

A Author Contributions

Data collection, annotation, and initial validation: Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Yikai Zhang, and Jiayi Lei collected the first version of the data, including some necessary manual annotation from PDF or Word documents. They then cross-validated the collected data. Junlei Zhang prompted GPT-4 to generate the explanation data, and then Yuzhen Huang, Yuzhuo Bai, and Junxian He manually revised the explanations.

Data processing: Yuzhen Huang did most of the data processing job including a lot of manual revisions, such as fixing typos/incorrect formats of the questions from PDF documents and ensuring all the \LaTeX notations can be correctly compiled. Jinghan Zhang did the data deduplication.

Evaluation: Yuzhen Huang tested ChatGPT and MOSS; Yuzhuo Bai tested ChatGLM; Junlei Zhang tested GLM-130B; Chuancheng Lv tested Bloomz-mt, LLaMA, Chinese-LLaMA, and Chinese-Alpaca; Yao Fu tested GPT-4 and the Claude model variants.

Website and submission system: Zhihao Zhu built the website and the online submission system.

Paper Writing: Yuzhen Huang, Yuzhuo Bai, and Junxian He wrote the main content of this paper, while other authors helped proofread.

Advising: Yao Fu, Maosong Sun, and Junxian He take advisor roles in this project. Junxian He is the main advisor, initializing and organizing this project.

B Detailed Stats of C-EVAL

Table 7 lists all the C-EVAL tasks and their broader categories, as well as the number of questions included in each task.

C Explanation Data Generation

We show an example of the automatic COT explanation generation via GPT-4 in Figure 5. We use five human-written question-explanation pairs to prompt GPT-4 to generate explanation giving question and its correct answer. The generated explanations are further revised manually to obtain the final explanations.

D Evaluation Prompts

We show the evaluation prompts of answer-only and chain-of-thought test in Figure 6 and Figure 7 respectively.

E Details of the models being evaluated

ChatGPT and GPT-4 are GPT-series models that are enhanced to follow human instructions and be more helpful, harmless and honest using Reinforcement Learning from Human Feedback. GPT-4 additionally enables image inputs and goes through well-designed post-training alignment process, as well as having a larger scale than most of the existing model. GPT-4 achieves human-level performance on various benchmark, and even scored to be the top 10% in some simulated exams.

Claude is the latest Anthropic-series LLM that also focuses on human intention alignment. Applying the constitutional AI approach (Bai et al., 2022), Claude manages to be both helpful and trustworthy. Claude-instant is the lighter version with less cost and faster inference of Claude.

BLOOMZ-mt (Muennighoff et al., 2022) is created by combining multitask prompted finetuning to the pretrained multilingual BLOOM model (Scao et al., 2022), using not only English prompts but also machine-translated prompts to match the language of multilingual tasks, and are found

Subject	Category	# Questions
Advanced Mathematics (高等数学)	STEM	197
College Chemistry (大学化学)	STEM	253
College Physics (大学物理)	STEM	200
College Programming (大学编程)	STEM	384
Computer Architecture (计算机组成)	STEM	219
Computer Network (计算机网络)	STEM	195
Discrete Mathematics (离散数学)	STEM	174
Electrical Engineer (注册电气工程师)	STEM	381
High School Biology (高中生物)	STEM	199
High School Chemistry (高中化学)	STEM	196
High School Mathematics (高中数学)	STEM	189
High School Physics (高中物理)	STEM	199
Metrology Engineer (注册计量师)	STEM	248
Middle School Biology (初中生物)	STEM	218
Middle School Chemistry (初中化学)	STEM	210
Middle School Mathematics (初中数学)	STEM	201
Middle School Physics (初中物理)	STEM	202
Operating System (操作系统)	STEM	203
Probability and Statistics (概率统计)	STEM	189
Veterinary Medicine (兽医学)	STEM	238
Business Administration (工商管理)	Social Science	339
College Economics (大学经济学)	Social Science	557
Education Science (教育学)	Social Science	304
High School Geography (高中地理)	Social Science	202
High School Politics (高中政治)	Social Science	200
Mao Zedong Thought (毛泽东思想和中国特色社会主义理论体系概论)	Social Science	248
Marxism (马克思主义基本原理)	Social Science	203
Middle School Geography (初中地理)	Social Science	125
Middle School Politics (初中政治)	Social Science	219
Teacher Qualification (教师资格)	Social Science	448
Art Studies (艺术学)	Humanities	336
Chinese Language and Literature (中国语言文学)	Humanities	237
High School Chinese (高中语文)	Humanities	202
High School History (高中历史)	Humanities	207
Ideological and Moral Cultivation (思想道德修养与法律基础)	Humanities	196
Law (法学)	Humanities	250
Legal Professional (法律职业资格)	Humanities	243
Logic (逻辑学)	Humanities	231
Middle School History (初中历史)	Humanities	234
Modern Chinese History (近代史纲要)	Humanities	240
Professional Tour Guide (导游资格)	Humanities	300
Accountant (注册会计师)	Other	497
Basic Medicine (基础医学)	Other	199
Civil Servant (公务员)	Other	481
Clinical Medicine (临床医学)	Other	227
Environmental Impact Assessment Engineer (环境影响评价工程师)	Other	317
Fire Engineer (注册消防工程师)	Other	318
Physician (医师资格)	Other	497
Plant Protection (植物保护)	Other	226
Sports Science (体育学)	Other	204
Tax Accountant (税务师)	Other	497
Urban and Rural Planner (注册城乡规划师)	Other	469

Table 7: Summary of all 52 subjects.

to be capable at task- and language-agnostic generalization. We evaluate the 176B version in our experiment.

LLaMA (Touvron et al., 2023) is a Transformer-architecture LLM that is trained on a mixture of several open sources. Applying several improvements on vanilla Transformers used by previous LLMs, and optimized to improve the training efficiency, LLaMA shows strong language abilities and can surpass models having 10x larger parameters than LLaMA. We evaluate the LLaMA-65B version in our experiment.

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

以下是中国关于初中物理的单项选择题，请根据以下题目和答案，生成一段正确，合理的答案解析。

The following are multiple-choice questions about middle school physics in China. Please generate a correct and reasonable explanation based on the following questions and answers.

下列属于可再生能源的是_____。

Which of the following is a renewable energy source?

A. 石油 B. 煤炭 C. 核燃料 D. 太阳能

A. petroleum B. coal C. nuclear fuel D. solar energy

答案：D

Answer: D

你生成的答案解析是什么？

What is your explanation generated?

答案解析：让我们一步一步思考，

1. (A) 石油是一种化石能源，是从地下深处开采出来的，不属于可再生能源。

2. (B) 煤炭也是一种化石能源，是从地下深处开采出来的，不属于可再生能源。

3. (C) 核燃料是指用于核反应堆中的燃料，包括铀、钚等，虽然可以通过核反应产生能量，但是核燃料是有限的，不属于可再生能源。

4. (D) 太阳能是指从太阳辐射中获取的能量，是一种可再生能源，因为太阳每天都会升起，不会因为使用而枯竭。

所以答案是D。

Explanation: Let's think step by step,

1. (A) Petroleum is a fossil fuel extracted from deep underground and is not a renewable energy source.

2. (B) Coal is also a fossil fuel extracted from deep underground and is not a renewable energy source.

3. (C) Nuclear fuel refers to the fuel used in nuclear reactors, including uranium, plutonium, etc. Although energy can be generated through nuclear reactions, nuclear fuel is limited and is not a renewable energy source.

4. (D) Solar energy refers to the energy obtained from solar radiation and is a renewable energy source, because the sun rises every day and will not be exhausted due to use.

Therefore, the answer is C.

Figure 5: An example of generating explanations via GPT-4. The red text is the autocompleted response from model, while the preceding text is the user-inputted prompt. We indicate English translation below the corresponding Chinese text for each paragraph.

以下是中国关于临床医学的单项选择题，请选出其中的正确答案。

The following are multiple-choice questions about clinical medicine in China. Please select the correct answer.

男性，65岁，昏迷10小时。血气分析：pH 7.26，PaCO₂ 82mmHg，PaO₂ 45mmHg。下列哪项处理不宜_____

A 65-year-old male patient has been in a coma for 10 hours. Blood gas test shows a pH of 7.26, PaCO₂ 82mmHg, and PaO₂ 45mmHg. Which of the following treatments is not appropriate? _____

A. 机械通气 B. 呼吸兴奋剂 C. 支持疗法 D. 高浓度给氧

A. mechanical ventilation B. respiratory stimulants C. supportive therapy D. high concentration oxygen therapy

答案：D

Answer: D

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

硅尘可导致硅沉着病，其中致病力最强的硅尘颗粒直径为_____

Silicon dust can cause silicosis, and the most pathogenic silicon dust particles have a diameter of _____.

A. > 5μm B. 4 ~ 5μm C. 1 ~ 2μm D. 2 ~ 3 / μm

答案：C

Answer: C

Figure 6: An example of few-shot evaluation in answer-only scenarios, while zero-shot evaluation is similar by removing the exemplars. The red text is the autocompleted response from model, while the preceding text is the inputted prompt. We indicate English translation below the corresponding Chinese text.

523 **GLM-130B and ChatGLM-6B** are based on the General language model (GLM) structure that
524 gain benefits from its bidirectional attention advantage. By using alterable number and length of
525 blanks, GLM can adapt to various tasks. GLM-130B is a bilingual pre-trained GLM that utilizes self-
526 supervised learning and multitask instruction pretraining. GLM-130B also realized INT4 quantization

以下是中国关于计算机组成考试的单项选择题，请选出其中的正确答案。
The following are multiple-choice questions about computer architecture in China. Please select the correct answer.

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

设寄存器R的内容(R)=1500H，内存单元1500H的内容为2500H，内存单元2500H的内容为3500H，PC的值为4500H，采用相对寻址方式，有效地址为2500H的操作数是_____。
If the content of register R is set to 1500H, the content of memory address 1500H is 2500H, the content of memory address 2500H is 3500H, and the value of the program counter (PC) is 4500H, using the relative addressing mode, the operand at the effective address of 2500H is _____

A. 1500H B. 2500H C. 3500H D. 4500H

答案：让我们一步一步思考，
1. 根据题目中的信息，内存单元1500H的内容为2500H，内存单元2500H的内容为3500H，表示当前指令的地址为2500H-1500H=1000H。
2. 根据相对寻址方式，有效地址为2500H的操作数需要加上当前指令地址1000H才能得到实际的操作数地址，即3500H。
所以答案是C。

Answer: Let's think step by step,
1. According to the information given in the question, the content of memory unit 1500H is 2500H, and the content of memory unit 2500H is 3500H. This indicates that the current instruction address is 2500H-1500H=1000H.
2. According to the relative addressing mode, in order to obtain the actual operand address of the effective address with the value of 2500H, the current instruction address of 1000H needs to be added to it, resulting in the actual operand address of 3500H.
Therefore, the answer is C.

Figure 7: An example of few-shot evaluation in chain-of-thought scenarios, while zero-shot evaluation is similar by removing the exemplars. The red text is the autocompleted response from model, while the preceding text is the inputted prompt. We indicate English translation below the corresponding Chinese text.

with little to no quality degradation that significantly accelerate its inference efficiency. ChatGLM-6B is a lightweight conversational version of the GLM family that has been specially optimized for Chinese contexts. ChatGLM-6B also applies quantization so it can be deployed with a consumer-grade graphic memory requirement as little as 6GB. We evaluate on the fp16 settings for both two models in our experiment.

Chinese-LLaMA is an adaptation of original LLaMA into Chinese language environments. Chinese-LLaMA expands the original LLaMA by adding 20K Chinese tokens into its vocabulary, and is secondarily pre-trained and instruction fine-tuned on Chinese data. We evaluate Chinese-LLaMA-13B in our experiment, the largest Chinese-LLaMA variant.

Chinese-Alpaca is based on the Chinese-LLaMA checkpoint that is further tuned on Chinese instruction tuning data. We evaluate Chinese-Alpaca-13B in our experiment, the largest Chinese-Alpaca variant.

MOSS is the first open-source Chinese LLM that matches ChatGPT on both the training scale and alignment techniques. MOSS is initialized with CodeGen(Nijkamp et al., 2022), being pretrained on 100B Chinese tokens and 20B English tokens, and has further integrated supervised fine-tuning and preference model, as well as plugin augmentation, but not all the version are publicly available. We evaluate the moss-moon-003-sft version in our experiment.

F Breakdown of Model Performance

Table 8 and Table 9 show the detailed breakdown of accuracy per subject of four representative models in zero- and five-shot settings respectively, while we refer the readers to our website leaderboard <https://cevalbenchmark.com/static/leaderboard.html> for a detailed breakdown of results for all models.

G Results on Validation Split

Since we do not publicly release the labels for our test split, we provide the average accuracy on the validation split as a reference for developers. The validation split comprises a total of 1346 questions,

Subject	GPT-4	ChatGPT	GLM-130B	Claude-instant-v1.0
Advanced Mathematics	48.6	38.2	28.9	36.4
College Chemistry	54.0	36.2	29.0	25.9
College Physics	50.6	34.1	29.5	33.5
College Programming	72.2	56.1	33.3	40.6
Computer Architecture	73.1	62.7	38.3	42.5
Computer Network	74.9	60.8	36.3	44.4
Discrete Mathematics	62.1	37.9	34.6	28.8
Electrical Engineer	48.7	38.9	31.6	32.7
High School Biology	69.1	49.1	37.7	40.0
High School Chemistry	55.2	45.3	36.0	39.5
High School Mathematics	41.0	28.3	34.9	24.7
High School Physics	65.1	36.6	26.9	40.0
Metrology Engineer	68.0	58.4	49.8	48.4
Middle School Biology	88.0	68.2	51.6	46.4
Middle School Chemistry	82.7	63.2	45.9	47.0
Middle School Mathematics	65.5	40.1	32.2	34.5
Middle School Physics	80.9	61.2	41.6	48.9
Operating System	79.3	70.4	42.5	46.4
Probability and Statistics	50.0	36.7	25.9	27.7
Veterinary Medicine	75.7	56.7	47.6	43.8
Business Administration	63.1	48.2	43.2	40.2
College Economics	67.6	47.3	38.6	40.8
Education Science	67.4	54.8	52.6	43.0
High School Geography	73.6	55.6	51.1	42.7
High School Politics	74.4	50.0	45.5	37.5
Mao Zedong Thought	74.9	56.6	67.6	49.3
Marxism	77.7	70.9	69.3	62.0
Middle School Geography	78.7	57.4	58.3	49.1
Middle School Politics	85.5	69.4	61.1	57.0
Teacher Qualification	84.5	69.9	70.7	54.9
Art Studies	57.4	47.7	49.3	33.2
Chinese Language and Literature	61.2	51.7	46.4	35.9
High School Chinese	38.8	37.6	25.8	31.5
High School History	66.5	48.4	48.4	41.2
Ideological and Moral Cultivation	73.8	65.1	59.3	58.1
Law	56.1	42.1	37.6	34.8
Legal Professional	55.3	40.5	35.8	36.7
Logic	62.3	38.7	34.3	36.3
Middle School History	83.6	65.2	65.2	48.8
Modern Chinese History	63.7	46.2	59.9	42.5
Professional Tour Guide	68.8	53.4	63.2	35.0
Accountant	63.4	46.3	39.7	38.1
Basic Medicine	78.9	61.1	46.9	39.4
Civil Servant	62.2	45.5	45.2	36.1
Clinical Medicine	70.5	51.0	43.5	38.0
Environmental Impact Assessment Engineer	59.4	52.0	45.6	41.3
Fire Engineer	45.0	41.5	35.8	29.4
Physician	75.6	59.4	44.7	41.5
Plant Protection	71.4	56.3	49.7	48.2
Sports Science	70.0	53.9	44.4	40.6
Tax Accountant	56.0	38.4	33.9	38.1
Urban and Rural Planner	59.1	49.0	43.5	38.0

Table 8: Zero-shot answer only accuracy per subject.

with each subject contributing fewer than 30 validation questions on average. Therefore, tracking accuracy on a specific subject may not yield significant insights. Instead, we report the average answer-only accuracy across all subjects in Table 10. The average validation accuracy closely mirrors the average test accuracy as presented in Table 3 and Table 4. Additionally, the model ranking on the validation split broadly corresponds to that on the test split. These observations suggest that developers may use the average validation accuracy as a good indicator for expedited development processes.

Subject	GPT-4	ChatGPT	ChatGLM	Claude-instant-v1.0
Advanced Mathematics	49.7 / 53.8	48.0 / 29.5	3.5 / 23.1	39.9 / 41.6
College Chemistry	59.4 / 55.8	39.7 / 33.9	17.0 / 28.6	33.5 / 32.1
College Physics	50.6 / 60.2	35.8 / 35.2	14.8 / 30.1	35.8 / 31.3
College Programming	78.1 / 77.5	57.9 / 57.9	18.7 / 24.0	48.5 / 50.3
Computer Architecture	74.6 / 75.6	68.9 / 63.7	30.1 / 27.5	52.3 / 50.8
Computer Network	77.8 / 77.2	58.5 / 59.6	36.3 / 28.1	46.8 / 53.8
Discrete Mathematics	66.7 / 62.7	43.1 / 30.7	17.0 / 26.1	32.7 / 33.3
Electrical Engineer	49.9 / 56.0	39.5 / 44.0	23.9 / 31.3	31.3 / 35.7
High School Biology	72.6 / 70.3	53.7 / 48.0	28.0 / 26.9	44.6 / 38.9
High School Chemistry	52.3 / 55.8	52.9 / 36.0	19.8 / 28.5	39.0 / 37.8
High School Mathematics	38.0 / 42.8	34.3 / 37.3	4.2 / 24.1	24.1 / 32.5
High School Physics	69.1 / 61.1	43.4 / 34.9	10.9 / 21.7	44.0 / 26.9
Metrology Engineer	70.3 / 71.2	63.0 / 59.8	41.1 / 48.4	50.7 / 55.3
Middle School Biology	91.2 / 86.5	67.7 / 63.5	30.7 / 38.0	55.2 / 57.3
Middle School Chemistry	81.1 / 71.4	71.4 / 60.0	36.8 / 42.2	58.4 / 59.5
Middle School Mathematics	62.7 / 66.7	44.6 / 42.4	3.4 / 20.3	28.8 / 33.3
Middle School Physics	81.5 / 78.7	66.9 / 57.9	20.2 / 41.0	52.8 / 53.4
Operating System	82.1 / 80.4	71.5 / 60.3	40.8 / 27.4	57.0 / 57.5
Probability and Statistics	53.6 / 62.0	33.7 / 42.8	27.7 / 26.5	34.9 / 31.9
Veterinary Medicine	80.0 / 80.0	64.3 / 58.1	39.5 / 35.2	51.0 / 53.3
Business Administration	67.1 / 65.4	52.8 / 46.8	32.6 / 34.9	44.2 / 42.9
College Economics	71.4 / 74.4	52.5 / 51.3	22.1 / 30.0	47.7 / 50.3
Education Science	68.5 / 69.3	59.6 / 54.4	37.8 / 33.0	51.1 / 51.9
High School Geography	74.7 / 70.2	58.4 / 55.1	29.2 / 38.2	52.8 / 49.4
High School Politics	77.8 / 69.3	52.3 / 51.1	25.0 / 29.6	38.1 / 39.8
Mao Zedong Thought	79.5 / 79.9	60.7 / 63.5	46.6 / 50.7	47.0 / 50.2
Marxism	81.6 / 82.7	71.5 / 70.4	40.2 / 49.2	66.5 / 63.7
Middle School Geography	83.3 / 83.3	60.2 / 56.5	33.3 / 44.4	62.0 / 52.8
Middle School Politics	87.1 / 85.5	74.1 / 67.4	41.5 / 40.9	60.6 / 64.3
Teacher Qualification	85.0 / 85.0	75.7 / 66.9	48.9 / 49.1	68.4 / 62.2
Art Studies	65.1 / 66.1	49.7 / 52.3	34.6 / 36.9	35.6 / 38.9
Chinese Language and Literature	61.2 / 67.0	50.2 / 51.7	32.5 / 37.3	41.2 / 44.0
High School Chinese	37.6 / 39.3	36.0 / 27.5	26.4 / 19.1	30.9 / 27.0
High School History	68.1 / 68.1	54.4 / 45.1	35.2 / 40.1	52.7 / 40.7
Ideological and Moral Cultivation	77.3 / 77.3	66.9 / 68.0	48.3 / 52.3	62.8 / 63.4
Law	60.6 / 54.8	43.9 / 34.8	23.5 / 29.9	38.0 / 33.0
Legal Professional	54.4 / 48.4	44.7 / 32.6	27.0 / 26.5	39.5 / 30.2
Logic	60.3 / 63.2	37.7 / 35.8	32.8 / 31.4	38.7 / 35.8
Middle School History	84.5 / 86.5	62.8 / 63.8	48.3 / 52.7	58.5 / 52.2
Modern Chinese History	68.9 / 67.0	52.8 / 51.4	36.3 / 39.2	44.8 / 44.8
Professional Tour Guide	71.8 / 70.3	61.3 / 62.0	44.0 / 51.9	43.6 / 44.0
Accountant	64.6 / 61.9	51.7 / 42.0	23.9 / 32.7	47.2 / 41.3
Basic Medicine	80.6 / 78.3	61.1 / 60.6	31.4 / 33.7	49.7 / 45.1
Civil Servant	62.5 / 59.0	46.9 / 43.1	24.5 / 29.6	42.0 / 41.3
Clinical Medicine	76.0 / 73.5	55.0 / 53.0	34.5 / 32.0	36.5 / 37.5
Environmental Impact Assessment Engineer	63.7 / 59.8	50.9 / 49.1	37.7 / 35.6	47.3 / 47.3
Fire Engineer	50.0 / 49.6	42.6 / 37.9	28.4 / 32.3	36.5 / 36.2
Physician	76.8 / 76.1	63.7 / 54.6	34.8 / 37.7	50.6 / 46.5
Plant Protection	78.9 / 80.4	65.8 / 57.8	43.2 / 36.2	56.3 / 50.3
Sports Science	72.2 / 70.0	58.3 / 50.6	42.2 / 41.1	45.6 / 46.7
Tax Accountant	58.0 / 58.2	42.0 / 35.7	16.3 / 30.9	42.7 / 31.8
Urban and Rural Planner	63.2 / 65.6	52.2 / 49.3	36.8 / 37.3	45.5 / 42.3

Table 9: Five-shot accuracy per subject. We report both the answer only (AO) and chain-of-thought (COT) accuracy in an AO / COT format.

558 H Compute and Resources Used for Evaluation

559 During our experiments to evaluate different LLMs on C-EVAL, we utilize a cluster with 8 A100-
560 80GB GPUs to run the inference for models with released weights, such resources are required due
561 to deploying three large models – Bloomz-mt (176B), LLaMA-65B, and GLM-130B. In most cases
562 the inference on C-EVAL is finished within one day. For models with API access, we just run the
563 inference with CPUs which finishes mostly within one day as well.

Model	Zero-shot	Five-shot
GPT-4	66.7	69.9
Claude-v1.3	52.1	55.5
ChatGPT	50.8	53.5
Bloomz-mt	45.9	38.0
GLM-130B	44.2	40.8
Claude-instant-v1.0	43.2	47.4
ChatGLM-6B	39.7	37.1
LLaMA-65B	38.6	39.8
MOSS	35.1	28.9
Chinese-Alpaca-13B	32.0	27.2
Chinese-LLaMA-13B	29.4	33.1

Table 10: Average accuracy on the validation split in the answer-only setting.