The Performance of AI in China's National Civil Service Exam (NCSE)

Guanlin Li[1], Kaijie Yang[1], Chuwei Ma[2], Jimao Mo[3], Yuanzhe Cai[*], and Hongqiang Ding[*]

[1]Guanlin Li, Shenzhen The Second People's Hospital, Shenzhen, 518037, China
[*]Yuanzhe Cai, Shenzhen Technology University, Shenzhen, 518118, China
[*]Hongqiang Ding, The Chinese University of Hong Kong, Shenzhen, Shenzhen, 518172, China
[1]info@lglws.com
[+]Feijuan Huang

ABSTRACT

This study explores the potential of large language models, specifically ChatGPT-4, in the context of high-stakes exams by evaluating its performance on the Chinese National Civil Service Exam (NCSE). With the rapid advancements in artificial intelligence, understanding how well AI can perform on exams designed to test comprehensive cognitive skills has become increasingly relevant. Inspired by previous studies that assessed ChatGPT's performance on the U.S. Bar Exam, we aimed to extend this evaluation to measure its "intelligence quotient" through the NCSE, a more diverse and demanding benchmark than traditional IQ tests.

The NCSE includes the Administrative Aptitude Test (AAT) and Argumentative Essay Writing (AEW), covering a range of cognitive domains such as logical reasoning, reading comprehension, quantitative analysis, and memory-based questions. This structure allows for a holistic assessment of ChatGPT-4's capabilities, including language processing, logical reasoning, mathematical calculations, and data analysis. Results indicate that ChatGPT-4 demonstrates considerable strengths in natural language comprehension and structured writing but reveals notable limitations in visual recognition and logical reasoning, especially in tasks requiring abstract thought and multi-step problem-solving.

Using the Fenbi grading platform for evaluation, ChatGPT-4's scores were compared against average human test-takers, providing a reliable benchmark of the model's performance relative to human standards. The study shows that ChatGPT-4 exceeds the human average in certain areas, yet falls short of the highest human scores, underscoring the need for continued development in AI's logical reasoning and response regulation capabilities.

Our findings suggest that with ongoing advancements, AI models like ChatGPT-4 could potentially serve as valuable tools in academic and professional assessments. The NCSE offers a robust framework for evaluating AI's practical cognitive skills, marking an innovative step in redefining intelligence metrics for AI in complex, real-world scenarios. This research contributes to the growing body of knowledge on AI assessment, setting the foundation for future applications and improvements in AI-driven evaluation systems.

Please note: Due to slight differences in semantics between Chinese and English, translating Chinese titles into English can sometimes lack emotional nuance, leading to differences in understanding. Everything should be based on the original meaning in Chinese.

Introduction

In today's era of rapid advancements in artificial intelligence, new and increasingly complex AI models are emerging at an incredible rate, each pushing the boundaries of human and machine intelligence. Amid these innovations, a key question arises: How accurate and reliable are AI models, like ChatGPT, in answering highly

specialized, domain-specific questions? This question has sparked the interest of researchers worldwide, with studies evaluating ChatGPT's potential across various fields, including a notable study in the United States that assessed ChatGPT's performance on the U.S. Bar Exam[1]. The results were promising, suggesting the model's competence in handling complex legal reasoning and case-based knowledge.

Inspired by this research, we aimed to investigate the "intelligence quotient" (IQ) of ChatGPT when benchmarked against traditional IQ assessments. To do this, we initially experimented with two widely regarded IQ tests: the Mensa and Raven's Progressive Matrices tests, which are commonly used to measure abstract reasoning, pattern recognition, and logical problem-solving skills. However, our findings revealed certain limitations: ChatGPT showed some deficiencies in visual pattern recognition and logical reasoning—areas critical for IQ testing that involve spatial and abstract thinking, such as identifying visual sequences or deducing solutions from intricate patterns. These limitations led us to explore alternative assessments that could offer a more comprehensive evaluation of the model's cognitive abilities.

This led us to an innovative idea: testing ChatGPT using China's National Civil Service Exam (NCSE). The rationale behind this choice lies in the exam's unique structure and broader range of question types. Unlike traditional IQ tests, the NCSE comprehensively covers various cognitive domains, including logical reasoning, reading comprehension, quantitative analysis, and memory-based questions. Additionally, it includes question formats that extend beyond those in standard IQ tests, providing a richer framework for evaluating the model's reasoning, memory, and problem-solving abilities in both straightforward and complex scenarios. By replacing traditional IQ tests with the NCSE, we're not only able to assess ChatGPT's skills across a broader spectrum of cognitive tasks but also gauge its performance on challenges that more closely resemble real-world applications of knowledge.

In summary, using the NCSE to test ChatGPT provides a novel and effective approach for measuring the model's intelligence. The diversity and depth of the NCSE's questions enable a robust assessment of ChatGPT's abilities in processing and reasoning with information, making this method an optimal choice for evaluating AI intelligence in today's complex, information-rich environment. This study, therefore, represents a pioneering effort to redefine the metrics of AI assessment, contributing to a broader understanding of AI intelligence in practical and applied contexts.

China's National Civil Service Exam (NCSE)

Most government departments in China require candidates to pass the Civil Service Examination (often called the National Exam) as a prerequisite for public sector positions[2]. This exam is a series of challenging tests designed to assess candidates' comprehensive abilities, administrative skills, and problem-solving capabilities[3]. Successful candidates are expected to demonstrate strong analytical skills in addressing complex policy issues, effectively manage public affairs, and apply laws and regulations proficiently.

In April 1995, China's former Ministry of Personnel issued a notification standardizing the content, level, and criteria of the recruitment exam for civil servants, based on Article 17, Clause 2 of the Interim Provisions: "Public subjects are uniformly determined by the personnel department of the State Council." This notification introduced a unified framework for exam content, which included two main public subjects: General Knowledge and the Administrative Aptitude Test (AAT). Experts and experienced personnel were assigned by the Ministry of Personnel to compile materials for these public subjects. Since the establishment of the National Civil Service Bureau in 2008, the public subject exams have remained largely unchanged, focusing on the AAT and Essay Writing, with adjustments in depth and complexity based on administrative levels[4].

The National Exam primarily consists of a written examination and an interview. The written portion is

divided into two main sections: the Administrative Aptitude Test (AAT) and Essay Writing. The AAT measures candidates' logical reasoning, data analysis, and language comprehension abilities, while the Essay Writing section requires an in-depth analysis of public policy or current social issues, with candidates expected to propose solutions.

China's AAT is an ideal framework for evaluating large AI models due to its diverse and highly standardized structure. The AAT includes various fields such as verbal reasoning, logical deduction, quantitative relations, data analysis, and general knowledge, requiring candidates to possess a broad range of skills. These question types cover a range of tasks—including natural language processing, logical reasoning, mathematical calculations, and data analysis—that emphasize language comprehension, rigorous reasoning, and a wide breadth of knowledge. Thus, the AAT allows for a multidimensional assessment of AI model performance, revealing strengths and limitations in handling varied tasks.

Furthermore, the AAT's strict time constraints require candidates to process information efficiently and make decisions quickly. For AI models, this testing environment assesses response speed and information-processing efficiency under time pressure. The test's complex reasoning and multi-step questions effectively evaluate an AI's reasoning ability, logical consistency, and precision. Meanwhile, the quantitative relations and data analysis sections test a model's mathematical reasoning and data interpretation skills, broadening the scope of evaluation in terms of computational and data comprehension capabilities.

Testing AI models with the AAT can yield valuable insights. First, the test outcomes contribute to evaluating a model's comprehensive performance across multiple domains, enabling a systematic understanding of its strengths in natural language processing, logical reasoning, and mathematical operations. Second, the AAT can help identify a model's strengths and weaknesses across various question types, providing data to support further model optimization. Additionally, the strict time limits in the AAT can assess model performance under high-pressure conditions, gauging its efficiency in information filtering, reasoning, and rapid response. Ultimately, this test allows us to evaluate the model's practicality and versatility, particularly in complex task scenarios that require combined abilities.

In this study, we tested several large AI models using the AAT's modules, primarily covering general knowledge, verbal comprehension and analysis, quantitative relations, logical reasoning, and data analysis, with a detailed score distribution shown in Table 1

CNCSE Component	Total CNCSE Points	Questions	Times
General Knowledge	10 Points	20	2 Hours
Words Understanding and Expression	30 Points	40	
Quantitative Relations	15 Points	15	
Sequitur	25 Points	35	
Data Analyzed	20 Points	20	

Table 1. Structure of the Administrative Aptitude Test (AAT) for the 2022 National Civil Service Exam - Administrative Law Enforcement Paper

Methodology

1. Data Type

The Administrative Aptitude Test (AAT) section of the National Civil Service Exam includes five distinct components: General Knowledge, Semantic Comprehension and Analysis, Quantitative Relations, Logical Reasoning, and Data

Analysis. To evaluate the performance of various GPT models on the AAT, we gathered relevant materials from each of these areas. Specifically, we obtained the full set of questions from the 2022 National Civil Service Administrative Aptitude Test, available on the official National Civil Service Exam website. The 2022 test consists of 130 multiple-choice questions covering a wide range of topics, including language, mathematics, logic, history, literature, and science.

2. Method

We used a three-step process to collect data for this study. First, we collected three questions from each participant. Second, we gathered responses from various GPT models. Third, we evaluated the responses provided by each model and calculated scores. Below, we describe these three steps in detail.

First, we manually collected the full set of questions from the 2022 National Civil Service Administrative Aptitude Test, available on the official National Civil Service Exam website.

Second, we submitted each question to different GPT models through their respective Question and Answer interfaces. No preset prompts were used; only the original question was entered, and each model was asked to provide a response. Once the response was received, we saved screenshots of the answer along with the inquiry record and logged the results in a database.

Finally, we scored each model' s responses in the database based on the correct answers, producing the final results.

Our study has several limitations to consider. First, ChatGPT is highly sensitive to prompt phrasing. If a response showed a misunderstanding of the question, we did not provide clarification. Additionally, when the model receives the same prompt repeatedly, it may produce different responses. We recorded only the first answer and shared it with participants. It' s unclear whether alternative responses would have been better or worse[5]. The human test-taker data referenced in the text is based on responses from the Chalk Civil Service Exam App' s big data.

Question Type Analysis

1. General Knowledge

First, general knowledge questions typically cover a broad range of fundamental information and everyday topics, requiring the model to have a solid grasp of common knowledge about the world. These questions tend to focus less on specialized or in-depth academic knowledge and more on basic, general knowledge across various areas, such as daily life, history, and culture. Consequently, general knowledge questions are often shorter and easier to understand. Compared to other types of questions, like verbal comprehension and logical reasoning, general knowledge questions are generally more straightforward for large models to understand and process.

1) ChatGPT 4.0: Scored 16 out of 20 on these 20 questions, performing quite well. It achieved a high score of 6/7 in the legal domain, indicating a strong understanding of legal knowledge;

2) Gemini: Scored 12/20, an average performance. In technology-related questions, it scored only 2/5, showing room for improvement in technical knowledge;

3) Copilot: Scored 8/20, performing poorly overall. It only managed to pass in certain areas like geography and law but scored low in technical and scientific knowledge;

4) Coze: Scored 16/20, tying with ChatGPT 4.0. It excelled in the legal domain, achieving a perfect score of 7/7;

5) ERNIE Bot: Scored the highest with 18/20, showing strong performance across multiple areas, including technology, humanities, and law, demonstrating broad knowledge coverage and depth of understanding;

Question Type	GPT-4o	Gemini	Copilot	Coze	ERNIE BOT	Human Avg.
Political	5/5	2/5	2/5	3/5	5/5	52.68%
Economic	4/5	3/5	2/5	5/5	3/5	34%
Technology	5/5	4/5	5/5	5/5	4/5	41.53%
Humanities	1/1	1/1	1/1	1/1	1/1	47.45%
Geography	0/1	0/1	0/1	0/1	1/1	38.40%
Laws General	2/2	1/2	1/2	2/2	2/2	38.89%

Table 2. Accuracy of Models on General Knowledge Questions in the Chinese Civil Service Exam' s Administrative Aptitude Test

After a comparative analysis of these models' scores, it is evident that there are significant differences in each model' s ability to handle general knowledge questions:

1) ERNIE Bot performed the best with a score of 18/20, showing high levels of understanding and answering ability across almost all areas, especially in technology and law.

2) ChatGPT 4.0 and Coze both scored 16/20, reflecting strong performance across multiple knowledge domains. Notably, Coze achieved a perfect score in legal questions, highlighting its expertise and depth in that field.

3) Gemini and Copilot scored relatively lower, with 12/20 and 8/20, respectively. This may indicate a lack of in-depth knowledge in certain fields or a need for optimization in answering strategies.

This evaluation clearly reveals the strengths and weaknesses of each AI model in handling general knowledge. For users, choosing the most suitable model according to their needs is crucial. For instance:

1) If users require accurate answers to legal or technical questions, ERNIE Bot is evidently the best choice.

2) For scenarios that require strong performance across multiple knowledge domains, ChatGPT 4.0 and Coze may be more suitable options.

3) For applications with specific cost and performance requirements, understanding the weaknesses of each model is equally important; for example, Gemini and Copilot show weaker performance in some knowledge areas and may need further adjustment and training.

Among general knowledge questions, legal knowledge and geographic knowledge are two typical categories, with consistent performance across the major models from Table 2.

Analysis of Typical Questions:

Example 1: On December 8, 2020, President Xi Jinping and Nepal' s President Bhandari exchanged messages and jointly announced the elevation measurement of Mount Everest. Regarding this measurement process, which of the following statements is incorrect?

A The elevation measurement started from a baseline at sea level and followed the standard procedures.

B The BeiDou Satellite Navigation System was used for high-precision positioning.

C This was the first time that aerial gravity measurement was conducted on the northern side of Mount Everest.

D The measurement of snow depth on the summit of Mount Everest was conducted using ultrasound detection.

Answer Analysis: This question tests knowledge of geography and national conditions.

1. **Option A is correct.** To measure the height of Mount Everest, an elevation control network was set up around Mount Everest and surrounding areas, conducting leveling measurements. Starting from a national first-class leveling point in Shigatse, Tibet, surveyors used precision leveling instruments to transfer the Yellow Sea elevation datum value step-by-step to the base of Mount Everest, ultimately obtaining an accurate height measurement.

2. **Option B is correct.** The 2020 Mount Everest elevation measurement achieved several breakthroughs, with more comprehensive and advanced technology employed. This included the integration of high-precision positioning via the BeiDou Satellite Navigation System, aerial gravity measurement, remote sensing, real-world 3D modeling, and centimeter-level quasi-geoid refinement.

3. **Option C is correct.** The average altitude of the Mount Everest region is over 5,000 meters, with extremely complex terrain, making it impossible to conduct ground gravity measurements in most areas. Gravity data is sparse, with many areas lacking gravity information. This measurement was the first in the world to conduct aerial gravity measurement on the northern side of Mount Everest, addressing gaps in gravity data and improving the accuracy of the elevation reference surface in the Mount Everest region.

4. **Option D is incorrect.** Snow depth radar primarily uses antennas to emit and receive high-frequency electromagnetic waves to detect ground snow depth. Snow depth radar observations were used to measure the thickness of the ice and snow layer at the summit of Mount Everest, and this value was subtracted from the summit' s snow surface elevation to determine the rock surface elevation at the summit.

This question requires identifying the incorrect statement; therefore, the correct answer is D.

Question Analysis:

This is a classic example question, where all models except GPT answered incorrectly, with most errors concentrated between options B and C.

Many large models tend to make mistakes on topics like the elevation measurement of Mount Everest. The underlying reasons can be understood through several key factors:

Firstly, information sources and knowledge updates are crucial factors. Large models are trained on vast amounts of pre-existing data, sourced from internet text, public documents, and more. Although this data is extensive, it does not guarantee that information in every domain is the latest or most accurate. For instance, new technologies used in the elevation measurement of Mount Everest, such as the BeiDou Satellite Navigation System and high-precision snow depth radar, might not have been fully covered or correctly labeled during training. Consequently, if the model lacks updated information on these recent advancements, it may rely on outdated or incomplete knowledge when responding to related questions.

Secondly, language models sometimes struggle with capturing fine details. Multiple-choice questions and their options often test a nuanced understanding of details. For instance, option D in the question mentions "ultrasound detection," a technical term that can be confusing. Ultrasound is widely used in other fields, such as medical imaging. If the model has formed a strong association with "ultrasound" through extensive training data without understanding its specific application in this context, it may misinterpret or make an incorrect judgment.

Analysis of Typical Questions:
Example 2: According to the "Data Security Law of the People' s Republic of China," which has been in effect since September 1, 2021, which of the following statements is incorrect?

A The Ministry of Industry and Information Technology is responsible for coordinating national data security policies and decisions.

B Provincial-level and higher governments should incorporate digital economic development into their local economic and social development plans.

C The state establishes a data classification and grading protection system, implementing classified and graded protection of data.

D Institutions providing intermediary services for data transactions should require data providers to specify the source of the data.

Answer Analysis: This question tests knowledge of legal regulations.

1. **Option A is incorrect.** According to Article 10 of the "Production Safety Law of the People' s Republic of China," "The Emergency Management Department of the State Council is responsible for the comprehensive supervision and management of production safety across the country according to this law. Local emergency management departments at or above the county level shall implement comprehensive supervision and management of production safety within their administrative areas according to this law. Relevant departments, such as transportation, housing and urban-rural development, water resources, and civil aviation, under the State Council, are responsible for supervising and managing production safety within their respective scopes of duties according to this law and other relevant laws and regulations. Local government departments at or above the county level are similarly responsible for the supervision of production safety in relevant industries and areas. For emerging industries or areas where safety supervision duties are unclear, the local government at or above the county level designates a supervisory department according to similar business principles. Departments with safety supervision duties must coordinate, share information, and use resources in common to strengthen safety supervision by law." This clarifies that the Emergency Management Department is

responsible for overall supervision of production safety, while other departments such as transportation, housing and construction, water resources, and civil aviation supervise within their respective scopes.

2. **Option B is incorrect.** According to Article 5 of the "Production Safety Law," "The primary responsible person of a production and operation unit is the first person accountable for the unit' s production safety and is fully responsible for its safety work. Other persons are responsible for safety within their respective scopes of duty." This means that the person in charge of production within the company is accountable for production safety within their area of responsibility.

3. **Option C is incorrect.** According to the second clause of Article 74 of the "Production Safety Law," "If a major accident hazard or a major accident occurs due to a violation of safety regulations, causing harm to national or public interests, the People' s Procuratorate may initiate a public interest lawsuit according to relevant provisions of the Civil Procedure Law and the Administrative Procedure Law."

4. **Option D is correct.** According to Article 114 of the "Production Safety Law," "If a production safety accident occurs, the Emergency Management Department shall impose fines on the responsible production and operation unit in accordance with the following standards, in addition to requiring them to bear corresponding compensation or other responsibilities: (1) For a general accident, a fine of between 300,000 and 1 million yuan; (2) For a major accident, a fine of between 1 million and 2 million yuan; (3) For a significant accident, a fine of between 2 million and 10 million yuan; (4) For an especially major accident, a fine of between 10 million and 20 million yuan. If the accident has extremely severe consequences and a particularly negative impact, the Emergency Management Department may impose a fine up to five times the amounts specified above on the responsible production and operation unit." Therefore, in cases of extremely serious production safety accidents with particularly negative consequences, the maximum fine imposed by the Emergency Management Department can reach five times 20 million yuan, totaling up to 100 million yuan.

Depend on the previous analyzed, the answer is D.

Question Analysis:
This is a typical legal knowledge question. Compared to the geography knowledge question above, legal knowledge questions are more straightforward and concise, with a very clear focus, leaving little room for misunderstanding by large models. In this type of question, most models perform exceptionally well. Only the Copilot model performed poorly, but this model has also shown poor performance in other question types, likely due to its low recognition accuracy in national civil service exam questions. It lacks basic question recognition capabilities and struggles to capture the main logic within the question.

For instance, in the question above, all large models answered correctly. This could be attributed to the fact that the question stem provides a very clear and specific scope, with time information and names presented concisely, allowing the models to easily understand the question' s intent and retrieve the required knowledge. Furthermore, there are no complex implications—it' s simply a matter of judging correctness.

Looking at the options for this question, each option is straightforward, with no need for additional logical reasoning. There is no requirement to infer new information from existing knowledge, so the models only need to retrieve information and compare each option one by one to select the correct answer.

On the right side of the table, we have also recorded the accuracy rates of human candidates for various general knowledge question types. It is evident that human performance is generally lower than the overall performance of the large models, especially in topics like political knowledge, legal knowledge, and economic knowledge. This discrepancy may arise because the models provide highly stable answers based on their knowledge base, while human candidates vary in skill level and, given the large sample size, there is often polarization in performance, with only a few candidates answering correctly. Large models, on the other hand, have been extensively tested, trained on numerous question types, and can readily access vast amounts of information, far surpassing humans in information retention.

However, in geography knowledge questions, human candidates slightly outperform the models. This is not because human performance is particularly outstanding in this question type, as their performance in geography is not the best across categories. Instead, the models' performance is consistently poor in this area, likely due to the reasons mentioned above—human candidates have better question comprehension and detail-capturing abilities than models. They also possess a capability to identify hidden logic and perform reasoning that the models lack.

In summary, large models outperform human candidates in most question types. However, their performance in a few categories is still lacking, with significant discrepancies in logical reasoning and question comprehension. To be considered a reliable tool for national exam preparation, further refinement and development of these models are necessary.

2. Words Understanding and Expression

In the Administrative Aptitude Test, the verbal comprehension and expression section includes 40 questions, each worth 0.8 points, making it one of the more heavily weighted sections. This section primarily assesses comprehension and fill-in-the-blank skills in Chinese, covering vocabulary, idioms, and sentence structure, with 20 questions in each part[6].

In the vocabulary and idioms portion, most questions require selecting the best option to fill in blanks with idioms or words. Test-takers need to pay close attention to specific meanings and semantic relationships within the context to choose the most suitable answer. There are seven types (Table 4) of fill-in-the-blank formats: single idiom, word + single idiom, double idioms, double words, triple idioms, word + double idioms, and double words + single idiom. This variety thoroughly evaluates test-takers' understanding of semantics and vocabulary usage.

In the sentence segment, there are three subtypes: fill-in-the-blank, analysis, and sequencing. For fill-in-the-blank questions, test-takers must first understand the main idea of the sentence, then use the context to select an option that makes the passage flow coherently. Analysis questions focus on comprehension and summarization, with some requiring identification of the main idea or missing details in a passage. Lastly, sequencing questions present six shuffled sentences, which test-takers must arrange in a logical, coherent order by finding internal connections, providing a comprehensive assessment of passage structure understanding.

Question Type	GPT-4	Copilot	Gemini	Coze	ERNIE BOT	Human
Words Understanding and Expression	75.0%	72.6%	70.6%	72.5%	87.8%	69.2%

Table 3. Accuracy of Each Model and Human Test-Takers on Verbal Comprehension and Expression Questions in the Chinese Civil Service Exam' s Administrative Aptitude Test

We tested five major models on these questions.

The experimental procedure involved:

1) inputting the question in text format into each model;

2) recording its results and explanations for analysis;

3) compiling and analyzing the data.

Question Type	GPT-4	Copilot	Gemini	Coze	ERNIE BOT
Single Word	1/3	2/3	1/3	2/3	2/3
Single Idiom + Single Word	4/5	4/5	2/5	5/5	3/5
Double Idiom	5/5	5/5	4/5	5/5	4/5
Double Words	1/1	1/1	1/1	1/1	1/1
Triple Idiom	0/1	0/1	0/1	0/1	1/1
One Word + Double Idiom	2/2	1/2	1/2	2/2	1/2
Two Words + One Idiom	2/3	2/3	2/3	3/3	2/3
Sentence Completion	3/3	3/3	3/3	3/3	3/3
Main Idea	3/5	4/5	5/5	5/5	5/5
Sentence Ordering	3/3	1/3	1/3	2/3	3/3
Semantic Detection	9/9	9/9	7/9	8/9	9/9

Table 4. Accuracy of Model by Question Type in the Verbal Comprehension and Expression Section of the Administrative Aptitude Test in the Chinese Civil Service Exam

Some differences in experimental procedures, primarily in step 1, arose from the fact that certain models, such as Coze and ERNIE Bot, do not support image recognition. For these models, we adjusted the input to a text-only format and continued with the remaining steps. During the initial setup, we informed the models, "You are about to take a test; please provide your answer and corresponding explanation." This approach was intended to ensure the models fully understood the questions without affecting the test outcomes.

The model test results are shown in Table 3, where most models demonstrated good accuracy. ERNIE Bot and Coze achieved the highest accuracy rates, both exceeding 85%, indicating a strong grasp of verbal comprehension and expression tasks. Among the other models, accuracy rates ranked from highest to lowest were GPT-4, Copilot, and Gemini. The lowest accuracy rate was achieved by Gemini, which still scored 67.5%, indicating a respectable level of semantic understanding.

Analysis of Typical Questions:

Example 3: The recent documentaries such as Restoration of Cultural Relics in the Forbidden City, which feature national treasures as their theme, provide a new _____ on understanding history. These documentaries not only _____ specific cultural relics but also depict the spiritual world of the era to which these relics belong. This leads the modern audience to experience _____ with traditional culture, inspiring emotional engagement, broadening historical perspectives, and allowing the audience to feel as if they are in a museum displayed on screen. The most suitable option to fill in the blank in the underlined part in order is:

A Direction, Restriction, Broad Vision.

B Angle, Stickiness, Attachment.

C Scope, Limitation, Presence.

D Path, Stop, Breath.

Question Analysis:
Example 3 is a typical example question that clearly encapsulates the key testing points and problem-solving approach for this question type. The testing points are often highly generalized and prominent, allowing readers to clearly identify the question's focus and solution direction. It also provides insight into the model' s approach and internal logic in question analysis. In the following content, we focus on analyzing this question. The representative question type here is a word-fill problem, requiring three blanks to be filled: two with two-character words and one with a four-character idiom. From the context of the question, we can infer that the parts of speech needed for the blanks are noun, verb, and adjective, respectively.

In the first blank, the prompt states that these documentaries "offer a new _____ for understanding history." Based on the structure of the sentence, the blank should be filled with a noun that implies "opening new perspectives or directions," such as "angle," "direction," or "pathway," indicating a sense of orientation. The second blank appears in "These documentaries use specific artifacts as clues but do not _____ artifacts." Here, an appropriate verb showing a level of moderation or control is needed to complete the sentence fluently and meaningfully. For the third blank, the sentence reads, "to guide contemporary audiences to engage with traditional culture, foster emotional involvement, and broaden historical perspectives." The verb before the third blank is "guide," indicating that an adjective expressing "the kind of new experience or expanded understanding the audience should gain" is needed.

In the responses of major models, we see how they perceive and interpret semantics. ChatGPT-4o, Coze, and ERNIE-BOT provided correct answers for this question. Taking ChatGPT-4o as an example, it analyzed each blank individually, testing each option within the specific context of each blank. Starting from the semantic meaning of each word, it assessed fit within the context and overall logic, finally choosing the optimal answer. ChatGPT-4o' s analysis for the first blank was as follows: "视角" (perspective) refers to an angle of observation, and documentaries offering a new perspective on history aligns well with the context. "方向" (direction) is too strong semantically; "境界" (realm) does not match the context well; "途径" (pathway) generally refers to a method or road, and documentaries are not a pathway to understanding history.

However, the other models, Gemini and Copilot, answered this question incorrectly, choosing options A and D, respectively. Taking Copilot as an example, it also analyzed each blank individually, but its explanation process was much simpler compared to ChatGPT-4o. It merely explained the meanings of the words without evaluating if they fit the context, leading to an incorrect answer. It is clear that ChatGPT-4o is more logical in answering questions, offering a more detailed reasoning process, showing a clear semantic assessment process and a structured, well-supported answer. In contrast, Copilot' s answer is simpler, with a single-threaded logic, lacking a deeper consideration of the question, indicating weaker semantic comprehension.

Overall, large language models demonstrate a certain level of ability in answering semantic comprehension questions. The main difference is whether the model performs contextual analysis. For instance, Copilot struggled to comprehend the semantics adequately, failing to capture the characteristics and key elements of the linguistic environment. ChatGPT-4o, on the other hand, showed a distinct semantic comprehension process, placing the word meanings into the text to assess contextual suitability. This follows a proper problem-solving approach for semantic comprehension

questions, demonstrating its ability to understand semantics in testing.

We believe the discrepancy among models stems from three main factors:

a training set size

b training model direction

c differing levels of training depth

These factors lead to variation among models in addressing semantic comprehension questions effectively.

3. Quantitative Relations

We tested most of the available AI models, including ChatGPT-4, ERNIE Bot (Wenxin Yiyan), Copilot, Coze, and Gemini. The accuracy of each model is summarized in Table 5.

Model Name	GPT-4o	Copilot	Gemini	Coze	ERNIE BOT
Accuracy	30%	50%	80%	50%	25%

Table 5. Accuracy of Each Model on Quantitative Reasoning Questions

Question Type	GPT-4o	Gemini	Copilot	Coze	ERNIE BOT
Engineering Questions	50.00%	25.00%	75.00%	25.00%	25.00%
Probability Questions	50.00%	0.00%	0.00%	0.00%	0.00%
Time-Line Questions	0.00%	33.00%	66.70%	33.00%	33.00%
Geometry Questions	0.00%	0.00%	0.00%	0.00%	0.00%
Ave.	25%	15%	35%	15%	15%

Table 6. Accuracy of Each Model on Subtypes of Quantitative Reasoning Questions

Table 6 presents the performance of different models on specific tasks, highlighting the differences in performance between models. While ChatGPT-4 previously showed the best results among earlier models, Copilot demonstrated notable improvement. Compared to earlier models, newer models like Coze, ERNIE Bot, and Gemini displayed more consistent results.

This also highlights the steady improvement of AI models in this type of task since 2019. Early models displayed relatively inconsistent performance, while more recent models, such as ChatGPT-4 and Copilot, have demonstrated

clear advantages over previous versions across a range of questions. However, some questions remain challenging, indicating there is still room for further enhancement in these models[7].

Overall, Copilot performed the best, with an average accuracy rate of 35%, outperforming other models in three of the four content categories. It excelled in engineering problems, achieving an impressive 75% accuracy rate—significantly higher than other models—and also performed well on time-related problems. In contrast, ChatGPT showed balanced performance across all categories, especially in probability problems, where it achieved 50% accuracy, demonstrating a strong capability in handling statistics and probability-related questions[7].

Notably, in the geometry category, all models performed poorly, with an accuracy rate of 0%. This may suggest that geometry problems present a significant challenge for these models or that current models still struggle with tasks involving spatial and shape-related reasoning. The limited spatial reasoning capabilities of large language models (LLMs) may be a contributing factor. Overall, LLMs show considerable potential in mathematical reasoning and a trend of continuous improvement, but they still face numerous challenges and limitations that call for further research and optimization[8]. In summary, ChatGPT-4 ranks in the middle among these models, showing particular strength in handling engineering and probability problems.

Analysis of Typical Questions:

Example 4: A cylinder has a height of 1, with a square inscribed in the circular base. The side length of the square is also 1. Now, the cylinder is sliced vertically into 4 equal parts, resulting in a prism with a height of 1 for each section. What is the total surface area of the removed parts?

A $\sqrt{2}(\pi+2)$

B $2\sqrt{2}(\pi-2)$

C $(\sqrt{2}+1)\pi+2$

D $2\sqrt{2}\pi-2$

Question Analysis:

As shown in the diagram, the diagonal of the inscribed square on the circular base is equal to the diameter of the circle, so the diameter of the cylinder's base is $\sqrt{1^2+1^2}=\sqrt{2}$. Therefore, the radius is $\frac{\sqrt{2}}{2}$.
To obtain a square prism with side length 1, we need to slice along $A_1B_1B_2A_2$, $B_1C_1C_2B_2$, $C_1D_1D_2C_2$, and $D_1A_1A_2D_2$, cutting it into 4 parts with equal angles. The surface area of the removed parts includes the side surface areas of the cylinder, the two bases of the cylinder minus the area of the inscribed square, and the four lateral surfaces of the prism.
The total surface area of the removed parts is:

$$S_{\text{total}} = S_{\text{cylinder}} + 2\times S_{\text{base}} + 4\times S_{\text{side}} = 2\pi\times\frac{\sqrt{2}}{2}\times 1 + 1\times 2\left(\pi\left(\frac{\sqrt{2}}{2}\right)^2 - 1^2\right) + 4\times 1\times 1^2 = (\sqrt{2}+1)\pi+2.$$

According to the Fig.1 explanation, the correct answer is C.

Let' s look at another two typical example questions, starting with a geometry problem:

The core of geometry problems lies in spatial reasoning, shape recognition, and logical relationship modeling. Unlike

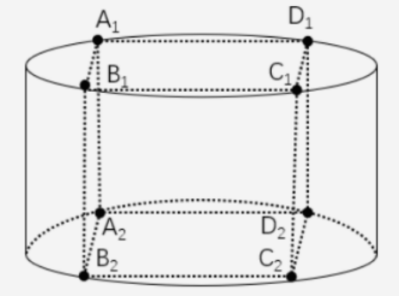

Figure 1. A graphical explanation for Question Example 4

textual reasoning and numerical computation, geometry problems often require a model to understand spatial relationships. For instance, calculating areas and volumes of geometric shapes, deriving angles and distances—these tasks are not commonly encountered in traditional natural language processing. For most large language models (LLMs), their training datasets rely heavily on natural language text, while geometry problems involve more spatial visualization and mathematical formula reasoning, which is not easily learned directly from large-scale text data. The limitations of models in handling geometry problems may be related to the following factors:

1. **Insufficient Training Data**: Geometry problems often require specialized mathematical or engineering knowledge, which is not widely found in conventional text corpora. Even if a model receives related mathematical training, its spatial reasoning abilities for geometry problems remain limited.

2. **Lack of Graphical Processing Capabilities**: Geometry problems typically rely on graphical reasoning, while large language models are primarily trained on text and symbols and lack an inherent mechanism for processing graphics. Current LLMs do not have the capability to handle and understand visual data, which restricts their performance on geometry problems.

3. **Complexity of Logical Reasoning**: Solving geometry problems often depends on a series of rigorous logical steps, requiring models to maintain a high level of precision and consistency in the reasoning process. However, current language models are better suited for semantic reasoning tasks, while precise modeling and reasoning in geometric logic pose particular challenges.

Analysis of Typical Questions:

Example 5: A project requires 3 hours for person A to complete independently, and person B requires 4 more hours than when working together with A to complete it independently. Together, A and B can complete the project in 4 hours. How many hours does it take for B to complete it alone?

A 10

B 12

C 6

D 8

Question Analysis: According to Question Example 5, engineering problem is a common question type in civil service exams and many other tests. It often involves complex calculations related to cooperation, efficiency, and time. This type of problem requires good logical reasoning and mathematical computation skills from the examinee

or model. AI models (such as ChatGPT-4o, ERNIE Bot, Copilot, Coze, and Gemini) show certain differences in handling such problems. This article explores the performance differences among various AI models in solving engineering problems by examining a specific example. The example problem is a typical cooperation-efficiency problem, where the examinee needs to understand the formula for calculating work efficiency and carry out multi-step reasoning to arrive at the answer.

1. Assume A's work rate is $\frac{1}{3}$ (i.e., A completes $\frac{1}{3}$ of the work per hour). If B takes x hours to complete the work independently, B' s work rate is $\frac{1}{x}$.

2. According to the question, B takes 4 more hours to complete the work alone than when working with A. Therefore, B' s time to complete the task independently is 4 hours + 4 hours = x hours.

3. The combined work rate of A and B is the sum of their work rates, i.e., $\frac{1}{3}+\frac{1}{x}=\frac{1}{4}$

4. Solving the equation $\frac{1}{3}+\frac{1}{x}=\frac{1}{4}$, we get $x=12$.

Thus, the answer is **12**.

Based on the solution process, we conclude:

1 **Performance of AI Models on Engineering Problems:** According to the data in Table 2, Copilot performs the best in handling engineering problems, achieving an accuracy rate of 75%. In contrast, the performance of ChatGPT-4o and other models is slightly inferior, especially ERNIE Bot and Coze, with accuracy rates of only 25%.

2 **Advantages of Copilot:** Outstanding performance may be related to its optimization in handling complex mathematical operations and logical reasoning. Its accuracy rate on engineering problems is significantly higher than that of other models, demonstrating its strong capability in deriving multi-step formulas and performing ratio calculations. For questions related to cooperation tasks between A and B, which often require multiple unit conversions and fraction calculations, Copilot performs exceptionally well.

3 **Balanced Performance of ChatGPT-4o:** ChatGPT-4o performs moderately on engineering problems, with an accuracy rate of 50%. This result suggests that, while the model has some reasoning abilities, it may still have gaps when facing multi-step complex calculations. This could be because the model does not fully utilize the premises in the reasoning process or lacks precision in handling ratio calculations.

4 **Limitations of Other Models:** Other models, such as ERNIE Bot, Coze, and Gemini, perform relatively poorly, with accuracy rates of only 25%. These models may encounter greater obstacles in handling engineering problems, particularly in multi-step reasoning and precise calculations. Geometry and engineering problems typically require models not only to understand language expressions but also to excel in tasks involving formulas and ratios. The performance of these models indicates that current AI still has considerable room for improvement in complex mathematics and logical reasoning.

5 **Complexity Reflecting:** The complexity of engineering problems is reflected in the multi-step reasoning process and the understanding of different work efficiencies. To achieve the final result, a model needs to handle multiple variables in the problem through conversions, allocations, and recalculations. AI models may exhibit performance differences in handling such problems due to the following factors:

a **Differences in Training Data:** Some models may lack data on mathematical problems related to engineering during training, leading to insufficient reasoning capabilities in handling such problems.

b **Limitations in Calculation Accuracy:** Some models perform poorly in calculations involving fractions, percentages, and ratios, affecting their accuracy in multi-step calculations.

c **Breakdown in Logical Reasoning Chains:** When solving multi-step problems, models may struggle to maintain the completeness of the reasoning chain, resulting in significant deviations in the final answer.

4. Logical Reasoning

This study tested a series of advanced GPT models on the logical reasoning section of the National Civil Service Exam to assess their performance across four reasoning tasks: figure reasoning, definition judgment, analogy reasoning, and logical judgment. The models tested included the latest GPT-4, Gemini, Coze, Copilot, and ERNIE Bot. Table 7 shows the accuracy of each model on these reasoning tasks.

The design and development of the logical reasoning section in the National Civil Service Exam have made significant contributions to national talent selection. Since 2000, the logical reasoning questions in the exam have been continuously refined, moving toward a more scientific and standardized format. Focused on assessing candidates' logical thinking skills, the logical reasoning section is relatively comprehensive and well-structured, with many questions exhibiting high levels of reliability, validity, and differentiation. This has enabled a large number of candidates to stand out and make important contributions to the country, achieving substantial social benefits from the examinable[9].

Reasoning section consists of 40 questions, divided into four modules—10 questions each on figure reasoning, definition judgment, analogy reasoning, and logical judgment. Overall, the Copilot model ranked highest with 21 correct answers, followed closely by ChatGPT-4 with 20 correct answers. Coze and Gemini each answered 19 questions correctly, while ERNIE Bot performed comparatively weaker, answering only 12 questions correctly. Table 7 indicate that, while the models demonstrate some capabilities in certain areas, they still face limitations in handling complex reasoning tasks, particularly in accurately interpreting visual information.

In this round of testing, ERNIE Bot was the top-performing model, with 23 correct answers. It also demonstrated specific strengths in individual tasks, ranking first in several areas. For figure reasoning and logical judgment tasks, it achieved accuracy rates of 30% and 70%, respectively, both the highest among the models. In the definition judgment task, it reached a high accuracy rate of 80%, and in analogy reasoning, it achieved 50% accuracy.

The second-highest performer was the Copilot model, which scored 70% accuracy in both the definition judgment and analogy reasoning tasks, highlighting its strong abilities in content extraction and knowledge matching. It achieved a 50% accuracy rate for logical judgment and 20% for figure reasoning.

ChatGPT-4 was close behind, performing exceptionally well in the definition judgment task with an impressive 90% accuracy rate. However, despite offering an image upload interface, it showed limited ability to extract information from images, scoring only 10% in the figure reasoning task. In the logical judgment and analogy reasoning tasks, it achieved 60% and 40% accuracy, respectively, ranking it overall as the second-highest in total accuracy.

The Coze and Gemini models each answered 19 questions correctly, tying for fourth place, though they displayed different strengths across specific tasks. Coze achieved an 80% accuracy rate in the definition judgment task, while Gemini scored 70% in this area. For analogy reasoning and logical judgment, Coze scored 60% and 50%, respectively, while Gemini scored 50% in both tasks. In figure reasoning, Gemini performed poorly with only 20% accuracy, while Coze did not answer any correctly, resulting in a 0% accuracy rate.

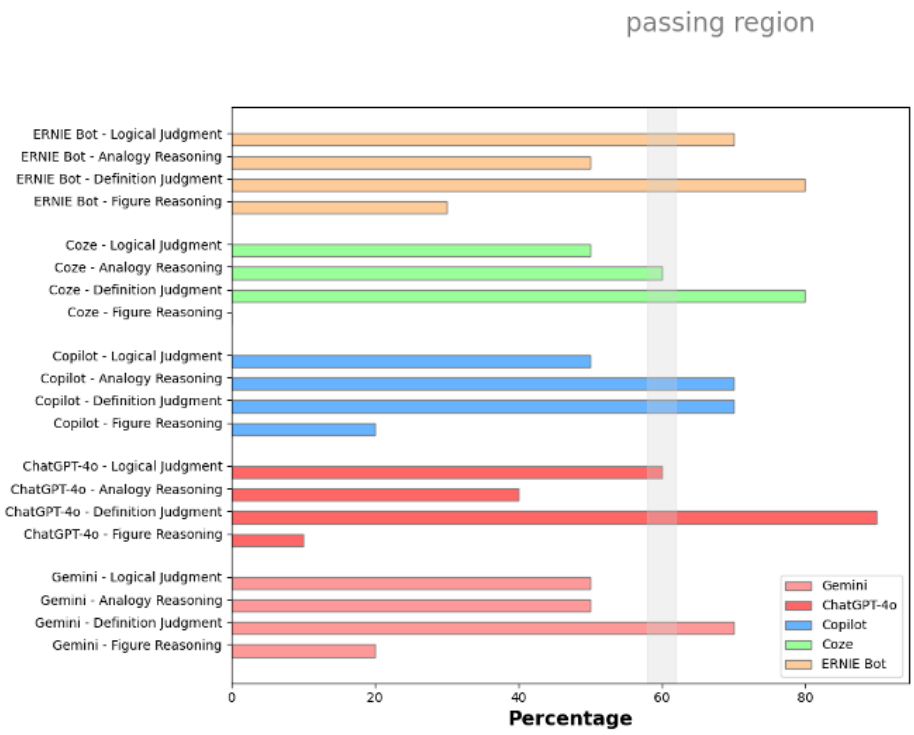

Table 7. Score Data of Each Model in the Logical Reasoning Section

Human test-takers achieved an overall accuracy rate of 69.35% in the logical reasoning section, outperforming all GPT models. The overall performance chart shows that human test-takers maintained relatively balanced accuracy across the four subcategories, with low variance. Each subcategory had an accuracy rate above 60%, with the highest accuracy in figure reasoning at 72.5% and the lowest in logical judgment at 66.3%.

Analysis of Typical Questions:

Example 6: Micro drones: Rotor drones

A Tropical plants: Spice plants

B Collective decision-making: Individual decision-making

C Visual thinking: Abstract thinking

D Open-loop system: Closed-loop system

Question Analysis: Analogy reasoning questions assess the ability to identify cross-relationships within logical connections, requiring evaluation of each option' s logical relationship and reasoning based on a clear understanding of these associations. For our example analysis, we selected **Question Example 6**, an analogy reasoning question. This question does not contain unique Chinese terms, historical references, or specific context, making it straightforward and suitable as an example. It assesses analogy reasoning by requiring recognition of the hierarchical and subordinate relationship between "micro drones" and "rotor drones." Both ChatGPT-4 and ERNIE Bot accurately identified this layered relationship, demonstrating strong semantic understanding and logical reasoning abilities in natural language processing—skills that are highly applicable to complex classification and analogy tasks. Both Coze and

Copilot answered incorrectly, highlighting areas for further improvement.

Analysis of Typical Questions:

Example 7: According to Fig.2, choose the correct one

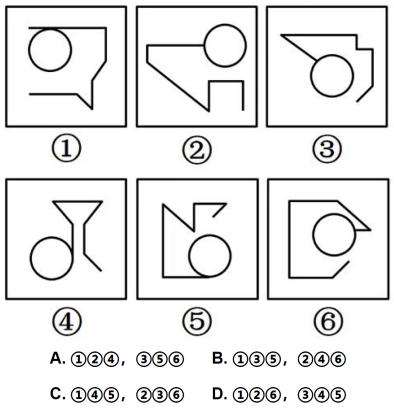

Figure 2. Question Example 7: Judgment reasoning

Question Analysis:

Figure reasoning questions assess abstract reasoning and pattern recognition skills, requiring the interpretation of graphical information to identify patterns and apply them to reasoning tasks. The "binary division" format is a key style in figure reasoning questions; it first appeared in the National Civil Service Exam in 2011 and has been a major component of this question type ever since. From a logical perspective, division serves as a method for clearly defining concept boundaries and distinctions, making it a valuable complement to, rather than a redundancy of, definition judgment. Additionally, binary division questions are designed to measure a test-taker's inductive reasoning ability. By grouping three figures into one category and three into another, the test-taker must use inductive reasoning to identify characteristics that are both shared and unique among the figures[6].

In this round of testing, it's evident that even the most advanced GPT models currently available still fall short in handling the logical reasoning questions on the Civil Service Aptitude Test. Their performance was lacking, with issues such as inaccurate image recognition, disorganized reasoning, and an inability to provide efficient and accurate answers. As a result, these models are not recommended as supplementary tools for the Aptitude Test.

Analysis of Typical Questions:

Example 8: In linguistic description, **anaphora** refers to a linguistic unit (*anaphoric term*) that points back to a previously expressed unit or meaning (*antecedent*) to achieve self-explanatory processes or results. Anaphora can be divided into **direct** and **indirect** anaphora[10]. Direct anaphora is when the anaphoric term and the antecedent have an obvious referential relationship, meaning the anaphoric term repeats the antecedent. Indirect anaphora is when the relationship between the anaphoric term and the antecedent is not explicit, and judgment must be made within a specific language context to confirm it. Based on the above definition, select the option where the relationship between the anaphoric term and the antecedent represents **direct anaphora**:

A Express gratitude to those who attended the meeting; their presence was our support.

B Some tricycles are parked at the door, and many cars have been parked there for years.

C This room is well-made; the door is made of mahogany.

D He went to the restaurant before lunch and ordered a cup of coffee. The server was an Italian.

Question Analysis: Definition Judgment Questions test the ability to logically assess concepts. They require determining the logical relationship among options based on sentence structure, then reasoning to answer after understanding the logical connection between the question stem and each option. Large models perform this task by typically transforming natural language text into structured data, such as class attributes and concepts, to represent the meaning of input statements, enabling a more accurate capture and extraction of meaning. We selected Question Example 8 for a typical example analysis. This question clearly distinguishes between direct and indirect **anaphora** without complex linguistic features or specific historical contexts, making it suitable for exemplary analysis. This question examines the ability to identify the relationship between **anaphora** language and antecedents, requiring accurate judgment of the connection between "**anaphora**" and "antecedents" in the sentence. From the answer results, Copilot, ERNIE Bot, and Coze correctly identified the correspondence between **anaphora** terms and antecedents, demonstrating strong natural language reasoning abilities. However, ChatGPT-4o failed to accurately recognize the relevant **anaphora** relationships in this question, indicating room for further optimization in handling complex language reasoning tasks.

These test results reveal differences in model performance on reasoning tasks and their potential application scenarios. Although the **Coze** model performed well in definition judgment tasks, its score of 0 in graphic reasoning indicates a significant shortfall in handling visual or graphic information. However, its moderate performance in analogy reasoning and logical judgment demonstrates that it retains some utility in certain reasoning tasks.

In contrast, **Gemini** and **Copilot** show more balanced performance across various tasks, particularly with **Copilot**'s high scores in analogy reasoning and definition judgment, showcasing its potential in semantic analysis and concept-matching tasks. **GPT-4o**'s performance appears to be task-dependent, as it excelled in definition judgment but lagged in graphic reasoning, limiting its application in vision-related tasks.

ERNIE Bot performed poorly in analogy reasoning, but its solid performance in logical judgment tasks highlights its potential in logical reasoning and decision support tasks. These results provide important references for future model optimization and improvement.

In this round of testing, it is evident that even the most advanced GPT models currently fall short in handling civil service exam judgment reasoning questions. Their performance remains unsatisfactory, with inaccurate recognition of image-based questions, confusing reasoning logic, and an inability to provide efficient and accurate answers. **Therefore, they are not recommended as an auxiliary tool for civil service exams.**

5. Data Analyzed

In this test, the models were required to answer a total of 20 data analysis questions (see Table 8). The results show that Copilot performed the best, achieving an accuracy rate of 60% and scoring 12 points, while ChatGPT-4 ranked second with 55% accuracy, scoring 11 points. Other models, including ERNIE Bot, Gemini, and Coze, performed significantly worse, with scores of 7, 6, and 3 points, respectively. The differences in performance among these models highlight the varying capabilities of current AI models in handling complex reasoning and data analysis tasks. Language models, especially when tackling real-world problems that require precise reasoning and identifying

Model	Accuracy(2022)	Accuracy(2023)	Accuracy(2024)
ChatGPT-4o	11/20	13/20	13/20
Copilot	12/20	9/20	9/20
ERNIE Bot	7/20	4/20	7/20
Gemini	6/20	6/20	4/20
Coze	3/20	3/20	4/20
Human	14.61/20	14.6/20	14.54/20

Table 8. Scores of Each Model on Data Analysis Section (2022-2024)

key information, show considerable limitations. This reflects a lack of ability to extract and understand complex, structured information, resulting in inaccurate recognition and flawed reasoning.

While Copilot and ChatGPT-4 performed relatively well in terms of accuracy, they still fell short of achieving a perfect score, indicating that even the most advanced models face challenges in comprehension and reasoning when dealing with complex analysis questions. Their specific performances are as follows:

1) Copilot: As the highest-scoring model, Copilot demonstrated strong information extraction and reasoning abilities on data analysis questions. It was able to locate answers accurately in questions involving charts and data; however, it showed some inconsistency with multi-step reasoning questions. Its score of 12/20 suggests difficulty in maintaining a consistent reasoning chain on certain questions, which affected its overall performance.

2) ChatGPT-4: With a score just below Copilot, ChatGPT-4 displayed some stability. Its primary issue lay in reasoning with complex data, particularly when dealing with questions involving multiple variables and conditions. The model struggled to accurately identify the core information, leading to errors in the reasoning process. These issues highlight current limitations in the model' s logical deduction abilities.

3) ERNIE Bot, Gemini, and Coze: These models performed significantly worse than Copilot and ChatGPT-4 on data analysis questions. Especially on multi-step reasoning questions, they often failed to correctly identify key data, resulting in scores far below the other models. This reflects these models' deficiencies in handling complex, structured information.

Another challenge for these language models is their difficulty with performing independent logical reasoning. Data analysis questions require not only basic calculations but also the ability to deduce accurate conclusions from the information provided. Although Copilot and ChatGPT-4 performed relatively well in terms of accuracy, their scores of 12 and 11 points indicate that they still struggle to make correct deductions on all questions. The lower scores of ERNIE Bot, Gemini, and Coze further highlight these models' limitations in handling complex reasoning and analytical tasks. When faced with multi-step reasoning, these models frequently make errors and struggle to extract essential information from complex data to perform effective reasoning, unlike human test-takers.

Analysis of Typical Questions:

Example 9: According to Fig.3, answer the following question: In 2019, what was the market size of China's advanced IC packaging industry in billions of yuan?

A 234

B 252

C 279

D 296

Figure 3. The market size of China' s IC packaging industry and the proportion of advanced IC packaging market from 2011 to 2020.

Question Analysis:
This question requires the model to extract data information for the year 2019 from the chart and select the correct market size. Such questions not only require the model to accurately read chart information but also to combine contextual information to make the correct judgment. Therefore, after a comprehensive analysis, **the answer to this question is B.**

Challenges of Data Analysis Question Types:
Data analysis questions are an important type in civil service exams, with the difficulty lying in the need to extract key information from a large amount of text and chart data, perform mathematical calculations, and conduct logical reasoning. For AI models, these questions present the following challenges:

1. **Multi-step Reasoning Ability:** Data analysis questions often involve associative reasoning across multiple data points. Models need not only to extract information but also to accurately connect this information and reach a final conclusion. Many models tend to experience logical chain breaks when handling multi-step reasoning, resulting in incorrect answers.

2. **Structured Information Extraction:** AI models still have limited capabilities in extracting information from charts, tables, and data. Although these models can obtain information from large amounts of text, their understanding and extraction abilities are relatively weak when dealing with visual data (such as charts).

3. **Logical Analysis Ability:** Data analysis questions require models not only to perform basic mathematical operations but also to have logical reasoning abilities. Current models still need improvement in this area, especially when handling data that requires synthesizing information from multiple sources, where models may easily overlook key points or make incorrect inferences.

From the test results, an important conclusion can be drawn: current language models still need further refinement and optimization to handle data analysis questions in civil service exams effectively. The performance of different models indicates that, as general-purpose models, they struggle when dealing with tasks in specific domains. These types of exam questions involve a great deal of specialized knowledge and logical analysis, and general-purpose models often cannot make efficient and accurate judgments when faced with multi-domain knowledge. Therefore, the future development of language models requires in-depth optimization for specific domains rather than relying solely on general knowledge bases. Through specialized training in a particular field, models can more accurately understand domain-specific knowledge and better handle complex tasks like data analysis questions.

Argumentative Essay Writing

In typical exams, essay writing is often considered significantly more challenging than multiple-choice questions, and this distinction becomes even more pronounced for computer systems. Essay questions require a high level of comprehension, structured reasoning, and creativity, all of which pose unique challenges for AI models. In the context of evaluating the cognitive abilities of large language models, such as ChatGPT, we attempted to assess their performance on the Argumentative Essay Writing (AEW) section of the Chinese National Civil Service Exam. This exam section demands not only analytical and writing skills but also an understanding of policies and the ability to apply them in a practical context. In preparation for the AEW test, we provided contextual prompts and specified writing requirements to the model to help it understand the expectations for the task at hand.

During the testing process, we observed that while ChatGPT-4 demonstrated strong comprehension and coherence in its responses, it struggled to precisely control the length of its answers. This challenge was particularly evident in longer responses. For example, when tasked with writing a 300-word passage, ChatGPT-4 generally stayed within the limit, occasionally exceeding it by around 10 words. However, for a 1,000-word essay, the model often overshot the target significantly, sometimes producing responses up to 1,400 words in length. This pattern suggests that while the model can comprehend general requirements, it faces difficulty with fine-grained control over response length, especially for extended writing tasks. To address this, we instructed the model to structure its answers in paragraphs, hoping to better align its output with the format and standards of the AEW exam. This adjustment allowed the model' s responses to more closely resemble human-generated essays, although the word count control remained a challenge.

To evaluate the quality of the model' s responses, we utilized the Fenbi grading platform, a dedicated assessment tool for the Chinese Civil Service Exam. Fenbi includes an automated grading system capable of assessing AI-generated outputs. Importantly, the platform not only provides raw scores but also offers the average scores of human test-takers. This feature enabled us to benchmark ChatGPT-4' s performance against that of actual human candidates, providing a meaningful reference point for assessing the model' s capabilities.

For context, the AEW section of the Chinese Civil Service Exam primarily tests candidates on their abilities in comprehensive analysis, logical thinking, writing, and policy understanding. This section is structured to simulate real-world tasks, often including several materials and a series of questions. Candidates must read and interpret the materials and then formulate answers based on their content. The AEW section is therefore focused on assessing practical application skills, rather than simple knowledge recall. This section includes five questions, totaling 100 points, with each question accompanied by materials of varying lengths. As candidates progress through the exam, the length of the required responses generally increases, with the longest response reaching approximately 1,000-1,300 words. The exam' s duration is three hours, making the AEW section a comprehensive assessment that not only tests knowledge but also logical reasoning and practical writing abilities.

We tested ChatGPT-4 on actual AEW questions from the past three years to evaluate its performance on this challenging section. We selected ChatGPT-4 specifically due to its consistent output quality, strong understanding of prompts, and suitability for detailed analysis. While other models generated correct answers, they lacked the consistency and depth demonstrated by ChatGPT-4, making it the most representative model for this evaluation. As shown in Figure 10, ChatGPT-4 performed impressively on the AEW exam, achieving scores that placed it in the upper tier, indicating its strong capabilities in text comprehension and structured writing.

Year	ChatGPT-4o	Human(Avg./Best)
2024	65.5	30.4/86
2023	60	28.36/85
2022	58	32.4/90

Table 9. ChatGPT-4o' s Score in the Argumentative Essay Writing (AEW) Section of the Chinese National Civil Service Exam

The data in the table further illustrates ChatGPT-4o' s performance over the years. From 2022 to 2024, ChatGPT-4o consistently outperformed the average scores of human test-takers in the AEW section. Specifically, ChatGPT-4o scored 58 in 2022, 60 in 2023, and 65.5 in 2024. In contrast, the average scores for human candidates during this period were significantly lower: 32.4 in 2022, 28.36 in 2023, and 30.4 in 2024. These results highlight ChatGPT-4o' s strong abilities in understanding complex prompts and generating coherent responses, consistently surpassing the human average each year. However, despite its impressive performance, ChatGPT-4o has not yet reached the top scores achieved by the highest-performing human candidates, who scored 90, 85, and 86 from 2022 to 2024, respectively. This gap suggests that while ChatGPT-4o excels in structured and coherent response generation, it may still lack the nuanced analysis and depth that top human candidates can achieve. This limitation is particularly evident in tasks requiring intricate reasoning, layered argumentation, and deep policy insights, which are often essential for excelling in AEW.

Interestingly, the upward trend in ChatGPT-4o' s scores over the three-year period suggests continuous improvements in its understanding and response generation abilities. This likely reflects advancements in large language model training and optimization, which have enhanced its capacity for high-level comprehension and structured writing. The observed improvement trajectory indicates ChatGPT-4o' s growing potential to handle complex language-based assessments, making it a promising tool for applications that require detailed understanding, structured responses, and policy-oriented writing.

Overall, ChatGPT-4o' s performance in the AEW section demonstrates its strong potential as an assistant for tasks that involve comprehensive text understanding and complex response generation. Although there remains a gap between the model and the highest-achieving human candidates, its consistent performance above the human average across multiple years indicates that it is well-suited for applications in academic, professional, and policy-based settings. As large language models continue to evolve, ChatGPT-4o' s strengths in these areas suggest it could become increasingly valuable for applications that require nuanced comprehension, logical analysis, and coherent, structured responses.

Conclusion

This study thoroughly examined the capabilities and limitations of large language models, specifically ChatGPT-4, within the context of the Chinese National Civil Service Exam' s Administrative Aptitude Test (AAT) and Argumentative Essay Writing (AEW) sections. Designed to assess a broad range of cognitive abilities, this exam

presents complex challenges in language comprehension, logical reasoning, data analysis, and policy understanding—skills essential for roles in public service. Our evaluation of ChatGPT-4' s performance on these tasks provides valuable insights into its strengths and areas needing improvement[11].

ChatGPT-4 demonstrated strong proficiency in several key areas, such as language comprehension, structured response generation, and basic data analysis. In the AEW section, the model produced coherent and well-structured essays on complex topics, reflecting its capability to handle high-level writing tasks. In the AAT section, it displayed competence in quantitative analysis and pattern recognition, often outperforming the average scores of human candidates. These findings indicate that large language models like ChatGPT-4 could be effective tools for tasks requiring natural language understanding and structured communication.

However, our analysis also revealed limitations. ChatGPT-4 struggled with precisely controlling response length, especially in the AEW section where word limits are critical. While it generally stayed within limits for shorter essays, the model often exceeded the word count on longer responses, suggesting a lack of fine-grained control. Furthermore, logical reasoning tasks requiring multi-step problem-solving and abstract thinking posed challenges for the model. This gap in logical reasoning highlights a broader limitation of current language models: while they excel at pattern recognition, they are less capable of handling tasks that require deep contextual understanding and causal inference.

To benchmark ChatGPT-4' s performance, we utilized the Fenbi grading platform, which provides comparative scores with human test-takers. This allowed us to see not only where ChatGPT-4 excelled but also where it fell short of top human performers, especially in nuanced reasoning and depth. Despite its impressive achievements, the model did not reach the highest scores seen among human candidates, emphasizing that further development is needed to meet the standards of human expertise in some areas[12].

Looking forward, this research has significant implications for the use of AI in education and assessment. The demonstrated ability of language models to handle complex language tasks suggests potential applications in standardized testing and automated evaluation. For example, AI could be used to standardize grading practices, reduce human bias, and offer consistent feedback, particularly in large-scale exams. In educational settings, such models could provide accessible feedback to students, especially in under-resourced areas, thereby supporting skill development in writing, analysis, and critical thinking.

In conclusion, while ChatGPT-4 and similar language models show considerable potential for handling sophisticated assessment tasks, this study highlights key areas for improvement, such as logical reasoning and adherence to strict guidelines. As these models continue to evolve, they are likely to become more valuable in applications requiring high-level comprehension and structured response generation, potentially transforming the landscape of educational and professional assessments[138]. This research thus provides a foundation for understanding both the current capabilities and future possibilities of AI in cognitive assessments, guiding the development of next-generation models equipped to tackle a broader range of cognitive challenges.

References

1. Katz, M., Bommarito, D., James, M., Gao, S. & Arredondo, P. D. Gpt-4 passes the bar exam, DOI: 10.6084/m9.figshare.c.7031287.v1 (2024). Collection.
2. Deng, Y. A study on the examination and recruitment system of chinese civil servants (2023).
3. 杨珊. 我国公务员考试录用制度的研究 (2017).
4. 石婷婷. 我国公务员“凡进必考”制度的创新实践与分析 (2018).
5. Schulze Balhorn, W. J. B. S. e. a., L. Empirical assessment of chatgpt’ s answering capabilities in natural science and engineering (2024).
6. Han, D. A brief analysis of the examination and recruitment system of contemporary chinese civil servants. In Economics and Management Science; Politics, Military and Law (2018).
7. Chang, Y. et al. A survey on evaluation of large language models (2024).
8. Yijia Xiao, T. L. W. W., Edward Sun. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts (2024).
9. 苏向荣. 近 20 年国考行测“判断推理”试题的特征分析及改进建议 (2023).
10. 赵睿卓. 大语言模型评估技术研究进展 (2024).
11. Michael R. King, V. U. N., Department of Biomedical Engineering. Administration of the text-based portions of a general iq test to five different large language models (2023).
12. Liu, Z. W., Z. A qualitative analysis of chinese higher education students’ intentions and influencing factors in using chatgpt: a grounded theory approach (2024).
13. Michael Hersche, T. H. A. S. A. R., Francesco di Stefano. Probabilistic abduction for visual abstract reasoning via learning rules in vector-symbolic architectures (2024).