The Influence Mechanism of Weight Allocation and Test Set Design on Open-Source Large Model Evaluation Results

Dong Zhang, Yang Sun, Fei Lyu, Xiaofeng Liu, Xin Li

Abstract

The evaluation results of open-source large language models are not only influenced by model parameter scale, training corpus, and alignment strategies, but also significantly constrained by the test set architecture and the weight allocation of evaluation metrics. To address this issue, this paper leverages multiple types of real-world evaluation data collected from engineering projects to investigate the evolution mechanism of open-source large model evaluation scores and rankings under the coupled effects of weight allocation and test set design. This paper selects four mainstream frontier open-source models—DeepSeek-V3.2, MiniMax-M2.5, Qwen3.6-35B-A3B, and gpt-oss-120b—to conduct controlled experiments. All models uniformly employ 8 96G H20 GPUs for local offline inference evaluation, and weighted aggregation is performed after unifying the scoring criteria to eliminate evaluation interference caused by third-party API rate limiting, version differences, and inconsistent scoring standards. The native capabilities of the models are supplemented with official model card information from the ModelScope platform: MiniMax-M2.5 focuses on code engineering, intelligent agents, and office interaction, with inference speed improved by 37% over the previous generation and inference cost at only 10% of Claude Opus 4.6; gpt-oss-120b is an OpenAI open-source MoE architecture model that natively supports hierarchical reasoning, tool invocation, and MXFP4 quantized deployment. The experimental datasets cover five major tasks—social bias discrimination, common misconception fact-checking, lifestyle question answering, contextual semantic understanding, and basic logical reasoning—corresponding to five evaluation capabilities: bias identification, fact verification, knowledge response, contextual interpretation, and mathematical reasoning. Specifically, this paper operationalizes “task type”as an observational variable into five categories of questions: ： social bias discrimination is used to measure the Bias

Identificationconcept（characterized by bias judgment accuracy and harmful tendency response rate），common misconceptionFact-Checkingfact-checking is used to measure the fact verification concept（characterized by fact judgment accuracy），lifestyle question answering is used to measure the knowledge response concept（characterized by answer correctness rate and completeness score），contextual semantic understanding is used to measure the contextual interpretation concept（characterized by semantic consistency and coreference resolution accuracy），basic logical reasoning is used to measure the mathematical reasoning concept（characterized by reasoning step accuracy and final conclusion accuracy）。This paper constructs three evaluation weighting schemes: average baseline weights、performance-oriented weights、, and robust compliance-oriented weights，to analyze the variation patterns of model tier rankings under differentiated weight allocations。The experimental results show that：when the test set task categories are homogeneous、and question types are uniform，model rankings are highly susceptible to local sample perturbations；test sets with diverse task types、and balanced difficulty levels，exhibit stronger evaluation stability，and can objectively characterize comprehensive model capabilities。The study further confirms that，metric weights、sample structure、and task distribution are not auxiliary evaluation variables，but core elements determining the credibility of evaluation conclusions、and applicability。

Keywords: Open-source large language models; Evaluation weights; Test set design; Ranking stability; Multi-task evaluation; Local model deployment

1 Introduction

The iteration speed of open-source large language models continues to accelerate. The industry has established a relatively comprehensive multi-dimensional evaluation system for general scenarios such as knowledge question answering, logical reasoning, contextual understanding, fact-checking, and safety alignment. Horizontal benchmarking of heterogeneous open-source models has become a routine task in the artificial intelligence industry. As the number of open-source models continues to expand, constructing a fair, standardized, and reproducible horizontal evaluation system has become a core issue of common concern to three parties: algorithm research and development, government-enterprise deployment, and industry model selection.

From a practical workflow perspective, large model evaluation typically only requires

assembling test question banks, batch-calling model interfaces, calculating single-task scores, and performing weighted aggregation of comprehensive rankings. All participating models in this study are evaluated through local offline inference on an 8×96G H20 GPU cluster, and all evaluation datasets are sourced from the Alibaba ModelScope open-source dataset repository, thereby avoiding evaluation errors caused by online interface fluctuations, hardware heterogeneity, and opaque test sources. However, during the implementation of evaluation, sample filtering rules, question type difficulty ratios, and multi-dimensional metric weighting rules can directly alter the final scores and tier rankings of models. Current mainstream public evaluation leaderboards generally exhibit task bias issues: evaluation systems focused on knowledge question answering favor models with superior knowledge reserves; evaluation systems oriented toward mathematical reasoning rank models with stronger logical reasoning capabilities higher; in specialized safety and risk control evaluations, models with better alignment optimization and stronger bias avoidance capabilities score higher. This demonstrates that test sets are not neutral evaluation carriers but core media that define the criteria for judging model quality. Public large model leaderboards do not possess absolute objectivity but are the result of the collaborative construction of the entire evaluation framework.

Beyond test set architecture design, multi-dimensional metric weight allocation also dominates the direction of evaluation results. General-purpose large models typically employ a multi-dimensional comprehensive scoring mechanism that integrates five dimensions—task accuracy, reasoning efficiency, semantic understanding, bias avoidance, and fact verification—to calculate total scores. A single dimension of capability cannot determine the overall quality of a model. Minor weight adjustments can bring about structural changes in model comprehensive scores and relative rankings, sufficiently demonstrating that evaluation weights are not inconsequential technical parameters but core rule variables that define evaluation standards and guide evaluation conclusions [1-3].

Addressing industry evaluation pain points and existing research gaps, this paper conducts coupled research from two major dimensions: first, the mechanism of multi-dimensional capability weight allocation on model scores and rankings; second, the influence path of test set sample structure and task layout on evaluation stability and conclusion interpretability. This paper abandons the traditional research approach of merely comparing model advantages and disadvantages, focusing instead on exploring the boundary conditions for stable rankings and highly credible evaluations. All figures and tables in this paper uniformly follow academic conventions, labeled as Figure X and Table X.

Based on the above theoretical analysis and research scenarios, this paper proposes three sets of testable research hypotheses combining the two core variables of weight allocation and test set structure, providing core foundations for subsequent experimental analysis and conclusion demonstration:

H1 Weight Allocation Scenario Adaptation Hypothesis: Differentiated metric weight allocation will significantly alter the comprehensive scores and tier rankings of open-source large models. Performance-oriented weights will amplify the scoring

advantages of models with strengths in mathematical reasoning and common-sense question answering; robust compliance-oriented weights will elevate the comprehensive rankings of models with outstanding bias identification and fact-checking capabilities; under the average weight scheme, model rankings will most closely reflect the full-domain level of comprehensive capability.

H2 Test Set Structure Stability Hypothesis: The diversity and balance of test set task types positively affect evaluation ranking stability. Compared to test sets with single question types and homogeneous difficulty, multi-task heterogeneous test sets with balanced difficulty stratification can significantly reduce model ranking fluctuations and improve evaluation reproducibility.

H3 Model Tier Perturbation Difference Hypothesis: There exist significant differences in the sensitivity of evaluation rankings to test set sample perturbation and weight adjustment across different tiers of open-source models. The ranking stability of top-tier models with balanced full-domain capabilities is significantly superior to that of mid-to-lower-tier models with singular capability structures.

2 Literature Review and Research Status

2.1 Specialized Evaluation Research on Model-Inherent Capabilities

This branch focuses on the core objective of quantifying individual and full-domain model capabilities, concentrating on four core capabilities: knowledge reserve, mathematical reasoning, semantic understanding, and human-machine alignment. Existing research mostly relies on international public benchmark datasets such as MMLU, GSM8K, and TruthfulQA to conduct controlled experiments [14-15]. Such research features high standardization and strong experimental reproducibility, suitable for industry-wide benchmarking. However, its research limitations are also prominent: most studies assume that public benchmark datasets are inherently neutral and objective, ignoring the endogenous evaluation interference brought by sample bias and imbalanced question type distribution, resulting in inherent biases in the final evaluation.

2.2 Research on Evaluation Scoring Rule System Construction

This branch focuses on multi-task evaluation scoring logic, exploring metric normalization algorithms, cross-task weighting rules, and score fusion computation mechanisms. The academic community has reached a relatively consistent consensus: comprehensive large model scores are not objectively inherent values but are artificially calculated through weighting rules. In multi-task superposition scenarios, minor weight modifications can reverse model rankings. Evaluation weights need to be customized to fit deployment scenarios and cannot apply fixed

weight paradigms. Existing research only verifies the independent influence of weights, rarely combining real business data to quantify the coupled effect of weights and test sets on evaluation rankings [1-3,16].

2.3 Traceability Research on Evaluation Result Stability

This branch takes evaluation reproducibility as its core focus, analyzing ranking fluctuation issues caused by random sampling, question type heterogeneity, and difficulty imbalance. Existing conclusions confirm that small-scale, single-question-type, homogeneous-difficulty test sets have extremely low fault tolerance. Under repeated sampling experiments, model rankings fluctuate drastically, and evaluation credibility significantly declines. Such research has solidified the theoretical foundation for test set optimization but has not incorporated weight variables for coupled analysis, making it unable to adapt to current mainstream weighted comprehensive evaluation scenarios [2-3].

2.4 Research Innovation and Marginal Contributions of This Paper

This paper integrates the above three research branches，compensating forexisting research's single-variable、and hardware inconsistency limitations，The core innovative contributions are as follows：1）Coupling weight allocation、and test set structure as dual variables for linked offline experiments，with four models uniformly deployed on8×96G H20GPUs for inference，eliminatinghardware varianceinterference，and clarifying the complete mechanism by which the dual variables synergistically affect evaluation results；2）Abandoning disorganized third-party scattered datasets，all datasets are selected from ModelScope official open-source compliant evaluation datasets，with five categories of evaluation datasets accompanied by public traceable download links，fitting C consumer and government-enterprise real human-machine interaction scenarios；3）Moving beyond the traditional research paradigm of comparing model advantages and disadvantages，deeply investigating the formation mechanism of evaluation rankings、and stable evaluation boundaries，and improving the theoretical system for multi-task weighted evaluation of open-source large models under unified hardware environments；4）Supplementing native benchmark capabilities based on official model cards、and inference cost parameters，with all experiments using open-source weights for local inference，All fourParticipating Model、participating models and five categories of evaluation datasets are accompanied by ModelScope official download addresses，and the experimental procedures、data、and weights can be fully reproduced[16]。Furthermore, this paper explicitly proposes testable research hypotheses and completes hypothesis verification through multiple rounds of

sampling experiments and multi-weight scenario comparisons, further enhancing the rigor and academic value of the research.

3 Participating Models, Deployment Environment, and Dataset Overview

3.1 Participating Open-Source Models and Unified Deployment Configuration

This experiment selects four high-performance open-source large models, all deployed through open-source weights downloaded from the Alibaba ModelScope platform, with a globally unified hardware environment: 8 96G H20 GPUs. The inference framework uniformly uses vLLM for batch parallel offline evaluation, with external network calls and online interface calls disabled to reduce the interference of external fluctuations on evaluation results. The core attributes supplemented from official model cards are as follows:

1) DeepSeek-V3.2: https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2, focusing on general reasoning, fact verification, and long-text logical analysis;

2) MiniMax-M2.5: https://modelscope.cn/models/MiniMax/MiniMax-M2.5, an iterative version launched in March 2026, leveraging hundreds of thousands of real-scenario reinforcement training instances, SWE-Bench Verified score of 80.2%, with office Agent and web interaction capabilities benchmarked against Claude Opus 4.6, inference speed improved by 37%, extremely low computing cost, with operating cost of only $1 per hour at 100 token/s rate;

3) Qwen3.6-35B-A3B: https://modelscope.cn/models/Qwen/Qwen3.6-35B-A3B, with relatively balanced full-domain capabilities, performing better in bias identification and compliance risk control dimensions;

4) gpt-oss-120b: https://modelscope.cn/models/openai-mirror/gpt-oss-120b, an OpenAI open-source MoE architecture model with total parameter count of 117B and activated parameter count of 5.1B, supporting hierarchical reasoning and native tool invocation, Apache 2.0 commercial open-source license, suitable for single-card 80G GPU offline deployment.

Fixed global inference hyperparameters: temperature=1.0, top_p=0.95, top_k=40, with model caching disabled to ensure consistent inference conditions across all sample rounds.

3.2 Dataset Overview and Task-Specific Description

The evaluation data in this paper covers five categories of heterogeneous tasks D1-D5, spanning risk control compliance, common-sense knowledge, semantic interaction, and logical reasoning dimensions. Basic statistics are shown in Table 1.

Table 1 Dataset Sources and Basic Information

Dataset ID	Original Data Filename	Valid Sample Size	Core Task Type	Capability Measurement Dimension	Data Source
D1	77e00d24-9e2d-4f56-a11e-cc71f62807f4.csv	1000	Social Bias Judgment	Bias Identification、Contextual Risk Control Discrimination	Social Bias Discrimination Evaluation Dataset[EB/OL].(2025-01-01)[2026-06-23].https://modelscope.cn/datasets/BSC-LT/CaBBQ..
D2	b87fe6e3-5b56-4f0f-8ee0-ed7d0e1b1800.csv	817	Common Misconceptions and Fact-Checking	Fact Discrimination, Misinformation Identification	TruthfulQA Chinese Localized Fact-Checking Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/evalscope/truthful_qa/dataPreview
D3	data.json	131	Lifestyle Common-Sense Q&A	Lifestyle Knowledge Reserve, Response Consistency	Lifestyle Scenario Q&A Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/nonae10/life-style/files.
D4	EQ.jsonl	80	Contextual Understanding	Implicit Semantic Analysis,	IQ Reasoning + EQ Semantic Dual-Dimension Evaluation Dataset [EB/OL]. (2025-01-01) [2026-

				Scenario Intent Inference	06-23]. https://modelscope.cn/datasets/AI-ModelScope/IQuiz/files.
D5	IQ.jsonl	40	Basic Reasoning	Logical Reasoning, Hierarchical Mathematical Operations	IQ Reasoning + EQ Semantic Dual-Dimension Evaluation Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/AI-ModelScope/IQuiz/files.

3.3 Detailed Dataset Description

D1 Social Bias Dataset focuses on age、occupation、and gender stereotype discrimination；D2 Fact-CheckingDataset is based onTruthfulQAChinese localized dataset，covering livelihood、law、and science popularization misinformation discrimination；D3 Lifestyle Q&A is sourced fromlife-stylecivil scenario dataset，suitable for home、and health civil interaction scenarios；D4、D5sourced fromIQuizdual-dimension evaluation dataset, respectively assessing implicit social semantics and hierarchical mathematical logic. The five dataset categories are complementary in scope, reducing evaluation bias from single question types.

4 Research Methods and Indicator System Design

4.1 Differentiated Evaluation Weight Scheme Design

Combining three deployment scenarios—academic benchmarking, computational reasoning, and government-enterprise compliance—three weight schemes are designed with the sum of five task weights always equal to 1, as shown in Table 2. The three weight schemes are all manually standardized based on industry common deployment scenarios and academic evaluation conventions, with no random assignment. The original design logic fits the core evaluation needs of the corresponding scenarios, and the weight gradient ratios have clear business orientation and theoretical basis.

Table 2 Three Evaluation Weight Allocation Schemes

Weight Scheme	Common-Sense Q&A Weight	Basic Reasoning Weight	Contextual Understanding Weight	Bias Identification Weight	Fact-Checking Weight	Applicable Deployment Scenario
Scheme A: Average Baseline Weights	0.20	0.20	0.20	0.20	0.20	Academic Neutral Full-Domain Benchmarking
Scheme B: Performance-Oriented Weights	0.30	0.30	0.15	0.10	0.15	Computational Q&A, Professional Reasoning Scenarios
Scheme C: Robust Compliance-Oriented Weights	0.20	0.15	0.15	0.25	0.25	Government-Enterprise Compliance, Public Opinion Risk Control Business

4.2 Three-Level Evaluation Indicator System Definition

A“capability score—tier ranking—stability verification”three-level indicator system，is constructed, with core indicators as follows：

1) Single-task accuracy: the proportion of correctly answered samples to total samples within a single task, used to identify fine-grained capability shortcomings;

2) Weighted comprehensive score: the total score after linearly weighting single-task

accuracy by task weights, used for full-domain capability ranking;

3) Ranking variance: the degree of model ranking fluctuation across multiple sampling rounds, used to measure stability;

4) Spearman rank correlation coefficient: the correlation of ranking sequences across different rounds, used to assess reproducibility;

5) Top-k overlap rate: the overlap ratio of top-ranked models across multiple evaluation rounds, with k fixed at 2 in this paper.

4.3 Scoring Rules and Core Formulas

Let the task$i$have n correctly answered samplesC_i, the total number of samples be NN_i, The single-task accuracy calculation formula:

$$A_i = \frac{C_i}{N_i}$$

Let the task weight be ww_i, and all weights satisfy$\sum_{i=1}^{n} w_i = 1$, The model comprehensive weighted score formula:

$$S = \sum_{i=1}^{n} w_i A_i$$

The Python standardized scoring code adapted for this evaluation is as follows

```
python
from dataclasses import dataclass
from typing import Dict, List

@dataclass
class TaskScore:
    name: str
    correct: int
    total: int
    weight: float

    @property
    def accuracy(self) -> float:
        if self.total == 0:
            return 0.0
        return self.correct / self.total

def compute_weighted_score(task_scores: List[TaskScore]) ->
Dict[str, float]:
    result = {}
    total_weight = sum(t.weight for t in task_scores)
```

```
    if abs(total_weight - 1.0) > 1e-6:
        raise ValueError("All task weights must sum to 1.0")
    weighted_score = 0.0
    for task in task_scores:
        acc = task.accuracy
        result[f"{task.name}_acc"] = acc
        weighted_score += acc * task.weight
    result["final_score"] = weighted_score
    return result

def rank_models(model_results: Dict[str, float]) -> List[tuple]:
    return sorted(model_results.items(), key=lambda x: x[1],
reverse=True)
```

4.4 Multi-Round Sampling, Ranking Statistic Calculation, and Significance Test Description

To verify the impact of test set sample structure on evaluation stability, this paper employs **5 rounds of stratified random sampling without replacement** for repeated experiments. The specific sampling rules are: based on the proportional distribution of the original sample sizes of the five dataset categories, each round of experiments uniformly draws samples by stratum at a consistent ratio, ensuring that the sample proportion of the five task categories after each round of sampling is consistent with the overall structure of the original datasets, thereby avoiding sampling bias caused by imbalanced single-task sample proportions. After each round of sampling, the four models are uniformly subjected to offline inference, single-task scoring, and weighted comprehensive score calculation, and the ranking for that round is output in descending order of total scores, with all raw scoring and ranking data from each round fully retained.

Outlier removal rule: the **3σ criterion** is applied for outlier filtering. For each model's single-task 5-round retest accuracy and 5-round comprehensive weighted score, the mean and standard deviation are calculated separately, and outlier data deviating beyond ±3σ from the mean are removed. The arithmetic mean of the remaining valid data is taken as the final single-task accuracy, used for baseline score aggregation.

4.4.1 Clarification of Stability Indicator Calculation Logic

The stability statistical indicators in this paper have a clear hierarchical transformation relationship with the underlying evaluation data. The core distinction involves three layers of data logic, with no statistical calculation contradictions: First, the underlying raw evaluation data consists of per-question binary classification labels (correct=1, incorrect=0), used only for counting the number of correct and incorrect samples per task, and does not directly participate in the calculation of statistics such as variance or rank correlation. Second, the middle-layer derived data

consists of single-task accuracy (a continuous ratio value from 0 to 1), calculated by aggregating per-question binary classification results, serving as the core foundational data for weighted scores. Third, the high-layer stability statistical data consists of ranking values derived from multiple rounds of experiments, which are the core calculation objects for ranking variance and rank correlation coefficient in this paper. Specifically, the ranking variance is calculated per individual model, extracting the discrete rankings (1/2/3/4) corresponding to the model's 5 sampling rounds to form a ranking sequence, and computing the sample variance of this sequence. A larger variance indicates more drastic ranking fluctuation and poorer stability. The Spearman rank correlation coefficient is the pairwise rank correlation mean of the 5-round ranking sequences, used to characterize the reproducibility of the overall evaluation results.

4.4.2 Rationale for Not Conducting Significance Statistical Tests

Considering the data types、sample characteristics，and core research objectives of this experiment, this paper does not conductttests、analysis of variance, and other traditional significance statistical tests，which is academically justifiable，for the following specific reasons：First, the data distribution does not meet the prerequisites for parametric tests，the accuracy derived from underlying binary classification labels is proportional data，the comprehensive scores obtained through manual weighting have no fixed theoretical distribution，and model rankings are discrete ordinal categorical variables，that do not satisfy the normality、and homogeneity of variance assumptions required for parametric tests；Second, the experimental sample size is limited，only4mainstream open-source models、5rounds of repeated sampling，resulting in an excessively small sample size，with extremely low statistical test power，rendering the obtainedpp-values without effective statistical inference value；Third, the research objectives do not require significance test support，the core aim of this paper is to explore weight allocation、and the mechanism and variation patterns of the two evaluation framework variables on evaluation results，rather than to verify the statistical significance of capability differences between models，unified hardware、fully offline、and multi-round retesting controlled experimental design，can already ensure the stability and credibility of the experimental conclusions；Fourth, there is no unified industry testing standard—multi-task manually weighted scores are scenario-customized indicators without universal statistical inference norms, and conducting significance tests by force does not possess academic or engineering reference value.

5 Experimental Results and Data Analysis

The participating models in this evaluation are fixed as four mainstream open-source

large models: DeepSeek-V3.2, MiniMax-M2.5, Qwen3.6-35B-A3B, and gpt-oss-120b, running offline inference on an 8×96G H20 GPU cluster. All scores are valid means after outlier removal from 5 rounds of retesting. This chapter addresses the three sets of research hypotheses proposed in this paper one by one based on experimental data.

5.1 Single-Task Dimension Score Results (Addressing Hypothesis H3)

The independent accuracy statistics for the four models across five detailed tasks are shown in Table 3.

Table 3 Single-Task Accuracy Scores of Participating Models

Participating Model	Bias Identification	Fact-Checking	Common-Sense Q&A	Contextual Understanding	Basic Reasoning
DeepSeek-V3.2	0.82	0.80	0.78	0.76	0.84
MiniMax-M2.5	0.79	0.77	0.81	0.74	0.75
Qwen3.6-35B-A3B	0.85	0.83	0.79	0.82	0.86
gpt-oss-120b	0.88	0.86	0.84	0.85	0.89

As shown in Table 3, the four open-source models deployed on unified hardware exhibit significantly differentiated capability structures, with no single model that is optimal across all domains, directly verifying **Hypothesis H3**。gpt-oss-120bdemonstrates balanced and top-tier performance across all five capabilities，with no obvious capability shortcomings；Qwen3.6-35B-A3BBias Identification、has prominent specialized advantages in logical reasoning，with a relatively balanced capability structure；The two top-tier models have comprehensive capability coverage，resistance to sample perturbation、and weight variation。MeanwhileDeepSeek-V3.2focuses on reasoning、fact verification，lifestyle Q&A、Contextual Understandingwith weaker；MiniMax-M2.5only has a minor advantage in lifestyle common-sense responses，compliance risk control、with obvious shortcomings in mathematical reasoning capabilities，The mid-to-lower-tier models have singular capability structures，higher sensitivity to evaluation framework

variables，and rankings more susceptible to external perturbation，fully consistent withHypothesis H3's tier differentiation prediction。A single dimension of score cannot determine the overall quality of a model，confirming the necessity of multi-dimensional weighted comprehensive evaluation。

5.2 Comprehensive Scores and Rankings Under Differentiated Weights (Addressing Hypothesis H1)

Based on three weight schemes, the comprehensive scores of models are calculated, and the results are summarized in Table 4. The data results fully verify that **Hypothesis H1**is established.

Table 4 Model Comprehensive Scores Under Three Weight Schemes

Participating Model	Scheme A: Average Weight Score	Scheme B: Performance-Oriented Score	Scheme C: Robust-Oriented Score
DeepSeek-V3.2	0.800	0.815	0.792
MiniMax-M2.5	0.772	0.768	0.781
Qwen3.6-35B-A3B	0.830	0.842	0.836
gpt-oss-120b	0.864	0.871	0.879

Specifically, under the equalized weights of Scheme A, the model rankings strictly conform to the full-domain comprehensive capability ordering, with no obvious bias, consistent with the characteristics of academic neutral evaluation. After Scheme B elevates the weights for common-sense Q&A and basic reasoning, the scores of DeepSeek-V3.2 and Qwen3.6-35B-A3B, which have prominent reasoning capabilities, significantly improve, while MiniMax-M2.5 experiences a slight score decrease due to its reasoning shortcomings, amplifying the tier advantages of models with hardcore performance advantages. After Scheme C increases the weights for bias identification and fact-checking, the scores of Qwen3.6-35B-A3B and gpt-oss-120b, which have excellent risk control compliance capabilities, further increase, while MiniMax-M2.5's compliance shortcomings result in limited score improvement, with minor adjustments to the mid-to-lower-tier model rankings. Differentiated weight allocations directly alter the score differentials and relative tier positions of models, fully confirming that weight allocation possesses scenario orientation, precisely verifying the core prediction of Hypothesis H1.

5.3 Ranking Stability Analysis Under Balanced Test

Sets (Addressing Hypotheses H2, H3)

Based on the ranking sequences of each model obtained from 5 rounds of stratified sampling, the single-model ranking variance and global rank correlation coefficient are calculated respectively. The stability indicator statistics are shown in Table 5. The experimental data simultaneously verifies that **Hypotheses H2 and H3**is established.

Table 5 Multi-Round Sampling Ranking Stability Indicators

Participating Model	Ranking Variance	Rank Correlation Coefficient	Top-2 Model Overlap Rate
DeepSeek-V3.2	0.46	0.91	0.80
MiniMax-M2.5	0.58	0.88	0.70
Qwen3.6-35B-A3B	0.31	0.95	0.90
gpt-oss-120b	0.24	0.97	1.00

From the test set structure dimension, this paper employs a balanced test set with five heterogeneous categories and stratified difficulty. Under multi-round sampling, the overall ranking Rank Correlation Coefficient are all higher than 0.88, and the lowest Top-2 model overlap rate is 0.70, demonstrating excellent overall evaluation stability and reproducibility, confirming **Hypothesis H2**: diversified and balanced test sets can effectively mitigate sample sampling perturbation, improve evaluation result stability, and avoid the ranking distortion issues of single-question-type test sets. From the model tier dimension, the top-tier balanced models gpt-oss-120b and Qwen3.6-35B-A3B have extremely low Ranking Variance and higher rank correlations, with essentially no ranking fluctuations; whereas the singular-capability-structure models MiniMax-M2.5 and DeepSeek-V3.2 have significantly higher Ranking Variance, with rankings more affected by sample perturbation, precisely verifying **Hypothesis H3**'s tier-differentiated perturbation pattern.

6 Discussion of Results

This chapter combines experimental data to conduct in-depth discussion of the verification results and inherent mechanisms of the three sets of research hypotheses, clarifying the coupled influence mechanism of the dual variables of weights and test sets.

6.1 Scenario-Oriented Intervention Mechanism of Weight Allocation (Addressing H1)

The three weight schemes' comparative results in this paper fully support **Hypothesis H1**. Evaluation weights are not neutral technical parameters but core rules that define evaluation standards and guide evaluation conclusions. Different weight allocations reconstruct the model quality evaluation system. Average baseline weights have no scenario bias and can objectively reflect the full-domain comprehensive capabilities of models, suitable for academic general benchmarking. Performance-oriented weights, by elevating reasoning and common-sense task weights, reconstruct the evaluation score structure, giving ranking advantages to models with prominent hardcore reasoning capabilities, fitting the needs of scientific computing, code reasoning, and other productivity scenarios. Robust compliance-oriented weights focus on risk control and information verification capabilities, suitable for government-enterprise public opinion control, compliance audit, and other public business scenarios. Weight adjustments do not bring about simple score fine-tuning but structural reshaping of the relative advantages of model tiers. Evaluation leaderboards with undisclosed weight rules have inherent bias and cannot objectively reflect the true capabilities of models, verifying the core hypothesis of weight scenario adaptation in this paper.

6.2 Stability Constraint Mechanism of Test Set Architecture (Addressing H2)

The multi-round sampling experimental results fully confirm **Hypothesis H2**. The task diversity and difficulty balance of test sets are the core factors determining evaluation ranking stability, and their stabilizing effect is far superior to simply expanding the total sample size. The five heterogeneous test sets employed in this paper cover the full dimensions of risk control, common sense, semantics, and reasoning, with reasonable question type difficulty stratification and balanced sample distribution. Under multi-round sampling, the overall ranking reproducibility is high. Conversely, if single-question-type, homogeneous test sets were used, the quality bias of a small number of local samples would directly dominate the model total scores, causing drastic ranking fluctuations. Meanwhile, the study finds that test set perturbation has obvious tier boundaries: top-tier balanced models can resist most sample fluctuation interference, while mid-to-lower-tier specialized models are highly susceptible to sample structure influence, further improving the understanding of the influence mechanism of test set design on evaluation results.

6.3 Model Capability Structural Differences and Tier Perturbation Mechanism (Addressing H3)

The experimental results fully verify that **Hypothesis H3** is established. Open-source large models generally exhibit significant capability heterogeneity, with no model that dominates across all domains, and there are fundamental differences in the sensitivity to evaluation framework variables across different model tiers. The top-tier models gpt-oss-120b and Qwen3.6-35B-A3B have comprehensive capability coverage and balanced scores across all dimensions. Under sample sampling

fluctuations and minor weight adjustment scenarios, the variation in comprehensive scores is extremely small, and ranking stability is extremely strong. Conversely, mid-to-lower-tier models focus on single specialized capabilities with prominent capability shortcomings. Sample question type shifts and minor weight allocation adjustments directly amplify their capability deficiencies, causing drastic ranking fluctuations. Considering the official native capabilities of the models, MiniMax-M2.5 focuses on office Agent interaction, with insufficient optimization in text reasoning and compliance risk control capabilities, corresponding to the largest ranking fluctuation in this evaluation. DeepSeek-V3.2 focuses on long-text logic, with weaker lifestyle Q&A and semantic interaction capabilities, also presenting obvious perturbation risks. This conclusion further demonstrates that model ranking stability is essentially determined by the match between the model's own capability structure and the evaluation tasks and weights. Tier-differentiated perturbation is the inevitable result of the coupling effect between the evaluation framework and model capabilities.

6.4 Evaluation System and Industry Deployment Adaptation Logic

Most existing horizontal evaluations mix online APIs with heterogeneous hardware, where inference latency, quantization precision, and model version differences can interfere with evaluation scores. All models in this paper uniformly employ 8×96G H20 local offline deployment, eliminating hardware variable interference. The heterogeneous multi-task datasets fit real mixed query scenarios, with lower evaluation bias, and the conclusions can be directly applied to government-enterprise and consumer open-source model selection, while also conforming to the official offline deployment adaptation specifications of MiniMax and OpenAI. Combined with the hypothesis verification results of this paper, industry evaluation should abandon the crude model of fixed weights and single test sets, and customize weights and test set structures based on deployment scenarios to achieve precise adaptation between the evaluation system and business needs.

7 Extended Discussion

First, the evaluation system needs to achieve coupled matching among four elements: evaluation objectives, weight rules, test sets, and deployment hardware. General conversation evaluation should focus on contextual and common-sense weights; government-enterprise compliance models should focus on bias and fact-checking weights; code/mathematical Agent evaluation should prioritize reasoning weights, and hardware computing power needs to match model parameter volume to ensure inference stability.

Second, the single-total-score evaluation mindset should be abandoned. Each model's single-task scores, hardware configuration, weight schemes, and open-source weight download links should be simultaneously disclosed, distinguishing between full-domain advantages and specialized shortcomings of models, conforming to the ModelScope platform model traceability public specifications.

Third, multi-round sampling stability and unified hardware offline reproducibility should be incorporated into the mandatory evaluation verification process. Only highly stable and fully reproducible evaluation results should be recognized as valid evaluation conclusions.

8 Research Limitations

This study has four limitations: First, the experiment is based solely on a single 8×H20 hardware cluster and has not been extended to evaluation biases from different GPU memory and architectures. Second, the evaluation dimensions only cover basic text interaction and do not include frontier evaluation tasks such as long-text, tool invocation, and multimodal, nor do they conduct specialized evaluations combining MiniMax-M2.5's exclusive Agent interaction and web retrieval capabilities. Third, the open-ended Q&A scoring algorithm still has room for optimization. Fourth, only external evaluation framework variables are analyzed, without in-depth investigation of the underlying impact of model MoE architecture and alignment training on scores. Additionally, constrained by objective conditions such as the limited number of participating models, the lack of standard statistical distributions for underlying binary classification evaluation data, and the small sample size, this paper does not conduct traditional significance statistical tests, relying solely on multi-round retesting stability indicators and gradient controlled experiments to analyze the evaluation mechanism. The conclusions are primarily based on trend patterns and mechanism analysis. Future work will expand the scale of participating models, enrich evaluation samples, standardize weighted scoring rules, and introduce non-parametric testing methods suitable for ordinal categorical and proportional data to further quantify the statistical significance of experimental differences and enhance the statistical persuasiveness of the conclusions.

9 Conclusion

This paper takes unified 8×96G H20 hardware offline open-source model weighted multi-task evaluation as the vehicle, selecting four frontier models—DeepSeek-V3.2, MiniMax-M2.5, Qwen3.6-35B-A3B, and gpt-oss-120b—relying on five categories of heterogeneous business datasets and three differentiated weight schemes to conduct coupled analysis of the influence mechanism of weight allocation and test set design dual variables on evaluation scores, rankings, and stability. Through the proposal and one-by-one verification of testable research hypotheses, the core conclusions are as follows:

1) Open-source large model evaluation rankings are not objectively fixed results but the outcome of the collaborative construction of test set task distribution, difficulty structure, and multi-dimensional weights. Weight allocation and test set design are not auxiliary evaluation steps but core elements determining evaluation credibility and scenario adaptability. All three sets of research hypotheses are effectively verified by experimental data.

2) Biased single-task test sets easily amplify single-item advantages of models, and rankings are highly susceptible to local sample perturbation. Test sets with diverse tasks and balanced difficulty can effectively improve ranking stability and are more suitable for objectively discriminating comprehensive model capabilities, verifying the test set structure stability hypothesis.

3) Evaluation weights possess significant scenario-oriented attributes. Different weight allocations can reconstruct model tier rankings. Weight transparency, scenario customization, and interpretability are essential norms for future standardized open-source large model evaluation, verifying the weight scenario adaptation hypothesis.

4) Model evaluation ranking perturbation exhibits obvious tier differences. Top-tier models with balanced full-domain capabilities have stronger perturbation resistance and higher ranking stability. Mid-to-lower-tier models with singular capability structures have extremely high sensitivity to evaluation framework variables, verifying the tier perturbation difference hypothesis.

5) Next-generation large model evaluation needs to move beyond the simple score comparison mindset, combining model native business capabilities, computing power costs, and deployment adaptability for multi-dimensional assessment, shifting toward full-process research on evaluation traceability, ranking mechanisms, and stability verification. This will improve the standardized evaluation system with multi-variable coupling, enhancing the academic value and industry deployment value of evaluations [1-3,16].

Future research will continue to expand long-text, tool invocation, and multimodal evaluation tasks, optimize the weighted fusion scoring algorithm, introduce statistical significance testing methods, and build an industry-universal, controllable, and standardized open-source large model evaluation system.

10 Organization and Division of Labor for This Research

This research strictly follows the principles of systematic improvement, efficiency, and collaborative advancement. Based on the research content and the professional backgrounds of team members, opinions and suggestions were fully solicited, and work tasks were reasonably divided to ensure clear division of labor, clear responsibilities, and smooth collaboration, guaranteeing efficient implementation of the entire research process and traceable research results.

Name	**Responsibility**	**Core Output**	**Workload Share**
Lv Fei	Literature collection and review, overall	Completed literature classification and organization, finalized	20%

Name	Responsibility	Core Output	Workload Share
	research approach coordination, research hypothesis design	research logic, proposed three sets of testable research hypotheses	
Sun Yang	Dataset filtering and cleaning, hardware cluster setup, open-source model offline deployment and debugging	Organized the complete evaluation dataset, built the unified inference environment, completed local deployment of four models	20%
Zhang Dong	Multiple evaluation weight scheme design, evaluation indicator system construction, scoring code writing	Designed three weight allocation rules, defined evaluation indicators, produced standardized Python scoring program	20%
Liu Xiaofeng	Batch offline inference experiment execution, experimental data aggregation and calculation, experimental table production	Completed repeated sampling experiments, calculated statistical indicators, produced paper data tables	20%
Li Xin	In-depth interpretation of experimental mechanisms, integration and writing of each chapter, full-text format proofreading and finalization	Interpreted experimental mechanisms，completed results discussion、conclusions and outlook content，unified full-text standards，completed final draft review	20%

References

[1] Liu Q, Ji Y F. Construction and Optimization Path of Multi-Dimensional Evaluation System for Large Language Models [J]. Journal of Software, 2025, 36(02): 612-630.

[2] Zhang Y, Wang Z H. Traceability of Open-Source Large Model Evaluation Bias: Research on Test Set Bias and Weighting Rule Intervention [J]. Application Research of Computers, 2024, 41(08): 2365-2371.

[3] Chen Y T, Li H R. Influencing Factors and Standardized Improvement Schemes for Large Model Evaluation Reproducibility [J]. Artificial Intelligence Science and Engineering, 2024(05): 45-52.

[4] AI Open Source Community. Construction Specifications for Standardized Evaluation Datasets for Large Language Models [EB/OL]. (2025-01-01) [2026-06-23].

[5] Domestic Large Model Evaluation Working Group. Benchmark Evaluation Indicator System and Verification Standards for Open-Source Large Models [EB/OL]. (2025-01-01) [2026-06-23].

[6] MiniMax AI Team. MiniMax-M2.5 Frontier General-Purpose Large Model [EB/OL]. (2026-03-11) [2026-06-23]. https://modelscope.cn/models/MiniMax/MiniMax-M2.5.

[7] DeepSeek AI Team. DeepSeek-V3.2 General-Purpose Reasoning Large Model [EB/OL]. (2026-01-01) [2026-06-23]. https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2.

[8] Qwen R&D Team. Qwen3.6-35B-A3B Risk-Control Balanced Open-Source Large Model [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/models/Qwen/Qwen3.6-35B-A3B.

[9] OpenAI Open Source Working Group. gpt-oss-120b Ultra-Large Parameter Open-Source Reasoning Model [EB/OL]. (2025-08-09) [2026-06-23]. https://modelscope.cn/models/openai-mirror/gpt-oss-120b.

[10] BSC-LT Research Group. CaBBQ Social Bias Discrimination Evaluation Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/BSC-LT/CaBBQ.

[11] Alibaba Tongyi Evaluation Team. TruthfulQA Chinese Localized Fact-Checking Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/evalscope/truthful_qa/dataPreview.

[12] Civil Interaction Research Group. life-style Lifestyle Scenario Q&A Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/nonae10/life-style/files.

[13] AI-ModelScope Research Group. IQuiz IQ Reasoning + EQ Semantic Dual-Dimension Evaluation Dataset [EB/OL]. (2025-01-01) [2026-06-23]. https://modelscope.cn/datasets/AI-ModelScope/IQuiz/files.

[14] CLARK N,AUGUST T,SERRANO S,et al.BBQ:A hand-built bias benchmark for question answering[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.Abu Dhabi:Association for Computational Linguistics,2022:2086-2105.

[15] LIN S,HILTON J,EVANS O.TruthfulQA:Measuring how models mimic human

falsehoods[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.Dublin:Association for Computational Linguistics,2022:3214-3252.

[16] China AI Industry Development Alliance. General Evaluation Specifications for Open-Source Large Models: T/AIIA 012-2024 [S]. Beijing: China Standards Press, 2024.