SuperCLUE-Math6: Graded Multi-Step Math Reasoning Benchmark for LLMs in Chinese

Liang Xu, Hang Xue, Lei Zhu, Kangkang Zhao

SuperCLUE team
contact@superclue.ai

Abstract

We introduce SuperCLUE-Math6(SC-Math6), a new benchmark dataset to evaluate the mathematical reasoning abilities of Chinese language models. SC-Math6 is designed as an upgraded Chinese version of the GSM8K dataset with enhanced difficulty, diversity, and application scope. It consists of over 2000 mathematical word problems requiring multi-step reasoning and providing natural language solutions. We propose an innovative scheme to quantify the reasoning capability of large models based on performance over problems with different reasoning steps. Experiments on 13 representative Chinese models demonstrate a clear stratification of reasoning levels, with top models like GPT-4 showing superior performance. SC-Math6 fills the gap in Chinese mathematical reasoning benchmarks and provides a comprehensive testbed to advance the intelligence of Chinese language models. [1].

1 Introduction

Recent advances in large language models like GPT-4 [1] have sparked great interest in evaluating their proficiency in solving reasoning problems. While benchmarks like GSM8K [2] have been influential, they are limited to English and do not sufficiently test multi-step inference. To overcome these limitations and systematically assess the mathematical reasoning of Chinese models, we introduce SC-Math6 as an upgraded Chinese version of GSM8K.

SC-Math6 has 1072 unique problems covering a diverse range of grade school math topics. Each problem is presented in a native Chinese context and accompanied by a detailed natural language solution walkthrough. Moreover, SC-Math6 provides a follow-up question for each initial query to assess the model's continuous reasoning ability during interaction with users (As shown in Figure 1. For more examples, please refer to the appendix A). We also propose a novel scoring scheme that combines performance over problems with different reasoning steps and overall accuracy to produce interpretable and fair reasoning levels from 1 to 5. Disparities and Correlations of SC-Math6 and GSM8K is presented in Table 1

Comparison Item	SC-Math6	GSM8K
Mathematical Logic Reasoning	YES	YES
Natural Language Solutions	YES	YES
Elementary Mathematical Knowledge	YES	YES
Multi-step Reasoning	YES	YES
Native Chinese Context	YES	NO
Multi-round In-depth Reasoning	YES	NO
Reasoning Steps in Problems	YES	NO
Interpretable Reasoning Level for LLMs	YES	NO
Number of Test Questions	2144 (1072 Pairs)	1300

Table 1: SC-Math6 and GSM8K: Disparities and Correlations

Our experiments on 13 major Chinese models demonstrate a clear stratification of reasoning capabilities. Advanced models like GPT-4 exhibit remarkably high accuracy on multi-step problems, while lower-level models show large performance gaps. The diversified grading scheme provides a reference for model selection and evaluation. SC-Math6 thus contributes the first comprehensive Chinese benchmark to assess and improve the mathematical reasoning abilities of Chinese language models.

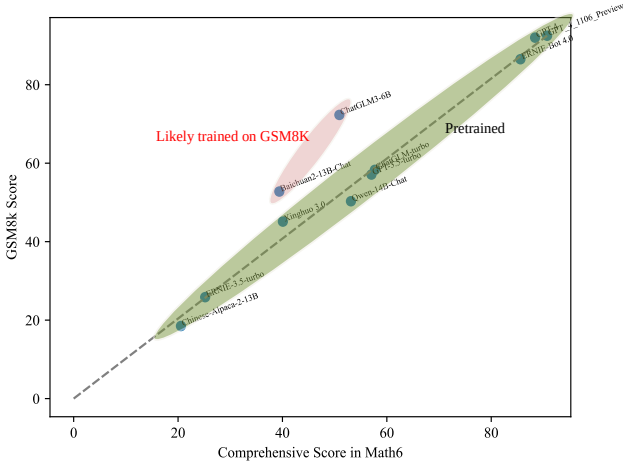

Figure 2: Trend analysis between GSM8k and SC-Math6. SC-Math6 aligns with GSM8K yet demands in-depth reasoning, whereas some models may struggle with SC-Math6.

Preprint.

[1] Our benchmark can be found at https://www.CLUEbenchmarks.com/superclue_math6.html

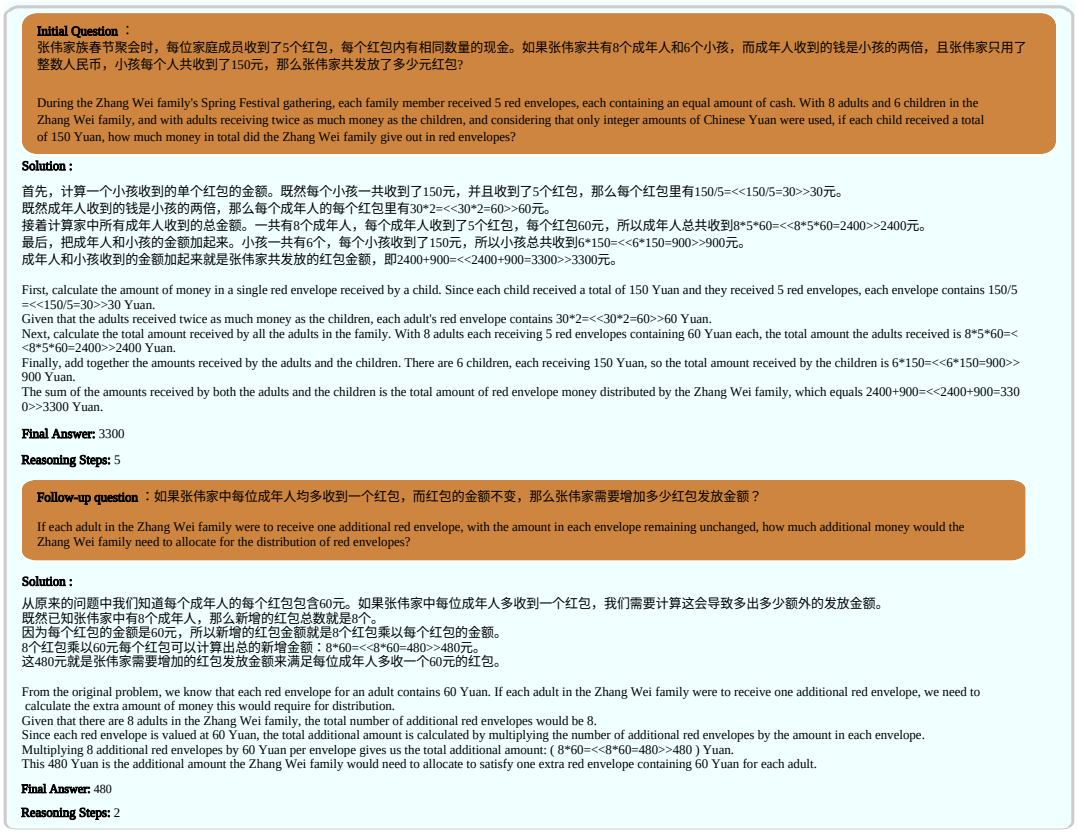

Figure 1: An example of a problem in SC-Math6

This work pioneers the systematic evaluation and benchmarking of mathematical reasoning capabilities of major Chinese language models. The key contributions are threefold:

First, the construction of SuperCLUE-Math6, the first native Chinese multi-turn, multi-step mathematical reasoning dataset for assessing model logical thinking and reasoning skills.

Second, the proposal of a novel transparent and consistent framework to parse and evaluate model reasoning levels, providing quantifiable metrics of model intellectual capabilities.

Third, comprehensive benchmarking and analysis of leading Chinese models on SuperCLUE-Math6, offering valuable insights into current model strengths, weaknesses and factors impacting reasoning performance.

Overall, this research fills the gap in Chinese mathematical reasoning evaluation and establishes an important benchmark for advancing the reasoning abilities of Chinese language models. The benchmark and insights lay a solid foundation for developing models with more human-like reasoning.

2 SuperCLUE-Math6

Data Collection

We first curated a large pool of Chinese math problems from elementary school exams and books and altered it manually to ensure it was unique. Two criteria were then applied to select problems requiring: 1) At least one step of reasoning, and 2) Error-free natural language solutions. This yielded 1072 unique arithmetic problems.

To evaluate the model's proficiency in sustained inferential reasoning throughout the interactive engagement, we designed multi-turn follow-up questions for each problem, bringing the total size to 2144.

Quality Control and Inspection of Questions. In the second round process, all questions were subjected to manual verification, which required the annotators to solve the questions themselves and record their answers. These were then compared with the provided reference answers and solution steps. If any inconsistencies were discovered, the problem was pinpointed and corrected. Corrections were made if there were issues with the answer or solution steps; if the question itself was ambiguous, it required clarification. Once consistency was confirmed, the process advanced to the next question. After the manual verification, a final round of random sampling checks was conducted. Out of 50 pairs of questions, 1 pair was found to potentially have ambiguities and needed to be corrected, giving us a sampling accuracy

of 98%.

The distribution of reasoning steps is controlled to prevent biases and test varied capabilities: 15-20% with 1 step, 15-20% with 2 steps, 45-50% with 3 steps, and 5-10% with 4-5 steps. The textual lengths of problems and solutions also exhibit high variability.

Scoring Scheme and Reasoning Level

To produce interpretable and fair quantification of reasoning capabilities, we propose a scoring scheme that combines:

- Reasoning Steps Score: Higher weight assigned for more steps based on the insight that longer reasoning chains are more difficult. Firstly, we separately compute the average score of the model corresponding to each inference step across the set of problems. Subsequently, we employ the number of inference steps as weights to calculate a weighted average of the scores at each step, thereby deriving a score that is weighted according to the number of inference steps.
- Overall Accuracy Score: The Overall Average Score is derived as the mean value of the Mean Accuracy and the Strict Interaction Accuracy. Mean Accuracy is computed by considering each question and its corresponding follow-up question as two separate items, thus calculating the average accuracy across 2144 questions. Conversely, Strict Interaction Accuracy is calculated by treating the question and its follow-up question as a unified interactive pair, with a point awarded only if both the question and the follow-up question are correctly answered, demonstrating proper reasoning. This Strict Interaction Accuracy is evaluated over a cohort of 1072 test pairs to establish the average interaction accuracy.
- Comprehensive Score: Calculated as the weighted sum of the Reasoning Steps Score and the Overall Accuracy Score, each component contributing equally to the final score.
- Reasoning Level: The Reasoning Level of a language model is based on the Comprehensive Score, with levels ranging from 1 to 5, where level 5 is the highest and level 1 is the lowest. A threshold of 5 points is used to determine the levels. If the composite scores of the two models differ by less than 5 points, they are considered to be within the same level. This provides a transparent system to classify model capabilities.

The accuracy score is a commonly used evaluation metric to quantify the mathematical reasoning abilities of language models, but it fails to account for the varying difficulty levels of individual questions. Solving a more challenging problem should be awarded more points than solving an easier one, and the number of reasoning steps involved usually correlates with the difficulty level. Therefore, we incorporate the Reasoning Steps Score into our assessment framework.

The advantage of the Reasoning Step Score lies in its ability to account for the varying difficulty levels of different questions. This score is validated by manual problem-solving, ensuring a high level of accuracy in the reasoning steps involved. Given that mathematical problems may have multiple solution methods, each with a potentially different number of reasoning steps, the reasoning step count is not necessarily unique. Therefore, the calculation of Reasoning Step Scores cannot eliminate the discrepancies in weighted precision.

On the other hand, the Overall Accuracy Score, while not considering the difficulty level of each question, ensures fairness and avoids bias that might be introduced during the weighting process. Therefore, we have not completely discarded the Overall Accuracy Score. Instead, we employ a weighted summation of Reasoning Steps Score and Overall Accuracy Score to calculate the unified score. This method balances the precision given by the Reasoning Steps Score with the fairness ensured by the Overall Accuracy Score, aiming to provide a more comprehensive evaluation of mathematical reasoning performance.

Experiments and Analysis

We evaluated 13 major Chinese models on SC-Math6 covering capacities from 13B to Proprietary APIs. Table 2 presents the overall accuracy, Reasoning Steps Score, Comprehensive Score, and resulting Reasoning Level. Information on models is shown in Table 3.

Model Name	R Level	Comp. Score	Reas. Steps Score	OvrAcc Score
GPT-4-1106-Preview	**5**	90.71	91.65	89.77
GPT-4	**5**	88.40	89.10	87.71
Ernie-bot 4.0	**5**	85.60	86.82	84.38
GLM-4	**5**	84.24	85.72	82.77
Xinghuo 3.5	**5**	83.73	85.37	82.09
ChatGLM-Turbo	**4**	57.70	60.32	55.09
GPT-3.5-Turbo	**4**	57.05	59.61	54.50
Qwen-14B-Chat	**4**	53.12	55.99	50.26
ChatGLM3-6B	**3**	40.90	44.20	37.60
Xinghuo 3.0	**3**	40.08	45.27	34.89
Baichuan2-13B-Chat	**3**	39.40	42.63	36.18
Ernie-3.5-turbo	**2**	25.19	27.70	22.67
Chinese-Alpaca2-13B	**2**	20.55	22.52	18.58

Table 2: SC-Math6 Model Reasoning Level. 'R Level' for Reasoning Level, 'Comp. Score' stands for Comprehensive Score, 'Reas. Steps Score' stands for Reasoning Steps Score, 'OvrAcc Score' stands for Overall Accuracy Score.

Model Name	Organization	Access
GPT-4-1106-Preview	OpenAI	API
GPT-4	OpenAI	API
Ernie-bot 4.0	Baidu	API
GLM-4	ZhiPu	Web Page
Xinghuo 3.5	Iflytek	API
ChatGLM-Turbo	ZhiPu	API
GPT-3.5-Turbo	OpenAI	API
Qwen-14B-Chat	Alibaba	API
ChatGLM3-6B	ZhiPu	Weight
Xinghuo 3.0	Iflytek	API
Baichuan2-13B-Chat	Baichuan	Weight
Ernie-3.5-turbo	Baidu	Weight
Chinese-Alpaca2-13B	Yiming Cui	Weight

Table 3: Model information

Top models like GPT-4 exhibit remarkably high performance on multi-step problems, demonstrating advanced reasoning skills. There is also a clear stratification of capabilities, with higher-scoring models significantly

outperforming lower ones. The diverse levels allow the selection of appropriate models based on application requirements.

We highlight several observations:

1) Comparison between GSM8k Score and Comprehensive Score in SC-Math6

The comparative analysis of performance on the GSM8k and SC-Math6 benchmark. SC-Math6(Average model score is 55.34) aligns with GSM8K(Average model score is 59.21) yet demands more advanced reasoning, whereas models excelling on GSM8K may struggle on SC-Math6. It suggests that the SC-Math6 benchmark presents a greater level of difficulty. It was observed that, across the board, models tend to score lower on the SC-Math6 benchmark compared to GSM8k, with this trend being particularly pronounced for the ChatGLM3-6B and Baichuan2-13B-Chat models(As shown in Figure 2).

2) Performance Declining during Multi-turn Interaction

In all models observed, the accuracy rate of the second iteration generally falls below that of the first, indicating a decline in model performance with increasing task complexity from the first to the second iteration(As shown in Table 4). This trend is ubiquitous across all models, suggesting that special attention should be given to the stability and adaptability of models in sustained tasks during their design and optimization processes.

Model Name	Accuracy of Turn 1	Accuracy of Turn 2	Difference
GPT-4-1106-Preview	95.43	89.37	**-6.06**
GPT-4	94.12	87.13	-6.99
Ernie-bot 4.0	91.98	83.96	-8.02
GLM-4	90.39	83.02	**-7.37**
Xinghuo 3.5	91.70	81.44	-10.26
ChatGLM-Turbo	73.69	52.80	-20.89
GPT-3.5-Turbo	70.99	53.59	-17.40
Qwen-14B-Chat	73.23	46.26	-26.97
ChatGLM3-6B	61.10	35.35	-25.75
Xinghuo 3.0	70.99	25.65	**-45.34**
Baichuan2-13B-Chat	58.86	33.77	-25.09
Chinese-Alpaca2-13B	35.63	16.62	-19.01
Ernie-3.5-turbo	43.00	20.43	-22.57

Table 4: SC-Math6 Accuracy during interaction

For instance, the GPT-4-1106-Preview model exhibited a first iteration accuracy rate of 95.43%, which decreased to 89.37% in the second iteration, marking an 6.06% reduction in accuracy. Similarly, the ERNIE_35_Turbo model's accuracy rate declined from 43.00% in the first iteration to 20.43% in the second, constituting a 22.57% decrease.

3) Correlation between Instruction Compliance Ratio and Comprehensive Score

Models with high compliance to the instructed output formats also tend to achieve higher Comprehensive Scores, suggesting instruction understanding as an important indicator(As shown in Figure 3).

4) The Potential Relationship between Mathematical Reasoning Proficiency and Response Length

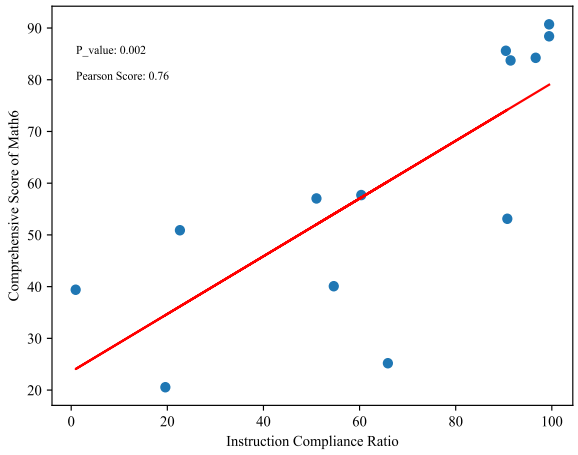

Figure 3: Correlation between instruction compliance ratio and Comprehensive Score

It has been observed that models yielding longer average response lengths tend to receive higher evaluation scores(As shown in Figure 4). In certain models, such as GPT-4-1106-Preview, a higher accuracy rate is accompanied by a longer average response length, which may suggest that these models are more precise when generating comprehensive responses. However, this trend is not consistently observed across all models, indicating that the relationship between response length and accuracy rate may be influenced by a multitude of factors, including the design of the model and the training data employed.

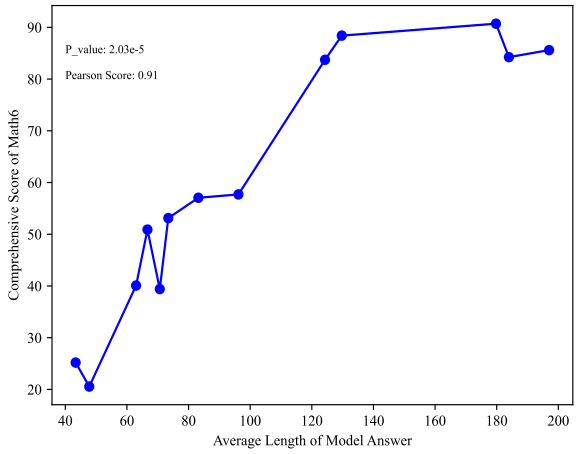

Figure 4: Relation between response length and Comprehensive Score

The GPT-4-1106-Preview exhibits an average response length of 179.78, correlating with a higher accuracy rate, while the ChatGLM3-6B shows a comparatively shorter

average response length of 66.66, with a corresponding lower accuracy rate. This implies that, in certain instances, there may be a correlation between response length and accuracy rate.

5) Performance Declining with Reasoning Steps Increasing

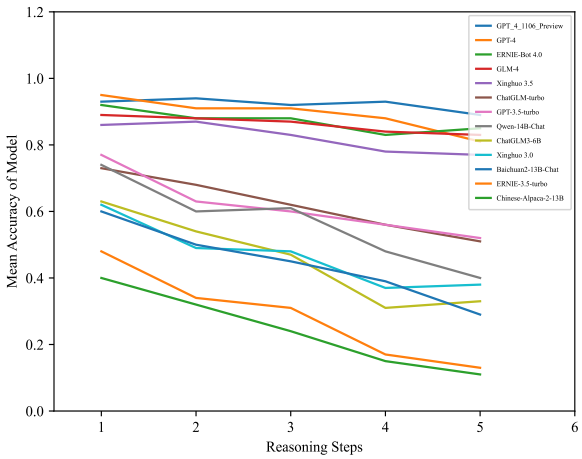

Figure 5: Relation between reasoning steps and Mean Accuracy

Analysis of the step-by-step scores reveals declining performance as problem complexity increases from 1 to 5 reasoning steps(As shown in Figure 5). This highlights the need to improve the model's capability to tackle more challenging problems requiring more reasoning steps. The results provide comprehensive insights to guide further progress on mathematical and general reasoning for Chinese language models.

3 Related Work

Benchmarks to evaluate the reasoning skills of language models have gained increasing research attention. Existing datasets mostly focus on English, including GSM8K for mathematical reasoning [2], MATH for complexity mathematical problem [3]. For general LLMs benchmark, we can find MT-bench [4], AlpacaEval [5]. We can find reasoning benchmarks for NLP, such as WinoGrad Schema Challenge for commonsense reasoning [6], and ARC for scientific question answering [7]. Our work aims to close the gap for the Chinese through a systematically designed mathematical reasoning benchmark. Our focus is to provide a benchmark to evaluate the general reasoning skills of Chinese language models. The diverse problems and reasoning patterns in SC-Math6 complement these methods to inspire new model designs and training strategies targeting enhanced mathematical intelligence.

4 Conclusion

We present SC-Math6 as the first native Chinese benchmark dataset for assessing the multi-step mathematical reasoning skills of language models in a multi-turn interaction. Developing human-like intelligence requires rich, diverse datasets like SC-Math6 that test sophisticated capabilities beyond pattern recognition. Our work aims to catalyze advances in Chinese models to better support real-world applications. The human-annotated natural language solutions also provide valuable data for training. We hope SC-Math6 will inspire exciting new model designs and training strategies targeting enhanced reasoning.

References

[1] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

[2] Cobbe, Karl and Kosaraju, Vineet and Bavarian, Mohammad and Chen, Mark and Jun, Heewoo and Kaiser, Lukasz and Plappert, Matthias and Tworek, Jerry and Hilton, Jacob and Nakano, Reiichiro and others, 2021. Training verifiers to solve math word problems. arXiv:2110.14168.

[3] Hendrycks, Dan and Burns, Collin and Kadavath, Saurav and Arora, Akul and Basart, Steven and Tang, Eric and Song, Dawn and Steinhardt, Jacob, 2021. Measuring mathematical problem solving with the math dataset. arXiv:2103.03874.

[4] Zheng, Lianmin and Chiang, Wei-Lin and Sheng, Ying and Zhuang, Siyuan and Wu, Zhanghao and Zhuang, Yonghao and Lin, Zi and Li, Zhuohan and Li, Dacheng and Xing, Eric and others, 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

[5] Li, Xuechen and Zhang, Tianyi and Dubois, Yann and Taori, Rohan and Gulrajani, Ishaan and Guestrin, Carlos and Liang, Percy and Hashimoto, Tatsunori B, 2023. Alpacaeval: An automatic evaluator of instruction-following models. GitHub repository.

[6] Keisuke Sakaguchi and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi, 2019. An Adversarial Winograd Schema Challenge at Scale. arXiv:1907.10641.

[7] Peter Clark and Isaac Cowhey and Oren Etzioni and Tushar Khot and Ashish Sabharwal and Carissa Schoenick and Oyvind Tafjord, 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

A Examples for problem in SC-Math6

Below are examples of problems in SuperCLUE-Math6.

Initial Question ：
在家里安排春节聚餐，妈妈准备的食材是10个包子，每个包子可以供2个人吃，她还准备了5个鸡翅，每个鸡翅只能供1人吃，那么妈妈准备的食材最多可以供多少个人吃饭？

For the Spring Festival family dinner at home, mom prepared 10 steamed buns, each of which can be shared by 2 people. She also prepared 5 chicken wings, with each wing enough for 1 person. How many people in total can be served with the food prepared by mom?

Solution :

首先，计算一下包子一共能够供多少人食用。
妈妈准备了10个包子，每个包子可以供2个人吃，所以包子总共可以供10*2=<<10*2=20>>20个人吃。然后，计算一下鸡翅一共能够供多少人食用。
妈妈准备了5个鸡翅，每个鸡翅可以供1个人吃，所以鸡翅总共可以供5*1=<<5*1=5>>5个人吃。
最后，将包子和鸡翅能供食用的人数相加，就是妈妈准备的食材最多可以供多少个人吃饭。
包子和鸡翅一共可以供20+5=<<20+5=25>>25个人吃饭。

First, let's calculate how many people can eat the steamed buns.
Mom prepared 10 steamed buns, and each bun can be shared by 2 people, so in total the buns can serve 102=20 people.Next, let's calculate how many people can eat the chicken wings.Mom prepared 5 chicken wings, and each wing is enough for 1 person, so the chicken wings in total can serve 51=5 people.
Finally, by adding together the number of people who can be served with the buns and the chicken wings, we get the maximum number of people who can be served with the food mom prepared.
Together, the steamed buns and chicken wings can serve 20+5=25 people.

Final Answer: 25

Reasoning Steps: 3

Follow-up question ：如果突然来了3个客人，妈妈打算只做鸡翅，请问她还需要做几个鸡翅。

If three more guests suddenly arrive and mom plans to prepare only chicken wings, how many more chicken wings does she need to make?

Solution :

每个鸡翅可以供1个人吃，来了3个客人，3/1=<<3/1=3>>3个。

Each chicken wing is enough for one person. With 3 guests arriving, that would be 3/1=3 chicken wings needed.

Final Answer: 3

Reasoning Steps: 1

Figure 6: An example of a problem in SC-Math6

Initial Question ：
张伟计划在暑假的60天内，每天阅读30页的书。如果张伟在假期的前15天每天阅读40页的书，那么他在剩下的假期天数里，每天应阅读多少页的书才能达到他的阅读计划?

Zhang Wei plans to read 30 pages of a book every day during the 60 days of summer vacation. If Zhang Wei reads 40 pages each day for the first 15 days of the vacation, how many pages per day should he read during the remaining days of the vacation to meet his reading goal?

Solution :

首先，我们需要计算张伟整个暑假计划要阅读的总页数，60天乘以每天30页，得到：60*30=<<60*30=1800>>1800页。
然后，计算前15天他实际阅读了多少页，15天乘以每天40页，得到：15*40=<<15*40=600>>600页。
接着我们需要找出剩下的假期天数，总天数60天减去已经过去的15天，得到：60-15=<<60-15=45>>45天。
现在，我们知道剩下的天数里张伟还需要阅读多少页，就是总计划阅读页数减去前15天已经阅读的页数，得到：1800-600=<<1800-600=1200>>1200页。
最后，将剩下需要阅读的页面数1200页平均分配到剩下的45天里，计算每天应该阅读多少页：1200/45=<<1200/45=26.67>>26.67页。
因为我们需要的答案是正整数，所以将小数点后面的数四舍五入，得到每天应该阅读27页。

First, we need to calculate the total number of pages Zhang Wei plans to read during the entire summer vacation, which is 60 days times 30 pages per day, resulting in: 60*30=1800 pages.
Next, we calculate how many pages he actually read in the first 15 days, which is 15 days times 40 pages per day, resulting in: 15*40=600 pages.
Then we need to find out the number of remaining vacation days, which is the total of 60 days minus the 15 days that have passed, giving us: 60-15=45 days.
Now, we know how many more pages Zhang Wei needs to read in the remaining days, which is the total planned pages minus the pages already read in the first 15 days, resulting in: 1800-600=1200 pages.
Finally, we distribute the remaining 1200 pages to be read over the remaining 45 days, calculating how many pages should be read per day: 1200/45=26.67 pages.
Since we need the answer to be a whole number, we round the number after the decimal point and get that he should read 27 pages per day.

Final Answer: 27

Reasoning Steps: 5

Follow-up question ：如果张伟决定在剩下的假期里，每天阅读30页的书，他的总阅读数量将超过他原先的计划多少页？

If Zhang Wei decides to read 30 pages a day for the rest of the vacation, by how many pages will his total reading amount exceed his original plan?

Solution :

首先找出张伟剩余假期阅读的页数，他计划在剩余的45天每天阅读30页，所以这45天他将阅读45 * 30 =<<45 * 30 =1350>>1350页。
然后，我们将这1350页与他在前15天阅读的600页相加，也就是说他的总阅读数量将是1350 + 600 =<<1350 + 600 =1950>>1950页。
最后，我们将这个数量与他原计划的1800页相比，也就是说他的阅读数量将超过他原先的计划1950 - 1800 =<<1950 - 1800 =150>>150页。

First, we find out how many pages Zhang Wei will read during the remaining vacation days. He plans to read 30 pages each day for the remaining 45 days, which means he will read 45 * 30 = 1350 pages in those 45 days.
Next, we add these 1350 pages to the 600 pages he read in the first 15 days. This means his total reading amount will be 1350 + 600 = 1950 pages.
Finally, we compare this total with his original plan of 1800 pages. This means that his reading amount will exceed his initial plan by 1950 - 1800 = 150 pages

Final Answer: 150

Reasoning Steps: 3

Figure 7: An example of a problem in SC-Math6