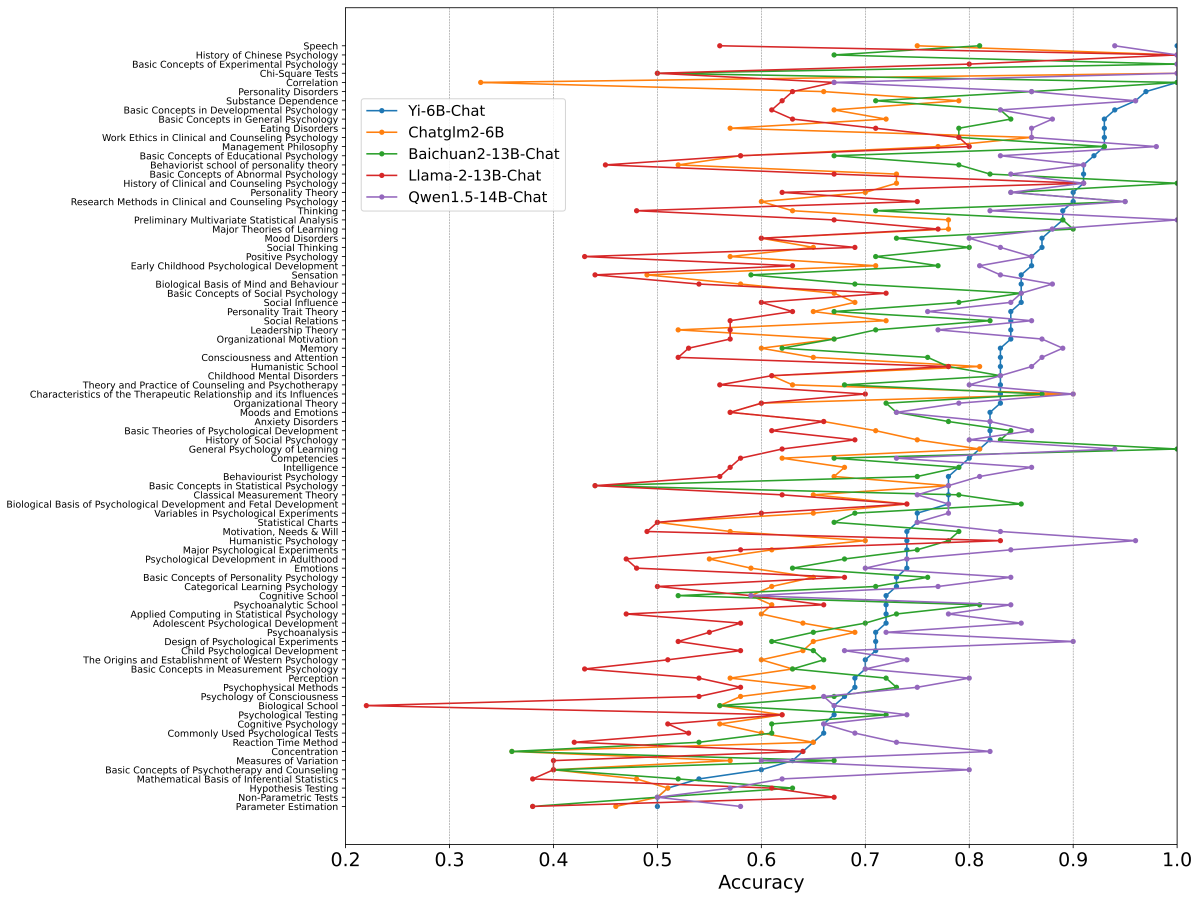

Figure 8: Concept-level results for models wi 14 Billion parameters

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 9: 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 10: Details of the number of Concepts and Questions in each Chapter (Cont.)

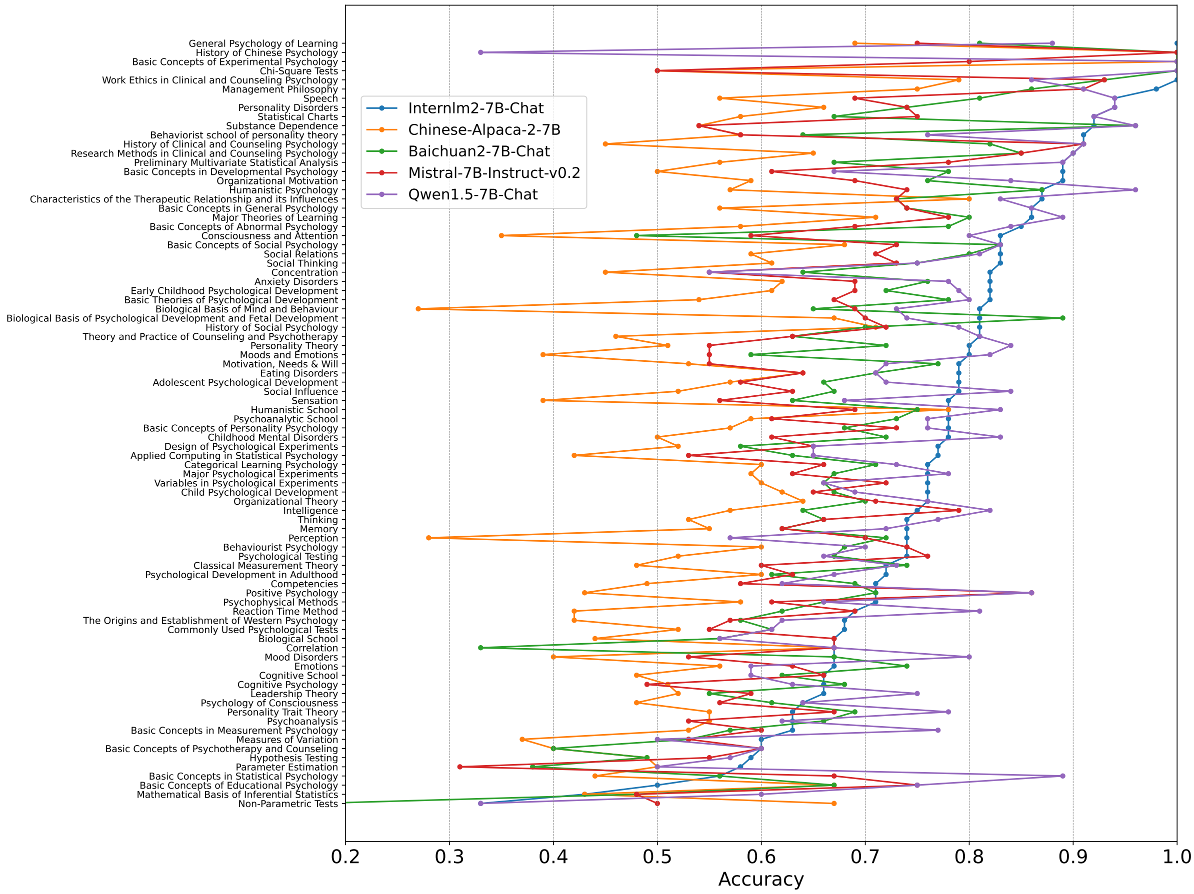

Figure 9: Concept-level results for some other models

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., 2023)	Qwen/Qwen1.5-7B-Chat
`Qwen1.5-72B-Chat`(Bai et al., 2023)	Qwen/Qwen1.5-72B-Chat
`Qwen1.5-14B-Chat`(Bai et al., 2023)	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 11: 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

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 type 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.