Concept Bias Analysis in Chinese MMLU and Introduction of PsyBench: A Benchmark with Reduced Concept Bias in Psychology

Anonymous ACL submission

Abstract

Many Chinese Massive Multitask Language Understanding (MMLU) benchmarks represent a subject by collecting multiple-choice questions and provide a score to reflect a model's ability in that subject. They have emphasized the comprehensiveness of subject variety but overlooked the comprehensiveness of concepts within individual subjects. We introduce the term *concept bias*, which refers to the bias caused by the collected questions covering only a portion of the concepts that a subject comprises. Our experiments shows that: 1) the final score can significantly vary depending on the sampled concepts, making it difficult to correlate the final score with the model's actual ability in the subject; 2) the reported model ranking can also be affected. To address this issue, we propose PsyBench: a concept-driven psychology benchmark generated by `GPT-4`. We generate high-quality questions for each required concept, thereby reducing concept bias. PsyBench not only fills the gap in the Chinese MMLU series benchmarks for the lack of comprehensive undergraduate-level psychology subjects but also reduces concept bias, offering developers scores that more accurately reflect the model's actual abilities across various subjects.

1 Introduction

Evaluation benchmarks play a core role in the process of AI development. Recently, Chinese Large Language Models (LLMs), capable of tackling a variety of tasks, have shifted the evaluation focus to more general and intricate skills, for example, the knowledge understanding and complex reasoning ability. To align with this new era, many Massive Multitask Language Understanding (MMLU) benchmarks (Hendrycks et al., 2020; Huang et al., 2023; Li et al., 2023; Zhong et al., 2023) have been proposed. These benchmarks are constructed by collecting questions to represent each subject. For example, Li et al. (2023) collects 105 questions and provides a score to reflect an LLMs' performance in college Mathematics. Unlike humans, who can only tackle a limited number of questions in an exam due to finite energy, LLMs can complete many more tests, hence the need for their assessments to be comprehensive. However, previous Chinese MMLU benchmarks have only emphasized comprehensiveness at the subject level while neglecting the comprehensiveness of the concepts (illustrated in Figure 2) covered for each subject. This oversight can lead to an incomplete representation of the subject. Figure 1 displays the scores of `GPT-3.5-Turbo` for different concepts. The horizontal axis represents concepts ordered by the sequence they are taught, with adjacent concepts on the axis being similar in content. A correct an-

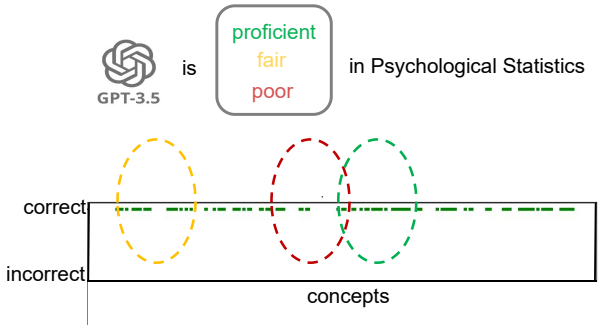

Figure 1: `GPT-3.5-Turbo`'s performance on all required concepts in Psychological Statistics. The horizontal axis corresponds to the question ID covering a particular concept, and the vertical axis indicates whether the question was answered correctly. Questions close to each other on the horizontal axis examine similar concepts. **`GPT-3.5-Turbo` tends to get sequences of similar concepts (connected on the axis) either correct or incorrect together.** This indicates that it performs similarly on similar concepts, yet there may be a significant gap between different concepts. If a benchmark only samples a subset of concepts (red, yellow, and blue ellipses), the scores can vary significantly, leading to a biased score that does not accurately reflect the LLM's true performance in the subject.

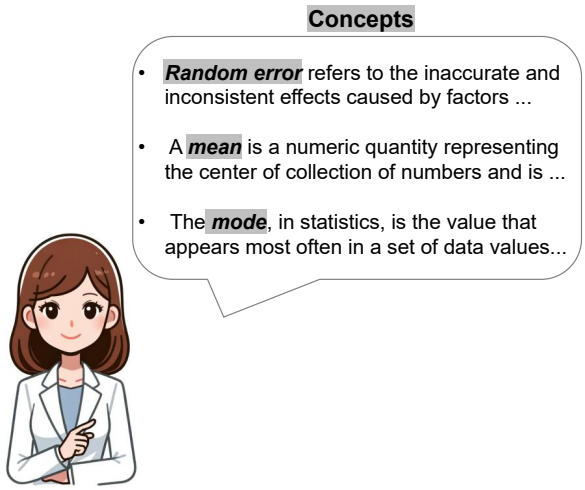

Figure 2: Examples of concepts. We define a "concept" as fundamental units of understanding that encapsulate specific knowledge within a broader field of study.

swer is marked as 1 on the vertical axis. It can be observed that the LLM tends to consecutively get similar concepts right or wrong, **indicating that the model's performance on similar concepts is alike, while there can be a significant gap between different concepts' scores**. Collecting questions can be seen as sampling a subset of concepts, and sampling from different chapters can lead to substantial variations in the final score. Such large discrepancies in scores make it challenging to correlate the final score of a subject with the model's actual ability in that subject. To quantify this, We introduce *"concept bias"*. We refer to a "concept" as a coherent snippet of knowledge (an example is illustrated in Figure 2), such as "random error", which is often used as a unit to educate human. Taking human learning as a reference, we use the concepts covered in the collected questions as a tool to analyze these MMLU benchmarks. "Concept bias" refers to the bias caused by uneven sampling of concepts within the whole concepts set.

We revisit the concept bias in the MMLU benchmarks and its influences on the reported score and ranking. We found that that the the unevenly sampled concepts can lead to a huge bias on the reported scores. To mitigate this issue, we propose a novel concept-driven dataset in which we utilize `GPT-4` to generate a comparable number of high-quality questions for each concept. We aim for our dataset to compensate for the lack of comprehensive psychological assessment in the Chinese MMLU benchmark and to provide a dataset with low concept bias to contribute to the community.

2 A Case Study on Chinese MMLU benchmarks

The MMLU series of benchmarks assess an LLM's performance of a subject by using a score reported on approximately 200 multiple-choice questions representative of that subject. As shown in Figure 1, a LLM's performance will vary across different knowledge areas. When the concepts of the sampled questions for each subject in these benchmarks are biased, it becomes difficult to provide developers with an accurate assessment of an LLM's performance in that subject. Moreover, relying solely on a single final score does not reflect the model's specific strengths and weaknesses within the subject. To analyze the impact of concept bias, we take Advanced Mathematics as an example. Its overall concepts can be obtained from the requirements of graduate entrance examinations. In §2.1, we analyze the concept coverage of two popular Chinese MMLU benchmarks, CMMLU and C-EVAL. Furthermore, in Section 2.2, we examine the performance variance of `GPT-3.5-Turbo` across different chapters of Advanced Mathematics.

2.1 Analysis on the Concept Coverage

Concept bias occurs when the concepts examined by the collected questions for a subject do not encompass all the required concepts for that subject, or when there is a disparity in the number of questions for each concept. Collecting the required concepts for a subject can be a resource-intensive task. For this reason, we choose Advanced Mathematics as a case study, a subject for which the comprehensive concepts are well-defined by existing academic requirements. Moreover, both CMMLU and C-EVAL include this subject in their benchmarks.

Setup We first analyze the $coverage\_rate$ of the subject: that is, the ratio of the number of concepts N_c covered by the collected questions to the number of concepts N_C required by the subject. We manually extracted the required concepts from the National Entrance Test for MA/MS Candidates. We collect 157 high-level concepts for the subject (e.g., the method of substitution for the first kind of integral). Additionally, we documented the chapters of the knowledge points to which each concept belongs. We employed a two-step process to tag which concepts are examined in the questions.In the first step, we utilized `GPT-4` to tag the chapters in which the concepts examined by each question were located. In the second step, given the concepts

	#questions	#concepts	coverage_rate
C-EVAL	173	94	0.54
CMMLU	104	36	0.35

Table 1: Analysis of concept coverage. CMMLU does not offer a dedicated advanced mathematics course; therefore, we assessed the coverage of advanced mathematics concepts within the college mathematics course at CMMLU.

	min	max	mean	std
C-EVAL	0.0	0.80	0.35	0.24
CMMLU	0.0	0.75	0.41	0.18

Table 2: Performance of `GPT-3.5-Turbo` on different chapters of Advanced Mathematics.

of the selected chapters, `GPT-4` is used again to tag the concepts covered by each question (each question can have multiple tags). Finally, we calculated the coverage rate as follows:

$$coverage\_rate = \frac{N_c}{N_C}$$

.

Experiments Within C-EVAL's advanced mathematics course, there are 173 questions, yet they encompass only 94 unique concepts (one question can cover multiple concepts). CMMLU's college mathematics curriculum, which includes multiple math courses, presents only 104 questions. The coverage rate for advanced mathematics concepts is alarmingly low at just 0.35. The questions from each subject within C-EVAL and CMMLU can be viewed as a random sampling of all possible concepts. When the number of covered concepts is too small, these concepts may not adequately represent the subject. In fact, it is quite possible to introduce this kind of bias during the question collection process. Many public question sets, which can be a source of questions for these benchmarks, often focus only on a subset of concepts—such as those found in end-of-chapter exercises. If the collection process is halted as soon as a target number of questions is reached, this can result in a lower coverage rate for the concepts. In §2.2, we will demonstrate that there is a significant disparity in model performance across different concepts. Furthermore, the quantity of questions per concept is not evenly distributed, resulting in a substantial bias in the scores reported by these benchmarks. Consequently, these scores do not provide a comprehensive reflection of the model's proficiency within the course.

2.2 Variations in Performance Across Different Concepts

Setup We employ `GPT-4` to categorize each question from the C-EVAL's Advanced Mathematics subject and the College Mathematics subject of CMMLU into the following chapters(following relevant entrance examinations): Functions, Basic Elementary Functions and Elementary Functions, Common Functions, Limits, Continuity of Functions, Derivatives and Their Relationship with Differentiability and Continuity, Mean Value Theorem, Monotonicity and Concavity/Convexity of Functions, Asymptotes and Curvature, Indefinite Integrals, Definite Integrals, Infinite Series, Ordinary Differential Equations and Difference Equations, Analytic Geometry of Space and Vector Algebra, Differential Calculus of Multivariable Functions, Multiple Integrals, Line Integrals, and Surface Integrals. A single question could be classified into multiple chapters. For C-EVAL, we directly utilize its Advanced Mathematics subject, while for CMMLU, we employ its College Mathematics subject and discard questions not pertaining to advanced mathematics. After classifying the questions with `GPT-4`, we compute only the for the chapters that are assigned questions.

Experiments `GPT-3.5-Turbo` demonstrate variability in performance across the chapters in the Advanced Mathematics subject, as illustrated in Table 2. The standard deviation between chapters reached as high as 0.24 for C-EVAL and 0.18 for CMMLU, which also indicates that sampling questions from different concepts can lead to significant fluctuations in scores. This variability can hinder the correlation between subject scores and model performance; for example, a score of 80 could imply proficiency or merely a collection of questions from comparatively simpler chapters.

Concepts bias can affect models ranking We also demonstrate the performance of different models under the same set of concepts (as shown in Figure 4), highlighting the issues that arise from the limited sampling of concepts. Such constraints can result in scenarios like the one depicted by the grey circle in the figure, where the ranking of model performance is influenced. For instance, within the grey circle, ChatGLM-6B surpasses GPT-3.5-

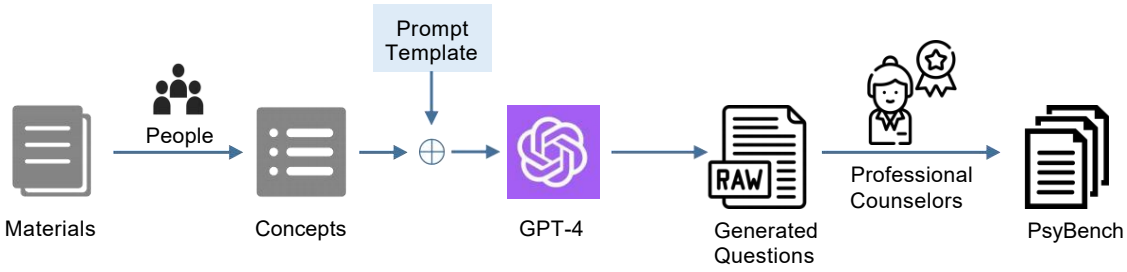

Figure 3: Overview of Our Concept-Driven Framework. We collect relevant concepts based on the requirements of corresponding examinations. To diversify the types of questions, we summarize three question patterns from these exams and design specific prompts for each type. Questions are then generated using `GPT-4`. Subsequently, we hire professional psychological counselors to review the questions for accuracy and relevance.

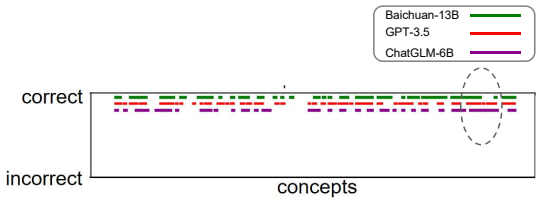

Figure 4: Performance discrepancies among models in Statistical Psychology become apparent when the sampling of concepts is limited. This can lead to situations like the one indicated by the grey circle, where the ranking of model performance is affected (within the grey circle, ChatGLM-6B outperforms GPT-3.5-Turbo).

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 counselors will filter, modify, and review the generated questions based on these requirements.

Turbo in performance. Increasing the coverage of concepts can mitigate this problem.

3 Concept-Driven Benchmark Generation

To address the issue of concept bias stemming from uneven concept sampling during question collection, we require a dataset that encompasses all necessary concepts with a similar number of questions per concept. However, due to copyright, cost, and other practical constraints, assembling a balanced collection of questions for each concept presents a significant challenge. Previous Chinese MMLU benchmarks (Li et al., 2023; Huang et al., 2023; Zhong et al., 2023) also did not emphasize concept coverage during their data collection process. To overcome this limitation, we introduce a concept-driven benchmark generation algorithm powered by the robust capabilities of *GPT-4*. Psychology, as an important field of study, has not been systematically evaluated in Chinese MMLU benchmarks such as C-EVAL and CMMLU. Although PsyEval Jin et al. (2023) provides psychological questions, it aslo focuses on subject-level comprehensiveness, without emphasizing the depth of concept coverage. Our proposed PsyBench aims to bridge this gap by providing a comprehensive benchmark for psychology within the Chinese MMLU framework.

3.1 Overview

We design PsyBench with the objective of providing a conceptually comprehensive benchmark. With this objective in mind, we propose a concept-driven benchmark generation framework and use it to create an undergraduate-level psychology benchmark: PsyBench. PsyBench differs from C-EVAL, CMMLU, and PsyEval in several key aspects:

- **Low Concept Bias:** PsyBench generates similar number, high-quality questions for each

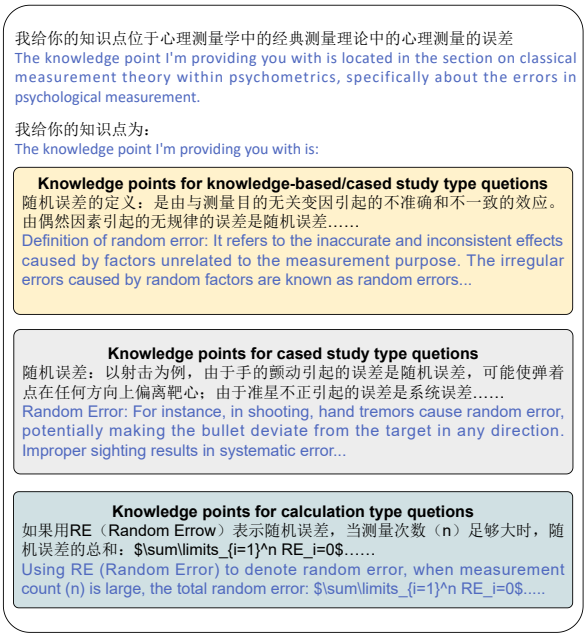

Figure 6: An example of an annotator assigning a suitable prompt to a concept. For the concept "random error", we collect multiple descriptions. The appropriate prompt is assigned based on the type of description provided.

Category	# C	# Q
In terms of subject		
Clinical & Counseling Psychology	56	156
Psychology of Personality	91	318
Abnormal Psychology	89	268
History of Psychology	126	472
General Psychology	183	605
Psychometrics	88	368
Social Psychology	169	559
Management Psychology	88	315
Psychological Statistics	99	311
Experimental Psychology	141	413
Developmental Psychology	159	580
Educational Psychology	67	208
In terms of split		
Dev	-	60
Valid	-	428
Test	-	4085
Total	1356	4573

Table 3: Statistics of psybench. The column "#C" indicates the number of concepts we have annotated for each subject, with eachconcepts generating 4 questions. The number of questions obtained after the review process is displayed in the column "#Q".

concept.

- **Comprehensive Coverage in the Field of Psychology:** Previous Chinese MMLU benchmarks lack detailed evaluating in the domain of psychology, and PsyBench serves as a complement to these.

3.2 Data Collection

Subject Selection Our selection of 12 core standard subjects within the discipline of psychology has been meticulously considered in accordance with both higher education standards[1] and professional qualification requirements. This inclusion is primarily guided by reference to the Comprehensive Examination for Psychological Counselors conducted by the Institute of Psychology of the Chinese Academy of Sciences (CAS).

Concepts Collection The Graduate Entrance Examination is an official test that evaluates the professional knowledge acquired by undergraduate students. We referenced the concepts required for the 12 subjects covered by this examination. An example of our method for recording knowledge points is depicted in Figure 6. To generate questions that are both accurate and of high quality, we provide the GPT-4 model not only with the concepts but also with the primary and secondary titles of the chapters where each concept is located. This additional context helps GPT-4 to achieve a more accurate understanding of the concepts. We attach chapter information to each generated question, enabling us not only to obtain a final score but also to gain more detailed scores for each chapter. This approach allows for a more fine-grained assessment of the model's performance in the subject.

3.3 Questions Generation

To generate high-quality questions, it is necessary to design prompt templates tailored to each subject. We drew inspiration from the CAS professional exams to design three types of multiple-choice question types: (1) Calculation; (2) Theory understanding; (3) Case Study. As illustrated in Figure 6, when generating questions, we assign an appropriate question type based on the category of the knowledge point. We employed four types of prompts (details can be found in Appendix A). Three of these are used to generate the aforementioned three types of questions for a given concept, and the fourth type is used to generate all three types of questions for the same concept simultaneously. Figure 7 presents an example of a calculation type multiple-choice question. To ensure a

[1] https://www.psy.pku.edu.cn/

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

relatively uniform number of questions for each concept, we generate four questions per concept.

4 Experiments

4.1 Dataset Statistics

We carefully select 12 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. Based on the knowledge points stipulated by the NEET, we generate four questions for each point. Subsequently, we filter out questions of low quality or containing errors. The specific quantities of knowledge points and the number of questions post-review are presented in Table 3.

4.2 Setup

We evaluated strong open-sourced Chinese models: `ChatGLM-6B` by Zeng et al. (2022); Du et al. (2022) and `Baichuan-13B-Chat` by (Yang et al., 2023), as well as the strong closed-sourced model: `GPT-3.5-Turbo` by (OpenAI, 2022). We conducted the evaluations using a 5-shot method (as illustrated in Figure 8). When generating, we set the temperature as 0, *top_p* as 1.0.

4.3 Results

In Table 4, we present the performance of various Large Language Models (LLMs) on PsyBench. Overall, `Baichuan-13B-Chat` achieved performance comparable to `GPT-3.5-Turbo`, which may be attributed to the fact that PsyBench is a Chinese dataset and `Baichuan-13B-Chat` has been trained on a more extensive corpus of Chinese text. Notably, in the computation-heavy chapters of Psychometrics and Psychological Statistics, `Baichuan-13B-Chat` performed slightly better than, or similar to, `GPT-3.5-Turbo`, and significantly outperformed `ChatGLM-6B`. This suggests that `ChatGLM-6B` may still lag behind in computational and reasoning capabilities.

5 Related Works

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), the development of similar benchmarks for the Chinese language environment remains underexplored. The CLUE benchmark (Xu et al., 2020) serves as the first large-scale Natural Language Understanding (NLU) benchmark for Chinese and is extensively used as a public dataset. More recently, the AGIEval benchmark (Zhong et al., 2023) has been introduced, consisting of questions derived

	GPT-3.5-Turbo	ChatGLM-6B	Baichuan-13B-Chat
Clinical & Counseling Psychology	0.78	0.71	0.80
Psychology of Personality	0.76	0.63	0.73
Abnormal Psychology	0.77	0.72	0.81
History of Psychology	0.65	0.61	0.68
General Psychology	0.72	0.66	0.73
Psychometrics	0.67	0.60	0.68
Social Psychology	0.81	0.75	0.82
Management Psychology	0.79	0.71	0.78
Psychological Statistics	0.63	0.51	0.60
Experimental Psychology	0.70	0.59	0.71
Developmental Psychology	0.77	0.68	0.74
Educational Psychology	0.81	0.80	0.82
Avg	0.74	0.66	0.74

Table 4: Performances on PsyBench with different LLMs

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 8: 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.

from publicly available Chinese College Entrance Exam, Chinese Lawyer Qualification Test, and Chinese Civil Service Examination. MMCU (Zeng, 2023) expands on this by collecting questions not just from the college entrance exams but also from a broader range of domains, such as medicine, law, and psychology. 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. In comparison with AGIEval, MMCU, and C-EVAL, our proposed psybench benchmark (1) represents the first deep evaluation within a single discipline, covering undergraduate core courses assessed by the postgraduate entrance examination and professional qualification tests; (2) our goal is not to gather as many questions as possible for each subject, but rather to evenly cover all necessary knowledge points, contributing to a more realistic and reliable model evaluation; (3) we introduce a semi-automated benchmark generation process, requiring the participation of trained annotators only during the ques-

tion review process. Compared to redesigning new questions from scratch, our approach enables the creation of new datasets at a significantly reduced cost for regular test set updates. This process, comprised of unpublished questions and regular updates, assists in mitigating dataset leakage issues.

6 Discussion

This paper addresses the issue of concept bias, which arises from the non-uniform sampling of concepts during the question collection phase of the MMLU benchmark. This bias may lead to reported scores that do not accurately reflect the true performance of Large Language Models (LLMs) on the subject in question. Moreover, since models may perform differently across various concepts, inappropriate sampling could result in changes to model rankings. To address this, we propose a concept-driven benchmark based on `GPT-4`, which can generate a domain-specific benchmark set at a lower cost. Utilizing this, we have created a psychology benchmark to address the shortcomings of previous Chinese MMLU benchmarks in this crucial field, covering most concepts required by undergraduate students.

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 the subject of Advanced Mathematics, without conducting a comprehensive analysis across multiple disciplines. 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 knowledge points prescribed by the Na- tional Entrance Examination for Postgraduates. Furthermore, the generated data is subjected to rigor- ous scrutiny by professional psychologists to ensure its ethical soundness.

References

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*.

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*.

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.

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. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

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*.

A 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 9, 10, 11, and 12 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 9: 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 10: 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 11: 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 12: multiple type prompt