CREA-Eval: An Evaluation Benchmark for Testing the Ability of Large Language Models to Understand Rare Earth Domain-Related Questions

Authors: Na Shihang, Yu Jiaxin, Shaoqing Ren, Gao Shuo, Wang Yu, Hongwei Yan, Hongwei Yan

Date: 2026-04-13T14:28:22+00:00

Abstract

This study aims to address the current lack of professional evaluation benchmarks for Large Language Models (LLMs) in the Chinese rare earth domain. To this end, the Chinese Rare Earth Ability Evaluation Benchmark (CREA-Eval) was constructed, covering 5 themes and 4 question types, and comprising a total of 2,443 high-quality corpora, which can efficiently evaluate the rare earth capability boundaries of various LLMs. The benchmark completed data collection and auditing through a combination of manual annotation, LLM assistance, and automated scripts, and adopted a hybrid evaluation strategy combining LLM-as-a-judge with regular expression matching.

Based on CREA-Eval, 22 mainstream LLMs from 6 platforms were systematically evaluated, and the accuracy rates of each model across different themes and question types were reported. The study further introduced the classification of subjective and objective questions from educational examinations and found that some models exhibited significant performance differences between the two categories; quantitative analysis through cosine similarity differences indicates that this phenomenon may stem from the fact that specific theme-related knowledge concepts or facts in model training originate from texts or thematic content in other fields, and specific theme-related knowledge has not been sufficiently organized through intra-thematic corpora, resulting in expression and reasoning abilities lagging behind the mastery of factual knowledge. CREA-Eval provides a standardized tool for the evaluation, selection, and fine-tuning of domain-specific LLMs in the rare earth field, contributing to the professional development of industry-specific large models.

Full Text

Preamble

CREA-Eval: An Evaluation Benchmark for Assessing the Capability of Large Language Models to Understand Rare Earth Domain-Related Issues

Institute of Rare Earth Magnetic Materials, Baotou Research Institute of Rare Earths, Baotou, 014030 nashihang@brire.com

Baotou Research Institute of Rare Earths, Baotou, 014030

Institute of Rare Earth Magnetic Materials, Baotou Research Institute of Rare Earths, Baotou, 014030

Hongwei Yan*, Institute of Rare Earth Magnetic Materials, Baotou Research Institute of Rare Earths, Baotou, 014030 hongw1125@126.com

April 13, 2026

Abstract

Abstract

This study aims to address the current lack of professional evaluation benchmarks for Large Language Models (LLMs) within the Chinese rare earth domain. To this end, we have constructed the Chinese Rare Earth Ability Evaluation Benchmark (CREA-Eval), which covers five major themes and four question types, comprising a total of 2,443 high-quality corpus entries. This benchmark enables the efficient evaluation of the capability boundaries of various LLMs in the rare earth field. Data collection and auditing were completed through a combination of manual annotation, LLM-assisted generation, and automated scripts. Furthermore, a hybrid evaluation strategy was adopted, combining LLM-as-a-judge with regular expression matching.

Based on CREA-Eval, we conducted a systematic evaluation of 22 mainstream LLMs from six different platforms, reporting the accuracy of each model across various themes and question types. The study further introduces a classification of subjective and objective questions common in educational examinations. We found that some models exhibit significant performance discrepancies between these two categories. Quantitative analysis using cosine similarity differences suggests that this phenomenon may stem from the fact that specific theme-related knowledge concepts or facts in model training are derived from texts in other domains or general thematic content. Consequently, specific domain knowledge is not sufficiently organized through intra-domain corpora, leading to a lag in expression and reasoning abilities compared to the mastery of factual knowledge. CREA-Eval provides a standardized tool for the evaluation, selection, and fine-tuning of domain-specific LLMs in the rare earth industry, contributing to the professional development of industry-specific large models.

Keywords: Large Language Models · Rare Earths · Evaluation Benchmark · Magnetic Materials

引言

Rare earth elements play an irreplaceable role in permanent magnets, laser crystals, fluorescent materials, catalysts, and new energy technologies due to their unique properties. As the global green transition accelerates, the demand for high-performance rare earth functional materials continues to rise. However, issues such as the high concentration of resources and the vulnerability of supply chains have become increasingly prominent, necessitating technological innovation to improve R&D efficiency and resource utilization. In this context, accelerating the rational design, performance prediction, and process optimization of new rare earth materials has become a strategic task to ensure national security regarding critical materials and industrial competitiveness. Facing the surging wave of artificial intelligence development, there are both opportunities and challenges; various industries are actively exploring digital and intelligent transformation paths to implement AI technology as early as possible, and the rare earth industry is no exception. Utilizing AI techniques to integrate industrial chain resources, optimize production management, and enhance innovation capabilities has become an inevitable choice for the high-quality development of the rare earth industry. Through this technological transformation, the industry can better adapt to market changes, improve resource efficiency, achieve sustainable development goals, and occupy a more favorable position in international competition.

In recent years, artificial intelligence (AI)—and particularly Large Language Models (LLMs)—has been widely regarded as a key tool for accelerating materials innovation. Some studies have already attempted to use AI to predict the properties of amorphous materials [**?**], analyze the microstructure of permanent magnets [**?**], or guide alloy design [**?**], initially demonstrating its potential. However, current LLM technology still exhibits significant deficiencies when applied to highly specialized material systems with complex physical mechanisms, such as rare earths: (1) Mainstream LLMs may lack a deep internalization of specialized knowledge within the rare earth field. Although LLMs demonstrate powerful capabilities on general corpora···

language generation capabilities, their training data regarding core rare earth concepts is sparse and fragmented. This leads to frequent "hallucinations" when the models answer professional questions—generating content that appears scientifically plausible but is factually incorrect. To address these hallucinations, benchmarks such as SimpleQA [**?**] and HaluEval [**?**] have been developed to measure the hallucination phenomena, factuality, and reliability of large language models. However, English-based benchmarks like SimpleQA are not suitable for evaluation within a Chinese linguistic context. (2) Existing Chinese LLM

evaluation benchmarks fail to reflect the cognitive depth required for the rare earth field. In recent years, several Chinese LLM benchmarks have been released, such as CLUE [?], C-Eval [?], CMMLU [?], and Chinese SimpleQA [?]. While these comprehensive benchmarks cover multiple disciplines, their question designs lean toward general knowledge and do not touch upon the unique physicochemical mechanisms and engineering practices specific to rare earth materials. Consequently, the industry cannot objectively measure the true capabilities of different LLMs in critical tasks such as magnetic material design, understanding luminescence mechanisms, or reasoning through metallurgical processes. (3) There is a lack of domain-aligned knowledge representation and verification. Current AI systems mostly employ black-box reasoning, which neither integrates the prior physical laws of rare earth materials nor includes feedback calibration mechanisms with experimental or computational data. This makes their outputs difficult for researchers and engineers to trust and adopt.

The aforementioned limitations severely hinder the credible implementation of AI technology within the rare earth industry chain. Without establishing a scientific, systematic, and verifiable professional capability evaluation system, it is impossible to identify which models truly possess the potential to support material design, process optimization, or technical consultation. To address these issues, evaluate the professional capabilities of Chinese LLMs in the rare earth field, and accelerate the digital intelligence transformation and application of AI in the industry, we propose the Chinese Rare Earth Ability Eval benchmark (CREA-Eval). This benchmark consists of 2,443 high-quality samples, including true/false, multiple-choice, fill-in-the-blank, and short-answer questions. It covers five major themes: rare earth magnetic materials (abbreviated as “Magnetism”), rare earth luminescent materials (“Luminescence”), rare earth hydrogen storage materials (“Hydrogen Storage”), rare earth pyrometallurgy (“Metallurgy”), and rare earth testing and analysis (“Analysis”). While most previously mentioned Chinese benchmarks are characterized by being high-quality, reliable, static, and easy to evaluate, CREA-Eval possesses these traits alongside the following unique characteristics: (1) Standardization: The benchmark is based on industry standards; the question content is accurate, the format is standardized, and the terminology and symbols follow industry regulations or conventions. It contains no discriminatory content regarding age, gender, ethnicity, or region. (2) Validity: The benchmark content fully reflects industry standard requirements and does not involve capability factors unrelated to the evaluation objectives. The criteria for judging answers are reasonable and consistent, and the question difficulty progresses through multiple levels with a balanced configuration to fully measure the differences in rare earth capabilities between models. (3) Domain Specificity: All question content centers on themes related to the rare earth field, covering core scientific knowledge such as the basic physical/chemical properties, electronic structures, spectral behavior, and magnetic properties of rare earth elements. It also encompasses professional knowledge in key application areas such as materials science and permanent magnets. The benchmark reflects a multi-level knowledge structure

Figure 1

Figure 1: Figure 1

ranging from basic theory to cutting-edge technology.

CREA-Eval

CREA-Eval encompasses five primary themes: “Rare Earth Magnetic Materials,” “Rare Earth Luminescent Materials,” “Rare Earth Hydrogen Storage Materials,” “Rare Earth Pyrometallurgy,” and “Rare Earth Testing and Analysis.” Each theme includes four types of questions: true/false, multiple-choice, fill-in-the-blank, and short-answer questions. According to educational examination classifications, true/false, multiple-choice, and fill-in-the-blank questions are categorized as objective questions, primarily assessing the Large Language Model’ s (LLM) mastery of objective facts or conceptual knowledge. Short-answer questions are categorized as subjective questions, intended to broaden the scope of the examination and primarily evaluate the LLM’ s expressive or reasoning capabilities; the evaluation of these questions may be influenced by the specific judge model employed. compares CREA-Eval with several other evaluation benchmarks, clearly demonstrating that CREA-Eval is the world’ s first professional competency assessment benchmark focused specifically on the field of rare earths.

The data collection process for CREA-Eval combines automated script processing with manual intervention. As illustrated in

, the process involves: (1) manually extracting and filtering relevant knowledge from publicly collected materials to ensure that the gathered content is both professionally accurate and subject-relevant; (2) automatically generating questions, answers, and options using an LLM; (3) utilizing an LLM to filter the generated data, where only data meeting specific criteria proceed to the next stage; and (4) performing a manual audit of the data samples to ensure correct content and formatting, while filtering out invalid, non-standard, or irrelevant content not related to the rare earth field.

First, high-quality texts related to the rare earth field were collected from the internet across the five aforementioned thematic directions. These collected materials were manually filtered to ensure that the terminology and symbols used conform to industry standards or paradigms and are free from discriminatory content. Subsequently, an LLM utilizing few-shot prompting extracted questions, answers, and options from the text materials in a formatted output. Specifically, the initial step for generating true/false, multiple-choice, and fill-in-the-blank questions is identical: one or more declarative sentences are extracted verbatim from the original text to serve as the basis for the question. To generate true/false questions, the LLM randomly negates the extracted content based on a random variable (with a value of 0 or 1); negated items are labeled “False,” while non-negated items are labeled “True.” For multiple-choice questions, the

LLM generates a question and one correct option based on prompt examples, while the three distractors are sourced from elsewhere in the text. For fill-in-the-blank questions, the LLM divides the item into two parts based on prompts, where the main body serves as the question and the omitted portion serves as the answer. The procedure for generating short-answer questions differs significantly: the text is first segmented into natural paragraphs, with a maximum number of questions set for each block based on its size. The LLM then generates questions and answers within these text blocks according to the prompts. Once all data samples are created, an LLM is used to filter out samples that do not meet the dataset standards. Finally, a manual audit is conducted to ensure the dataset' s overall quality. Furthermore, formulas and units appearing in the data samples are manually proofread and converted into LaTeX or plain text format, with plain text preferred when it does not result in ambiguity.

Dataset Audit Criteria: (1) All questions must belong to the rare earth field and its related topics to ensure domain specificity. (2) All terms and symbols involved in the data samples must follow industry regulations or paradigms to ensure standardization. (3) All answers must be time-invariant. (4) All answers must align with the professional background; specifically, answers for true/false and multiple-choice questions must be correct and unique, answers for fill-in-the-blank questions must be clear, complete, and consistent, and answers for short-answer questions must be objective and consistent with the source text. During the LLM filtering stage, the LLM was required to audit the data based on these criteria, while scripts were employed to check for correct formatting and missing attributes. During the manual audit stage, auditors compared data samples against their original source texts, allowing them to easily verify compliance with the audit standards. Following the LLM filtering and manual audit, we randomly selected 100 samples for a secondary manual review; the results showed that all sampled data met the audit criteria. This high-quality, low-noise characteristic is sufficient to measure differences in rare earth domain capabilities among models, particularly for high-accuracy models.

Dataset Statistics

The statistical data for CREA-Eval is shown in . It consists of 5 topics, 4 question types, and a total of 2,443 samples, providing an efficient means to evaluate the boundaries of various LLMs' capabilities in the rare earth domain. Among these, the "Magnetic Materials" topic has the highest representation with 1,014 samples, as it constituted the largest portion of materials collected during the public data acquisition phase. The "Analysis" topic has the lowest representation with 347 samples, while "Pyrometallurgy," "Hydrogen Storage," and "Luminescence" contain 360, 362, and 360 samples, respectively. Regarding question types, there are 668 multiple-choice questions, 649 fill-in-the-blank questions, 661 true/false questions, and 465 short-answer questions.

We also analyzed the distribution of question types across different topics. Tak-

ing "Magnetic Materials" as an example, it includes 270 multiple-choice, 258 fill-in-the-blank, 268 true/false, and 218 short-answer samples. Across all topics, the proportion of short-answer questions is significantly lower than other types, primarily due to the high exclusion rate of this category during the LLM filtering and manual audit stages.

For the evaluation of fill-in-the-blank and short-answer questions, we utilized an LLM classifier to score and analyze the results of each model, specifically employing qwen3-max as the evaluator.

The classifier for fill-in-the-blank questions takes the question, the reference answer, and the model' s predicted answer as input, outputting either "Correct" or "Incorrect." The classifier for short-answer questions takes only the reference answer and the model' s predicted answer as input, outputting "Correct," "Incorrect," or "Not Attempted." The evaluation criteria for the different classifiers are as follows:

(1) Fill-in-the-blank evaluation criteria:

- Correct: The number and sequence of predicted answers correspond one-to-one with the reference answers, and each answer is equivalent to its corresponding reference answer within the context of the question.
- Incorrect: The number of predicted answers does not match the reference answers, the sequence differs, or the answers are not equivalent to the reference answers within the context of the question.

(2) Short-answer evaluation criteria (referenced from SimpleQA [**?**]):

- Correct: The predicted answer completely contains the reference answer and does not contradict it in any way.
- Incorrect: The predicted answer contradicts the reference answer in any respect, even if the contradiction is expressed in uncertain or speculative language.

- Not Attempted: The predicted answer does not fully provide the reference answer, yet it does not contradict the reference answer.

For multiple-choice and true/false questions, we employ a scoring method that combines regular expression matching with an LLM classifier. Specifically, the LLM classifier is utilized only when regular expression matching fails to yield a result. During our experiments, however, the LLM classifier was rarely triggered; such instances occurred almost exclusively when evaluating models with smaller parameter counts.

实验

The models participating in this evaluation were sourced from six different platforms, totaling 22 models. These include qwen3-max [**?**], qwen-flash, qwen-

plus, qwen2.5-72b-instruct [?], qwen2.5-32b-instruct, qwen2.5-14b-instruct, qwen2.5-7b-instruct, qwen2.5-3b-instruct, and Moonshot-Kimi-K2-Instruct [?] from Alibaba' s Bailian Cloud; hunyuan-t1-latest, hunyuan-turbo-latest [?], and hunyuan-large [?] from Tencent Cloud; deepseek-chat [?] and deepseek-reasoner [?] from the DeepSeek open platform; doubao-seed-1-6-250615, doubao-seed-1-6-thinking-250715 [?], and doubao-seed-1-6-flash-250828 from Volcano Engine; glm-4.5 [?, ?], glm-4.5-x, and glm-4.5-air from the Zhipu AI open platform; and ernie-4.5-turbo-latest [?, ?] and ernie-x1.1-preview from Baidu Qianfan. Some of these are commercial closed-source models, while others, such as the qwen-2.5 series and Moonshot-Kimi-K2-Instruct, are open-source. Notably, all model API calls were executed concurrently using a common thread pool with default settings.

Results and Discussion

The comprehensive performance of the models in addressing problems within the rare earth field is presented in . Several key observations can be made: (1) doubao-seed-1-6-thinking-250715 achieved the highest overall accuracy at 76.4%, ranking first and securing the top accuracy across all question types and themes, with the exception of "Analysis."However, this accuracy level suggests that significant knowledge gaps and risks of "hallucination" persist within the rare earth domain, making it difficult for the model to handle actual production tasks. (2) doubao-seed-1-6-250615 and Moonshot-Kimi-K2-Instruct ranked second and third in overall accuracy, respectively, while the judge model for this evaluation, qwen3-max, ranked sixth. (3) Excluding overall accuracy, the models that frequently appeared in the top three for individual statistical categories were doubao-seed-1-6-thinking-250715 (9 times), doubao-seed-1-6-250615 (7 times), Moonshot-Kimi-K2-Instruct (3 times), glm-4.5-x (3 times), ernie-x1.1-preview (4 times), and hunyuan-t1-latest (1 time). (4) qwen2.5-3b-instruct achieved an accuracy of 63.5% on true/false questions, which is only slightly higher than a random result (50%). (5) The performance of qwen2.5-72b-instruct was unexpected, as it outperformed deepseek-chat in the "Pyrometallurgy"and "Luminescence" themes. (6) All models exhibited significant performance variance across different themes, with most performing poorly in "Hydrogen Storage." Furthermore, accuracy across question types varied significantly; all models performed better on "Multiple Choice" and "True/False" questions than on "Fill-in-the-blank" and "Short Answer" questions. For "Fill-in-the-blank" questions, even the top-ranked model achieved only 59.8%, and the rate of improvement across rankings was lower than that for "Short Answer" questions, indicating that the models' grasp of rare earth industry concepts is neither precise nor robust. summarizes the results for the CREA-Eval short-answer questions, analyzing the models' ability to express factual knowledge or their reasoning capabilities. As shown in : (1) The two Doubao models performed best on subjective questions, holding the top F-score across all themes; specifically, doubao-seed-1-6-thinking-250715 ranked first in the "Correct," "Correct after Attempt," and F-score categories. (2) qwen2.5-3b-instruct had the highest proportion of "Incorrect"

Figure 3

Figure 2: Figure 3

responses, highlighting the limitations of small-parameter models in expression and reasoning. (3) hunyuan-turbo-latest had the highest "Unattempted" rate. (4) Regarding subjective questions, all models showed significant performance differences across themes; most performed better in "Magnetic Materials" and "Analysis" than in the other three themes, with the worst performance occurring in "Hydrogen Storage."

分析

We estimate the calibration of Large Language Models (LLMs) by requiring them to output a confidence score simultaneously when answering relevant questions. Generally, a well-calibrated model exhibits an actual accuracy rate that matches its reported confidence level. As shown in [FIGURE:2], with the exception of hunyuan-turbos-latest, all models demonstrate a positive correlation between declared confidence and actual accuracy. This phenomenon indicates that these models possess a certain degree of self-awareness regarding their own confidence levels. Furthermore, the Qwen2.5 series shows improved calibration as the parameter scale increases; specifically, within the $(40, 90]$ declared confidence interval, the performance follows the order: qwen2.5-72b-instruct > qwen2.5-32b-instruct > qwen2.5-14b-instruct > qwen2.5-7b-instruct > qwen2.5-3b-instruct. This observation is consistent with previous research findings. Within the $(90, 100]$ confidence interval, only qwen2.5-14b-instruct and qwen2.5-3b-instruct exhibit anomalous behavior. However, the curves for all models in [FIGURE:2] remain significantly below the ideal calibration line $(y = x)$, indicating that all models are overconfident and substantially overestimate their own certainty.

Analysis of Subjective and Objective Questions

Regarding the objective question accuracy of the Qwen2.5 and Qwen3 series, as well as the subjective question F-scores for the Doubao, Qwen2.5, and Qwen3 series shown in

(d), (e), and (f) respectively, several observations can be made: (1) The objective question performance of the Doubao and Qwen3 series appears more consistent across different topics.

The performance is balanced across other top-ranked models not listed here, which exhibit similar characteristics. (2) All models show uneven performance across different topics in subjective questions, reflecting that certain topics in CREA-Eval still pose significant challenges for current Large Language Models

(LLMs). (3) Most models perform relatively better on objective questions within the "Pyrometallurgy" topic but significantly worse on subjective questions. For convenience, we refer to this distinct performance gap between objective and subjective questions as "Objective-Subjective Decoupling." A similar phenomenon is observed in the "Hydrogen Storage" and "Luminescence" topics, though it is less pronounced than in "Pyrometallurgy." (4) Within the Qwen2.5 series, performance generally increases with model parameter scale across each topic, with only a few exceptions. Notably, the Qwen2.5 series also exhibits a strong "Objective-Subjective Decoupling" phenomenon in the "Hydrogen Storage" topic.

To eliminate the influence of difficulty variations between topics and question types as much as possible, we utilize the difference in cosine similarity to describe the "Objective-Subjective Decoupling" phenomenon for a specific topic j. The formula is as follows:

$$\mathrm{Diff}_j(A,F) = 10 \times \left(\frac{a_j f_j + \sum a_i f_i}{\sqrt{a_j^2 + \sum a_i^2}\sqrt{f_j^2 + \sum f_i^2}} - \frac{\sum a_i f_i}{\sqrt{\sum a_i^2}\sqrt{\sum f_i^2}} \right)$$

In Equation (**??**), A and F represent the objective question accuracy vector and the subjective question F-score vector, respectively. $D_j(A,F)$ represents the degree of "Objective-Subjective Decoupling" for topic j, while a_i and f_i are the corresponding values of objective accuracy and subjective F-score for different topics, where $i \in$ {"Magnetic Materials", "Analysis", "Hydrogen Storage", "Luminescence"}. The final calculation results are shown in . From these observations, we can conclude: (1) hunyuan-t1-latest, hunyuan-large, and doubao-seed-1-6-thinking-250715 rank in the bottom three for average decoupling values, indicating they exhibit the weakest overall "Objective-Subjective Decoupling." However, since the first two rank lower than the latter in total accuracy, doubao-seed-1-6-thinking-250715 can be considered the overall optimal model in this evaluation. (2) The performance of the Doubao and Qwen2.5 series is consistent with the results observed in Figure 3. (3) Models with a total accuracy greater than 70% that also rank among the bottom three (lowest decoupling) for specific topics include: ernie-4.5-turbo-latest and qwen-plus for "Magnetic Materials" ; hunyuan-t1-latest for "Pyrometallurgy" ; doubao-seed-1-6-flash-250828 and doubao-seed-1-6-250615 for "Analysis" ; and glm-4.5-x for "Luminescence." These models simultaneously demonstrate weak decoupling and high total accuracy. Notably, no model with a total accuracy exceeding 70% ranks in the bottom three for "Hydrogen Storage," suggesting that mainstream models tend to neglect this topic. This also implies that applying LLMs to problems related to "Hydrogen Storage" may carry certain risks. (4) The models exhibiting the strongest "Objective-Subjective Decoupling" for each topic are hunyuan-turbos-latest, glm-4.5-air, qwen2.5-72b-instruct, qwen2.5-7b-instruct, and glm-4.5-air, respectively.

Regarding the causes of this "Objective-Subjective Decoupling" phenomenon, we

hypothesize that it stems from a lack of expressive or reasoning capabilities in most LLMs for specific topics. During the model training process, the knowledge concepts or facts related to a specific topic may be derived from other fields or topics, and this knowledge has not been directly integrated into the context of the current topic.

These results present the objective question accuracy and subjective question F-score for the Doubao series, Qwen2.5 series, and Qwen3 series, respectively.

This leads to the organization of "supporting evidence" based on irrelevant content, text, or corpora, resulting in a deficiency in the model' s expressive or reasoning capabilities that fails to match its grasp of knowledge concepts regarding the subject.

The horizontal axis represents the number of models that answered correctly, while the vertical axis represents the corresponding number of data samples. In these plots, a peak near the right side indicates low difficulty, whereas a peak near the left side indicates high difficulty. As shown in [FIGURE:4] (f) and (h), the difficulty distributions for multiple-choice and true/false questions decrease progressively, with distinct peaks appearing only at the right end. In contrast, the distributions in (g) and (f) show peaks at both the left and right ends, indicating that fill-in-the-blank and short-answer questions contain significant numbers of both high-difficulty and low-difficulty samples. A similar difficulty distribution is observed across different topics: low-difficulty samples are the most numerous, with the sample count decreasing as difficulty increases, followed by a sudden surge in the high-difficulty range. These distributional characteristics stem from the domain-specific requirements of the rare earth field during dataset construction. Because current Large Language Models (LLMs) are not specifically adapted for the rare earth domain, a large number of professional questions challenge all models. Conversely, knowledge or questions that overlap between the rare earth field and general knowledge can typically be answered correctly by at least some models. This ultimately results in a "heavy at both ends, light in the middle" difficulty distribution.

结论

Abstract

This paper introduces CREA-Eval, the first Chinese evaluation benchmark specifically focused on the field of rare earth science and technology, filling a critical gap in assessing the professional capabilities of Large Language Models (LLMs) within the domain of rare earth materials. Characterized by its standardization, validity, and domain specificity, CREA-Eval achieves domain-aligned knowledge representation and verification, reflecting the cognitive boundaries of current Chinese LLMs in the rare earth sector. The benchmark comprises 2,443 samples across five core themes—including "Pyrometallurgy"

and "Analysis" —and four question types. It is designed to systematically evaluate the factual accuracy, conceptual understanding, and reasoning abilities of LLMs in this highly specialized field.

By calculating the accuracy and F-score of various models categorized by theme and question type, our evaluation results demonstrate that even the highest-ranking Chinese LLMs exhibit significant knowledge gaps and risks of "hallucination" when processing concepts related to rare earths. Furthermore, through a comparative analysis of subjective and objective questions, we identified a "subjective-objective dissociation"phenomenon in several models, which we quantitatively described using the difference in cosine similarity. Most models exhibit a strong "subjective-objective dissociation" in the "Hydrogen Storage" theme, suggesting that mainstream models tend to overlook this area. This finding implies that applying LLMs to problems related to hydrogen storage carries inherent risks.

The establishment of CREA-Eval not only provides a standardized tool for screening and optimizing domain-specific LLMs for the rare earth industry but also highlights the necessity of integrating specialized scientific knowledge into AI systems to ensure scientific reliability. This benchmark paves the way for developing trustworthy, knowledge-grounded AI assistants capable of effectively supporting material design, process optimization, and technical decision-making in practical scenarios, or serving as a scoring standard for specialized rare earth LLM competitions. We acknowledge certain limitations in this work: the current version focuses on textual knowledge and has not yet integrated experimental data or multi-modal inputs.

A preprint - April 13, 2026

"Magnetic Materials," "Pyrometallurgy," "Analysis," and "Hydrogen Storage" ; (f) multiple-choice questions, (g) fill-in-the-blank questions, (h) true/false questions, and (i) short-answer questions. The horizontal axis represents the number of models that correctly answered the corresponding data samples, while the vertical axis represents the number of data samples that were answered correctly.

Future work will expand CREA-Eval to include dynamic reasoning tasks and computational materials databases. Furthermore, we will explore fine-tuning strategies to embed rare-earth-specific physical constraints directly into the architecture of large language models.

HaluEval[5] SimpleQA[4] AlpacaEval[22] MMLU[23] WebQA[24] CMMLU[8] C-Eval[7] Chinese SimpleQA[9] CREA-Eval

Dataset size (records)

Alpaca Alpaca

LLM-as-a-Judge: Evaluating Large Language Models and Their Accuracy

The rapid advancement of Large Language Models (LLMs) has necessitated more sophisticated evaluation frameworks. Traditional metrics, such as BLEU or ROUGE, often fail to capture the nuances of human-like reasoning, creativity, and instruction-following capabilities. Consequently, the "LLM-as-a-Judge" paradigm has emerged as a scalable and effective alternative, utilizing powerful models like GPT-4 to evaluate the outputs of other models.

The Mechanism of LLM-based Evaluation

The core principle of LLM-as-a-Judge involves prompting a high-capability model to act as an evaluator. This judge model is typically provided with a prompt, a generated response, and sometimes a reference answer or a set of evaluation criteria (e.g., helpfulness, honesty, and harmlessness). The judge then provides a score or a qualitative assessment. This approach addresses the limitations of static benchmarks by allowing for a more dynamic and semantic understanding of model performance.

Accuracy and Alignment with Human Judgment

A critical concern in this paradigm is the accuracy of the LLM judge itself. Research indicates that while LLMs can achieve high levels of agreement with human annotators—often surpassing 80% correlation in specific tasks—they are not immune to systematic biases. These include:

- **Position Bias:** A tendency to favor the first response in a side-by-side comparison.
- **Verbosity Bias:** A preference for longer responses, regardless of their actual information density or correctness.
- **Self-Preference Bias:** A slight inclination toward outputs that mirror the judge model' s own stylistic patterns.

To mitigate these issues, researchers employ techniques such as swapping the order of responses, providing few-shot examples of high-quality evaluations, and using chain-of-thought (CoT) prompting to force the judge to justify its score before finalizing it.

Challenges in Measuring Absolute Accuracy

Measuring the "accuracy" of an LLM judge is inherently complex because, in many generative tasks, there is no single "ground truth." Instead, accuracy is often defined as the degree of alignment with expert human consensus. As models evolve, the gap between human judgment and LLM judgment continues to narrow, making LLM-as-a-Judge an indispensable tool for rapid iteration in machine learning pipelines. However, for high-stakes applications, human-in-the-loop verification remains essential to ensure the reliability

分析

Among these, magnetic materials, pyrometallurgy, analytical chemistry, hydrogen storage, and luminescence represent the accuracy rates categorized by subject theme, while multiple-choice, fill-in-the-blank, true/false, and short-answer questions represent the accuracy rates categorized by question type.

The models evaluated include: doubao-seed-1-6-thinking-250715, doubao-seed-1-6-250615, Moonshot-Kimi-K2-Instruct, ernie-x1.1-preview, glm-4.5, qwen3-max, hunyuan-t1latest, qwen-plus, glm-4.5-x, doubao-seed-1-6-flash-250828, deepseek-reasoner, ernie-4.5-turbolatest, glm-4.5-air, qwen-flash, deepseek-chat, hunyuan-turboslatest, qwen2.5-72b-instruct, qwen2.5-32b-instruct, hunyuan-large, qwen2.5-14b-instruct, qwen2.5-7b-instruct, and qwen2.5-3b-instruct.

分析

The statistical results include the number of correct, incorrect, and unattempted responses. From these primary metrics, the accuracy of attempted questions and the F-score are derived. Furthermore, thematic F-scores are calculated for specific subject areas, including Magnetic Materials, Pyrometallurgy, Chemical Analysis, Hydrogen Storage, and Luminescent Materials. Finally, the cosine similarity is utilized to measure the correlation between the accuracy of objective questions and the F-score of subjective questions.

Fscore

Magnetic Materials F-score

Pyrometallurgy F-score

Analysis F-score

Hydrogen Storage F-score

Luminescence F-score

doubao-seed1-6-thinking250715 doubao-seed1-6-250615 MoonshotKimi-K2Instruct qwen3-max glm-4.5 qwen-plus glm-4.5-x hunyuan-t1latest ernie-x1.1preview doubao-seed1-6-flash250828 qwen-flash

A preprint - April 13, 2026

The statistical results include the number of correct, incorrect, and unattempted responses. From these primary metrics, the accuracy of attempted questions and the F-score are derived. To provide a more granular analysis, F-scores are categorized by specific research themes, namely: Magnetic Materials, Pyrometallurgy, Chemical Analysis, Hydrogen Storage, and Luminescent Materials. Additionally, the cosine similarity is calculated to evaluate the correlation between the accuracy of objective questions and the F-score of subjective questions.

The performance of several models was evaluated, including: ernie-4.5turbo-latest, glm-4.5-air, deepseek-reasoner, deepseek-chat, qwen2.5-72b-instruct, hunyuan-turbo-latest, qwen2.5-32b-instruct, hunyuan-large, qwen2.5-14b-instruct, qwen2.5-7b-instruct, and qwen2.5-3b-instruct.

glm-4.5-air qwen2.5-7b-instruct qwen2.5-3b-instruct deepseek-chat qwen2.5-72b-instruct hungyuan-turbos-latest deepseek-reasoner ernie-x1.1-preview qwen2.5-14b-instruct doubao-seed-1-6-250615 glm-4.5 qwen3-max doubao-seed-1-6-flash-250828 Moonshot-Kimi-K2-Instruct qwen-plus ernie-4.5-turbo-latest glm-4.5-x qwen-flash qwen2.5-32b-instruct doubao-seed-1-6-thinking-250715 hungyuan-large hungyuan-t1-latest

分析

A preprint - April 13, 2026

References

[1] Chengcheng Liu, Xuandong Wang, Weidong Cai, and Hang Su. Prediction of magnetocaloric properties of fe-based amorphous alloys based on interpretable machine learning. 625:122749.

[2] Amit K. Choudhary, Andreas Jansche, Tvrtko Grubesa, Florian Trier, Dagmar Goll, Timo Bernthaler, and Gerhard Schneider. Grain size analysis in permanent magnets from kerr microscopy images using machine learning techniques. 186:111790.

[3] Akshat Chaudhari, Chakradhar Guntuboina, Hongshuo Huang, and Amir Barati Farimani. AlloyBERT: alloy property prediction with large language models. 244:113256.

[4] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models.

[5] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: a large-scale hallucination evaluation benchmark for large language models.

[6] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A chinese language understanding evaluation benchmark. In Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics.

[7] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models.

[8] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: measuring massive multitask language understanding in chinese. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, Findings of the Association for Computational Linguistics:

ACL 2024, pages 11260–11285. Association for Computational Linguistics.

[9] Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Dekai Sun, Zhicheng Zheng, Wenbo Su, and Bo Zheng.

Chinese SimpleQA: a chinese factuality evaluation for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), pages 19182–19208.

Association for Computational Linguistics.

[10] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. May 2025.

[11] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. January 2025.

[12] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu,

Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu,

A preprint - April 13, 2026

Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi K2: Open Agentic Intelligence. July 2025.

[13] Tencent Hunyuan Team, Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, Dian Jiao, Dong Du, Dong Wang, Feng Zhang, Fengzong Lian, Guanghui Xu, Guanwei Zhang, Hai Wang, Haipeng Luo, Han Hu, Huilin Xu, Jiajia Wu, Jianchen Zhu, Jianfeng Yan, Jiaqi Zhu, Jihong Zhang, Jinbao Xue, Jun Xia, Junqiang Zheng, Kai Liu, Kai Zhang, Kai Zheng, Kejiao Li, Keyao Wang, Lan Jiang, Lixin Liu, Lulu Wu, Mengyuan Huang, Peijie Yu, Peiqi Wang, Qian Wang, Qianbiao Xiang, Qibin Liu, Qingfeng Sun, Richard Guo, Ruobing Xie, Saiyong Yang, Shaohua Chen, Shihui Hu, Shuai Li, Shuaipeng Li, Shuang Chen, Suncong Zheng, Tao Yang, Tian Zhang, Tinghao Yu, Weidong Han, Weijie Liu, Weijin Zhou, Weikang Wang, Wesleye Chen, Xiao Feng, Xiaoqin Ren, Xingwu Sun, Xiong Kuang, Xuemeng Huang, Xun Cao, Yanfeng Chen, Yang Du, Zhen Yang, Yangyu Tao, Yaping Deng, Yi Shen, Yigeng Hong, Yiqi Chen, Yiqing Huang, Yuchi Deng, Yue Mao, Yulong Wang, Yuyuan Zeng, Zenan Xu, Zhanhui Kang, Zhe Zhao, ZhenXiang Yan, Zheng Fang, Zhichao Hu, Zhongzhi Chen, Zhuoyu Li, Zongwei Li, Alex Yan, Ande Liang, Baitong Liu, Beiping Pan, Bin Xing, Binghong Wu, Bingxin Qu, Bolin Ni, Boyu Wu, Chen Li, Cheng Jiang, Cheng Zhang, Chengjun Liu, Chengxu Yang, Chengzhong Xu, Chiyu Wang, Chong Zha, Daisy Yi, Di Wang, Fanyang Lu, Fei Chen, Feifei Liu, Feng Zheng,

Guanghua Yu, Guiyang Li, Guohua Wang, Haisheng Lin, Han Liu, Han Wang, Hao Fei, Hao Lu, Haoqing Jiang, Haoran Sun, Haotian Zhu, Huangjin Dai, Huankui Chen, Huawen Feng, Huihui Cai, Huxin Peng, Jackson Lv, Jiacheng Shi, Jiahao Bu, Jianbo Li, Jianglu Hu, Jiangtao Guan, Jianing Xu, Jianwei Cai, Jiarong Zhang, Jiawei Song, Jie Jiang, Jie Liu, Jieneng Yang, Jihong Zhang, Jin Lv, Jing Zhao, Jinjian Li, Jinxing Liu, Jun Zhao, Juntao Guo, Kai Wang, Kan Wu, Lei Fu, Lei He, Lei Wang, Li Liu, Liang Dong, Liya Zhan, Long Cheng, Long Xu, Mao Zheng, Meng Liu, Mengkang Hu, Nanli Chen, Peirui Chen, Peng He, Pengju Pan, Pengzhi Wei, Qi Yang, Qi Yi, Roberts Wang, Rongpeng Chen, Rui Sun, Rui Yang, Ruibin Chen, Ruixu Zhou, Shaofeng Zhang, Sheng Zhang, Shihao Xu, Shuaishuai Chang, Shulin Liu, SiQi Wang, Songjia Feng, Songling Yuan, Tao Zhang, Tianjiao Lang, Tongkai Li, Wei Deng, Wei Li, Weichao Wang, Weigang Zhang, Weixuan Sun, Wen Ouyang, Wenxiang Jiao, Wenzhi Sun, Wenzhuo Jia, Xiang Zhang, Xiangyu He, Xianshun Ren, XiaoYing Zhu, Xiaolong Guo, Xiaoxue Li, Xiaoyu Ma, Xican Lu, Xinhua Feng, Xinting Huang, Xinyu Guan, Xirui Li, Xu Zhang, Xudong Gao, Xun Luo, Xuxiang Qi, Yangkun Chen, Yangyu Tao, Yanling Xiao, Yantao Mai, Yanze Chen, Yao Ding, Yeting Yang, YiFan Song, Yifan Yang, Yijiao Zhu, Yinhe Wu, Yixian Liu, Yong Yang, Yuanjun Cai, Yuanlin Tu, Yue Zhang, Yufei Huang, Yuhang Zhou, Yuhao Jiang, Yuhong Liu, Yuhui Hu, Yujin Lin, Yun Yang, Yunhao Wang, Yusong Zhang, Zekun Wu, Zelong Zhang, Zhan Yu, Zhaoliang Yang, Zhe Zhao, Zheng Li, Zhenyu Huang, Zhiguang Liu, Zhijiang Xu, Zhiqing Kui, Zhiyin Zeng, Zhiyuan Xiong, Zhuo Han, Zifan Wu, Zigang Geng, Zilong Zhao, Ziyan Tang, Ziyuan Zhu, Zonglei Zhu, and Zhijiang Xu. Hunyuan-TurboS: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. July 2025.

[14] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, Jiahao Bu, Zhongzhi Chen, Xuemeng Huang, Fengzong Lian, Saiyong Yang, Jianfeng Yan, Yuyuan Zeng, Xiaoqin Ren, Chao Yu, Lulu Wu, Yue Mao, Jun Xia, Tao Yang, Suncong Zheng, Kan Wu, Dian Jiao, Jinbao Xue, Xipeng Zhang, Decheng Wu, Kai Liu, Dengpeng Wu, Guanghui Xu, Shaohua Chen, Shuang Chen, Xiao Feng, Yigeng Hong, Junqiang Zheng, Chengcheng Xu, Zongwei Li, Xiong Kuang, Jianglu Hu, Yiqi Chen, Yuchi Deng, Guiyang Li, Ao Liu, Chenchen Zhang, Shihui Hu, Zilong Zhao, Zifan Wu, Yao Ding, Weichao Wang, Han Liu, Roberts Wang, Hao Fei, Peijie Yu, Ze Zhao, Xun Cao, Hai Wang, Fusheng Xiang, Mengyuan Huang, Zhiyuan Xiong, Bin Hu, Xuebin Hou, Lei Jiang, Jianqiang Ma, Jiajia Wu, Yaping Deng, Yi Shen, Qian Wang, Weijie Liu, Jie

Liu, Meng Chen, Liang Dong, Weiwen Jia, Hu Chen, Feifei Liu, Rui Yuan, Huilin Xu, Zhenxiang Yan, Tengfei Cao, Zhichao Hu, Xinhua Feng, Dong Du, Tinghao Yu, Yangyu Tao, Feng Zhang, Jianchen Zhu, Chengzhong Xu, Xirui Li, Chong Zha, Wen Ouyang, Yinben Xia, Xiang Li, Zekun He, Rongpeng Chen, Jiawei Song, Ruibin Chen, Fan Jiang, Chongqing Zhao, Bo Wang, Hao Gong,

Rong Gan, Winston Hu, Zhanhui Kang, Yong Yang, Yuhong Liu, Di Wang, and Jie Jiang. Hunyuan-large: An open-source MoE model with 52 billion activated parameters by tencent. November 2024.

[15] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 technical report. February 2025.

[16] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li,

Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X.

Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via

A preprint - April 13, 2026

reinforcement learning. January 2025.

[17] ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, Lingjun Liu, Zhiqi Lin, Mingxuan Wang, Chengyi Wang, Xiangpeng Wei, Wenyuan Xu, Yufeng Yuan, Yu Yue, Lin Yan, Qiying Yu, Xiaochen Zuo, Chi Zhang, Ruofei Zhu, Zhecheng An, Zhihao Bai, Yu Bao, Xingyan Bin, Jiangjie Chen, Feng Chen, Hongmin Chen, Riwei Chen, Liangqiang Chen, Zixin Chen, Jinsong Chen, Siyan Chen, Kaiyuan Chen, Zhi Chen, Jin Chen, Jiecao Chen, Jinxin Chi, Weinan Dai, Ning Dai, Jiahui Dai, Shihan Dou, Yantao Du, Zhengyin Du, Jianhui Duan, Chen Dun, Ting-Han Fan, Jiazhan Feng, Junda Feng, Ziyuan Feng, Yuwei Fu, Wenqi Fu, Hanjie Fu, Hao Ge, Hongyi Guo, Mingji Han, Li Han, Wenhao Hao, Xintong Hao, Qianyu He, Jerry He, Feng He, Wen Heng, Zehua Hong, Qi Hou, Liang Hu, Shengding Hu, Nan Hu, Kai Hua, Qi Huang, Ziyue Huang, Hongzhi Huang, Zihao Huang, Ting Huang, Wenhao Huang, Wei Jia, Bin Jia, Xiaoying Jia, Yuhua Jiang, Haobin Jiang, Ziheng Jiang, Kaihua Jiang, Chengquan Jiang, Jianpeng Jiao, Xiaoran Jin, Xing Jin, Xunhao Lai, Zheng Li, Xiang Li, Liyi Li, Hongkai Li, Zheng Li, Shengxian Wan, Ya Wang, Yunshui Li, Chenggang

Li, Niuniu Li, Siyu Li, Xi Li, Xiao Li, Aoyan Li, Yuntao Li, Nianning Liang, Xinnian Liang, Haibin Lin, Weijian Lin, Ye Lin, Zhicheng Liu, Guanlin Liu, Guanlin Liu, Chenxiao Liu, Yan Liu, Gaohong Liu, Juncai Liu, Chundian Liu, Deyi Liu, Kaibo Liu, Siyao Liu, Qi Liu, Yongfei Liu, Kang Liu, Gan Liu, Boyi Liu, Rui Long, Weiqiang Lou, Chenwei Lou, Xiang Luo, Yao Luo, Caiping Lv, Heyang Lv, Bole Ma, Qianli Ma, Hongzhi Ma, Yiyuan Ma, Jin Ma, Wenchang Ma, Tingting Ma, Chen Mao, Qiyang Min, Zhe Nan, Guanghan Ning, Jinxiang Ou, Haojie Pan, Renming Pang, Yanghua Peng, Tao Peng, Lihua Qian, Lihua Qian, Mu Qiao, Meng Qu, Cheng Ren, Hongbin Ren, Yong Shan, Wei Shen, Ke Shen, Kai Shen, Guangming Sheng, Jinlong Shi, Wenlei Shi, Guang Shi, Shuai Shuai Cao, Yuxin Song, Zuquan Song, Jing Su, Yifan Sun, Tao Sun, Zewei Sun, Borui Wan, Zihan Wang, Xiaohui Wang, Xi Wang, Shuguang Wang, Jun Wang, Qinlong Wang, Chenyuan Wang, Shuai Wang, Zihan Wang, Changbao Wang, Jiaqiang Wang, Shihang Wang, Xuwu Wang, Zaiyuan Wang, Yuxuan Wang, Wenqi Wang, Taiqing Wang, Chengzhi Wei, Houmin Wei, Ziyun Wei, Shufa Wei, Zheng Wu, Yonghui Wu, Yangjun Wu, Bohong Wu, Shuang Wu, Jingqiao Wu, Ning Wu, Shuangzhi Wu, Jianmin Wu, Chenguang Xi, Fan Xia, Yuqiao Xian, Liang Xiang, Boren Xiang, Bowen Xiao, Zhen Xiao, Xia Xiao, Yongsheng Xiao, Chao Xin, Shulin Xin, Yuwen Xiong, Jingjing Xu, Ziwen Xu, Chenyin Xu, Jiayi Xu, Yifan Xu, Wei Xu, Yufei Xu, Shikun Xu, Shipeng Yan, Shen Yan, Qingping Yang, Xi Yang, Tianhao Yang, Yuehang Yang, Yuan Yang, Ximing Yang, Zeyu Yang, Guang Yang, Yifan Yang, Xuesong Yao, Bairen Yi, Fan Yin, Jianian Yin, Ziqiang Ying, Xiangyu Yu, Hongli Yu, Song Yu, Menghan Yu, Huan Yu, Siyu Yuan, Jun Yuan, Yutao Zeng, Tianyang Zhan, Zheng Zhang, Yun Zhang, Mofan Zhang, Wang Zhang, Ru Zhang, Zhi Zhang, Tianqi Zhang, Xinyi Zhang, Zhexi Zhang, Sijun Zhang, Wenqiang Zhang, Xiangxiang Zhang, Yongtao Zhang, Yuyu Zhang, Ge Zhang, He Zhang, Yue Zhang, Renjie Zheng, Ningxin Zheng, Zhuolin Zheng, Yaowei Zheng, Chen Zheng, Xiaoyun Zhi, Wanjun Zhong, Cheng Zhong, Zheng Zhong, Baoquan Zhong, Xun Zhou, Na Zhou, Huan Zhou, Hang Zhu, Defa Zhu, Wenjia Zhu, and Lei Zuo. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. April 2025.

[18] Glm-4 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai

Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang,

A preprint - April 13, 2026

Juanzi Li, Yuxiao Dong, and Jie Tang. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. August 2025.

[19] Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. July 2024.

[20] Baidu-ERNIE-Team. ERNIE 4.5 technical report, 2025.

[21] PaddlePaddle/ERNIE. PaddlePaddle, November 2025.

[22] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators, March 2025.

[23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, January 2021.

[24] Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. September 2016.