International Journal of Innovative Computing, Information and Control
Volume **22**, Number **3**, June **2026**

ICIC International ©2026 ISSN 1349-4198
pp. **817–836**

METROEVAL: A LARGE LANGUAGE EVALUATION FRAMEWORK IN THE METRO DOMAIN

Tianzhen Lin, Hengyu Liu, Baiping Liu, Jifeng Liu, Cui Wang and Ning Li

R&D Center
Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd.
No. 9, Jinxia Road, High-Tech Zone, Qingdao 266111, P. R. China
{ fortune157; fortune158; fortune779; fortune159; fortune805 }@bnfortune.com; qdliubp@126.com

Received July 2025; revised November 2025

Abstract. *Existing studies on the evaluation benchmarks and approaches of large language models (LLMs) primarily concentrate on accuracy, robustness, bias and security. As the application of LLMs in the metro domain intensifies, it is imperative to construct a specific LLM evaluation framework and benchmark for evaluating LLMs. We collected the official information released via metro companies' official websites, WeChat official accounts, Weibo, Douyin, as well as other professional sources in the metro domain. After data cleaning, the original data were categorized into four types: passenger service, line operation and maintenance, emergency management, and safety operation, and the training dataset, evaluation benchmark and criteria were constructed. Then, multiple agents are constructed through fine-tuning the Qwen-7B model and prompt engineering, and MetroEval framework is established by multi-agent collaboration to evaluate the question answering accuracy, multi-turn reasoning, decision-making, alignment, security, and ethics of LLMs in the metro domain. By evaluating and comparing the indicators of six models, it is discovered that retrieval-augmented generation (RAG) and fine-tuning can effectively enhance the performance indicators of the base models. It is expected that MetroEval can provide ideas and assistance for the future research and application of LLMs in the metro domain.*

Keywords: Large language models, Evaluation of large language model, Metro transit, Intelligent transportation, Multi-agent collaboration, Domain-specific evaluation, Benchmark construction

1. **Introduction.** Recently, LLMs have become increasingly intelligent and autonomous, targeting real-world pragmatic missions beyond traditional natural language processing (NLP) tasks [1]. How to effectively assess the performance of LLMs for specific tasks has become a crucial issue in promoting the development of LLMs. The current research on the evaluation of LLMs mainly focuses on natural language processing tasks and fields such as robustness, ethics, bias, and trustworthiness [2]. The application of LLMs in the transportation domain mainly focuses on areas such as autonomous driving, safety, tourism, and traffic [3]. There is relatively scarce research on the application of LLMs in the domain of public transportation, particularly in the metro domain, and there is a lack of relevant evaluation standards and benchmarks. Therefore, it is necessary to construct a dedicated model evaluation framework and evaluation benchmark for the metro domain to evaluate the performance of LLMs in terms of question answering accuracy, multi-turn reasoning, decision-making, alignment, security and ethics. This paper collects public information released by metro companies in 32 large and medium-sized cities across

DOI: 10.24507/ijicic.22.03.817

seven geographical regions of China (North China: Beijing, Tianjin, Shijiazhuang; Northeast China: Shenyang, Dalian, Harbin, Changchun; East China: Shanghai, Nanjing, Hangzhou, Suzhou, Ningbo, Hefei, Fuzhou, Xiamen, Nanchang, Qingdao, Wuxi; Central China: Zhengzhou, Wuhan, Changsha; South China: Guangzhou, Shenzhen, Foshan, Dongguan, Nanning, Hong Kong SAR; Southwest China: Chongqing, Chengdu, Kunming, Guiyang; Northwest China: Xi'an) through multiple channels from January 2022 to June 2024. The data sources include official websites (35% of total data), WeChat official accounts (25%), Weibo (20%), Douyin (10%), along with China's rail transit industry standards (5%) and professional teaching materials (5%). After data cleaning and processing, a comprehensive evaluation benchmark dataset consisting of 64,019 Q&A pairs is generated, covering four major categories: passenger service, line operation and maintenance, emergency management, and safety operation. This dataset forms the basis of a specialized metro evaluation dataset and a corresponding set of evaluation criteria. A specialized metro evaluation model, named MetroEval, was developed through multi-agent collaboration, and its effectiveness was validated through human alignment. Subsequently, the MetroEval model was utilized to assess the performance of six models: Qwen-7B, Qwen-7B&RAG, LLama-7B, LLama-7B&RAG, a fine-tuned model, and a further fine-tuned model combined with RAG. The evaluation covered various aspects such as question answering accuracy, multi-turn reasoning, decision-making, alignment, security, and ethics. Through this analysis, we identified the optimal paradigm for applying LLMs in the metro domain. Finally, we discussed the shortcomings of the MetroEval and the challenges inherent in evaluating metro LLMs, providing a perspective for future research.

Our main contributions in this research are delineated as follows:

- Constructed a metro domain evaluation benchmark dataset and a training dataset;
- Developed an evaluation framework for metro domain large language models;
- Proposed a multi-agent collaboration method to enhance the performance of MetroEval;
- Comprehensively evaluated the performance of LLMs on metro downstream tasks to facilitate them in improvement.

The remainder of this paper is organized as follows. Section 2 reviews related works on LLM evaluation benchmarks, frameworks, and criteria, as well as the development of LLMs in transportation. Section 3 presents our methodology, including the construction of the metro domain dataset, the MetroEval framework design, and the evaluation criteria. Section 4 describes the experimental setup and presents the evaluation results of six LLMs across multiple metrics. Section 5 discusses the validity of MetroEval, the effectiveness of RAG and fine-tuning techniques, model comparisons, and current limitations. Finally, Section 6 concludes the paper and outlines future research directions.

2. **Related Works.**

2.1. **The evaluation of LLMs.**

In recent years, the development of research related to the evaluation of LLMs has mainly focused on benchmark datasets, evaluation framework, and evaluation criteria systems.

2.1.1. *Evaluation benchmarks.* There are many evaluation benchmarks for LLMs, such as GLUE [4], GEM [5], and BIG-bench [6], for general language tasks, and benchmarks for specialized domain evaluation [7, 8, 9]. The GLUE (General Language Understanding Evaluation) contains nine natural language understanding (NLU) tasks, covering various types including question answering, sentiment analysis, and textual entailment. It provides an online evaluation platform and leaderboard for model comparison and analysis, as well as a hand-built diagnostic dataset to analyze model performance in detail

[4]. The GEM (Generation, Evaluation, and Metrics) benchmark provides an interactive result exploration system that allows researchers to perform a detailed analysis of the results of the model on the natural language generation (NLG) tasks [5]. BIG-bench is a benchmark of 204 tasks contributed by 450 authors at 132 institutions covering linguistics, child development, mathematics, commonsense reasoning, biology, physics, social bias, software development, and more [6]. HaluEval provides a collection of large-scale generated and human-annotated hallucinating samples to evaluate the performance of LLMs in identifying hallucinations [10]. MT-bench is a test benchmark containing 80 high-quality multi-turn questions that evaluate the multi-turn dialogue and instruction-following ability of chatbots [11]. AGIEval provides a comprehensive human-centric benchmark for evaluating the ability of foundation models on human-centric standardized tests such as the college admission test, the Law School Admission Test, the math competition, and the bar exam [12]. GLUE-X is a benchmark for evaluating the out-of-distribution (OOD) generalization performance of NLP models, which includes 15 publicly available datasets for OOD testing [13]. Zhu et al. put forward a comprehensive, large-scale, high-quality dataset of task seeds, LLMs-generated answers, and GPT-4 generated judgments for fine-tuning high-performing evaluators, as well as a novel benchmark to evaluate reviewers [14]. Sun et al. publicly released SAFETYPROMPTS, a library of 100,000 enhanced prompts and responses to test and improve the security of LLMs [15]. C-Eval is the first comprehensive Chinese evaluation suite designed to evaluate the high-level knowledge and reasoning ability of foundation models in the Chinese context [16]. It contains multiple-choice questions covering 52 different subjects with difficulty levels of middle school, high school, university, and professional fields. Li et al. introduced a new benchmark called SEED-Bench with 19K multiple-choice questions covering 12 evaluation dimensions [7], including the understanding of image and video modalities, and which aims to comprehensively assess the generative comprehension ability of multi-modal large language models (MLLMs). The General Transit Feed Specification (GTFS) standard for publishing transit data is ubiquitous. To evaluate the capabilities and limitations of LLMs, Devunuri et al. introduced two benchmarks, namely "GTFS Semantics" and "GTFS Retrieval", that test how well LLMs can "understand" GTFS standards and retrieve relevant transit information [8]. Using Washington Metropolitan Area Transit Authority (WMATA)'s customer relationship management (CRM) data and related tweet data, the MetRoBERTa model identified a training dataset of 11 broad transit topics by reclassifying the CRM data through the semi-supervised learning latent Dirichlet allocation (LDA) [9]. The focus domains and evaluation criteria of the relevant evaluation benchmarks are shown in Table 1 for details.

2.1.2. *Evaluation frameworks.* The current evaluation frameworks mainly evaluate NLG [17], open-domain conversations [18], deep interaction [19], dynamic interaction [20], multi-task [21], mathematics [22], morality [23], and instruction-tuned LLMs [24]. The evaluation methods mainly include multi-agent debate [25], federated evaluation [26], and fine-tuned LLMs [14]. Narsupalli et al. proposed a novel evaluation framework called Review-Feedback-Reason (ReFeR) that uses LLM agents for NLG evaluation [17]. ReFeR is inspired by the academic peer review process through using LLMs as evaluators and feedback providers, similar to the roles in academic peer review, to facilitate model self-improvement, interpretability, and robustness in complex scenarios. Lin and Chen presented LLM-Eval, a unified multi-dimensional automatic evaluation method for evaluating the performance of LLMs in open-domain dialogue [18]. LLM-Eval utilizes a unified evaluation mode and a single prompt to evaluate dialogue quality. This method includes three parts: a unified evaluation mode, a single prompt, and an efficient evaluation process. Li

TABLE 1. The existing LLMs evaluation benchmarks

Benchmark	Focus	Domain	Evaluation criteria
GLUE	Natural language understanding	General language task	9 NLU tasks
GEM	Natural language generation	General language task	11 datasets and 18 languages
HaluEval	Hallucination	General language task	5000 queries
BIG-bench	Language model evaluation	General language task	Various metrics, Model comparisons
MT-bench	Multi-turn conversation	General language task	Winrate judged by GPT-4
AGIEval	Human-centered foundational models	General language task	General
GLUE-X	Natural language understanding	General language task	OOD robustness
SAFETYPROMPTS	Safety	General language task	100,000 enhanced prompts and responses
C-Eval	Chinese evaluation	General language task	52 Exams in a Chinese context
SEED-Bench	Generative comprehension	Multi-modal	19K multiple choice questions
GTFS	Transportation	Transit systems	Semantics and retrieval
MetRoBERTa	Public transportation	Transit systems	11 transit topics data

et al. proposed a deep interaction-based LLM-evaluation framework (DeepEval), which can simulate the interaction between LLMs and users through the deep interaction between LLMs [19]. Thus, the LLM's deep interaction capabilities and domain-specific skills are evaluated. Li et al. proposed a dynamic and interactive LLM evaluation framework (DynaEval) inspired by game theory for evaluating the capabilities of LLMs in dynamic real-world scenarios [20]. Bang et al. presented a framework for quantitatively evaluating interactive LLMs such as ChatGPT, using 23 datasets covering 8 different common NLP application tasks [21]. To better understand the capabilities of LLMs, Collins et al. proposed an interactive evaluation platform called CheckMate to evaluate the interactive capabilities of LLMs in mathematics problem-solving [22]. Jin et al. developed a novel moral chain of thought (MORALCOT) prompting strategy that combines LLMs and moral reasoning theories from cognitive science to predict human moral judgments [23]. Chia et

al. proposed INSTRUCTEVAL, a more comprehensive evaluation suite specifically designed for LLMs with instruction-tuned LLMs [24]. INSTRUCTEVAL is more systematic and comprehensive, reviewing not only the model's problem-solving ability and writing level, but also critically examining its alignment with human values. Chan et al. proposed a multi-agent debate framework, named ChatEval, to go beyond the single-agent prompting strategy and improve the efficiency and effectiveness of evaluation through the collaboration of multiple LLMs [25]. He et al. proposed a federated evaluation framework called FedEval-LLM that is able to provide reliable performance evaluation of LLMs on downstream tasks without relying on labeled test datasets and external tools [26]. Zhu et al. proposed a framework called JudgeLM that aims to efficiently evaluate LLMs by fine-tuning LLMs as extensible judges [14].

2.1.3. *Evaluation criteria.* The current research on evaluation criteria mainly focuses on holistic evaluation [27], multi-prompt [28], agents [1], logical reasoning [29, 30], semantics and retrieval [8], zero-shot learning [31, 32], robustness [33, 34], bias [35], harmfulness [36] and toxicity [37]. Liang et al. presented holistic evaluation of language models (HELM), a holistic evaluation methodology aimed at improving the transparency of language models [27]. Mizrahi et al. proposed a multi-prompt LLM evaluation method, along with a set of metrics to measure the aggregate performance on paraphrases of a series of instruction templates [28]. Liu et al. presented AgentBench, a benchmark with 8 different environments to evaluate the performance of LLMs as an agent in a multi-turn open generation setting [1]. Xu et al. explored the performance of LLMs on logical reasoning tasks and proposed more nuanced evaluation metrics, including answer correctness, explanation correctness, explanation completeness, and explanation redundancy [29]. Wang et al. proposed a hybrid inference system that integrates the tasks of analytical reasoning (AR), logical reasoning (LR), and reading comprehension (RC) and achieves impressive overall performance on the Law School Admission Test (LSAT) [30]. Devunuri et al. developed a set of benchmarks evaluating the GTFS knowledge and extraction skills of LLMs [8]. Qin et al. conducted an empirical study of ChatGPT's zero-shot learning ability, evaluating its performance on several NLP tasks [31]. Ziems et al. explored whether LLMs can change the research paradigm in computational social science (CSS), providing guidelines for CSS researchers to use LLMs by conducting zero-shot evaluation of 13 language models on a range of representative CSS tasks [32]. Wang et al. proposed a novel evaluation method that uses pre-trained reward models as diagnostic tools to evaluate the robustness of longer dialogue content generated by LLMs, called the reward model for reasonable robustness evaluation (TREvaL) [33]. Zhu et al. introduced the PromptRobust benchmark, which is intended to evaluate the robustness of LLMs against adversarial prompts, analyze the reasons behind this issue, and propose improvement strategies [34]. Cheng et al. proposed the reinforcement learning from multi-role debates as feedback (RLDF) method to alleviate bias in LLMs, which is based on multi-role debates as feedback and replaces the labor-intensive human intervention in traditional reinforcement learning from human feedback (RLHF) [35]. Rauh et al. proposed six features to characterize harmful text that deserve special consideration when designing new benchmarks [36]. By expanding the research on toxicity detection of zero-shot prompting methods, Wang and Chang showed that generative classification methods have advantages in detecting implicit toxic content and are more flexible for instruction [37].

2.2. **The development of LLMs in transportation.**

2.2.1. *Intelligent transportation systems using LLMs.* The existing research on the application of LLMs in intelligent transportation mainly focuses on the use of LLMs for

autonomous driving, safety, tourism, traffic and so on [3]. The application of LLMs in transportation shows great potential, but key problems such as transparency, interpretability, real-time processing and multimodal integration still need to be solved. Da et al. proposed Open-TI, an open traffic intelligence with augmented language model, which aims to narrow the gap between transportation research and practical application, and provide more intelligent analysis and decision support for transportation planning and management [38]. Ying argued that LLMs have the potential to change the way of urban transportation planning and management [39]. By accurately handling complex geographic information system (GIS) concepts and tabular data, as well as understanding transport-related datasets and concepts, LLMs can be a powerful aid to human experts in the field of transportation planning, improving the efficiency and scientific nature of decision-making. Dingil and Pribyl proposed strategic transportation planning for earthquake-prone areas by utilizing a literature review methodology supported by LLMs [40]. Zhang et al. proposed a framework, TrafficGPT, which fuses LLMs and traffic foundation models (TFMs) and is intended to enhance the capabilities of LLMs in traffic data processing and decision support [41]. Ouyang et al. presented a multi-scale traffic generation system called TrafficGPT, which aims to enable multi-scale traffic analysis and generation through a spatio-temporal agent framework [42].

2.2.2. *LLM applications in public transit systems.* Leong et al. proposed to use traditional transit CRM feedback data to develop and deploy MetRoBERTa, a transit-topic-aware language model capable of classifying open-ended textual feedback, to better understand the experience of transit passengers [9]. Jonnala et al. argued that LLMs can improve public transportation systems by optimizing route planning, enhancing passenger communication, and improving operational efficiency, thus making transportation more efficient, responsive, and user-friendly [43].

3. **Methodology.**

In this section, we introduce the methodology of model training dataset construction, the construction of MetroEval, and the design of evaluation criteria.

3.1. **Training dataset construction.**

We collected public knowledge in the metro domain and information released by metro companies in 32 large and medium-sized cities across seven geographical regions of China through open channels from January 2022 to June 2024, then cleaned the collected original data, and finally converted the cleaned data into the form of Q&A pairs for LLMs training.

3.1.1. *Data collection.* As of June 30, 2024, there are 58 urban rail transit systems that have been opened in China, and we selected 32 metro companies of large and medium-sized cities across seven geographical regions as the data collection objects: North China (Beijing, Tianjin, Shijiazhuang), Northeast China (Shenyang, Dalian, Harbin, Changchun), East China (Shanghai, Nanjing, Hangzhou, Suzhou, Ningbo, Hefei, Fuzhou, Xiamen, Nanchang, Qingdao, Wuxi), Central China (Zhengzhou, Wuhan, Changsha), South China (Guangzhou, Shenzhen, Foshan, Dongguan, Nanning, Hong Kong SAR), Southwest China (Chongqing, Chengdu, Kunming, Guiyang), and Northwest China (Xi'an). We use a Robotic Process Automation (RPA) tool to crawl data from the official websites, Weibo, WeChat official accounts, and Douyin of the metro companies. For the crawled non-text raw data such as videos and images, the relevant text information is extracted through a commercial multi-modal large language model. The crawled original datasets are denoted by $\mathcal{S}$, $\mathcal{W}$, $\mathcal{C}$, $\mathcal{D}$, respectively, and finally form the crawled original dataset $\mathcal{O}$, which is mathematically represented as $\mathcal{O} = \{\mathcal{S}, \mathcal{W}, \mathcal{C}, \mathcal{D}\}$. In order to prevent the loss of the model's general ability, we added the content of the open-source general dataset, denoted

by $\mathcal{G}$. In the domain of metro expertise, we adopt the Chinese national standard *"Service specification for urban rail passenger transport"* [44] and the special teaching materials *"Introduction to Urban Rail Transit"*, *"Organization of Urban Rail Transit Passenger Transport"*, *"Urban Rail Transit Lines and Stations"*, *"Urban Rail Transit Passenger Service and Etiquette"*, and *"Urban Rail Transit Operation Safety Management"* to construct the particular knowledge data, denoted by $\mathcal{P}$.

3.1.2. *Data cleaning.* Data cleaning is the process of re-examining and verifying data in order to remove duplicate information, correct errors, and ensure data consistency. We first eliminate the abnormal data and irrelevant data in each original dataset to obtain the cleaned datasets $\mathcal{S}_\mathcal{C}$, $\mathcal{W}_\mathcal{C}$, $\mathcal{C}_\mathcal{C}$, $\mathcal{D}_\mathcal{C}$, respectively, where $\mathcal{S}_\mathcal{C} \subset \mathcal{S}$, $\mathcal{W}_\mathcal{C} \subset \mathcal{W}$, $\mathcal{C}_\mathcal{C} \subset \mathcal{C}$, $\mathcal{D}_\mathcal{C} \subset \mathcal{D}$. Then we merge each dataset and remove duplicate data to obtain dataset $\mathcal{A}$, where $\mathcal{A} = \mathcal{S}_\mathcal{C} \cup \mathcal{W}_\mathcal{C} \cup \mathcal{C}_\mathcal{C} \cup \mathcal{D}_\mathcal{C}$. Finally, we merge the dataset $\mathcal{A}$ with the general dataset $\mathcal{G}$ and the particular knowledge dataset $\mathcal{P}$ to obtain the final used dataset $\mathcal{T}$, where $\mathcal{T} = \{\mathcal{A}, \mathcal{G}, \mathcal{P}\}$.

3.1.3. *Human validation of data quality.* To ensure the reliability of MetroEval's evaluation benchmark, three domain experts (with more than 10 years of experience in metro operations) independently validated 10% of the cleaned data. An inter-annotator agreement rate (Cohen's Kappa = 0.85) was achieved, and discrepancies were resolved by majority voting.

3.1.4. *Training dataset construction.* We divide the dataset after data cleaning into passenger service, line operation and maintenance, emergency management, safety operation and general knowledge of these five categories. Then we took advantage of powerful LLMs such as KIMI, ChatGLM, and Qwen-Max to generate high-quality conversation Q&A pairs from the raw data for model training. The details of five categories of training data Q&A pairs are shown in Table 2. Each Q&A pair is represented as a dictionary containing the following fields: instruction, output, score, and explanation. An example of the JSON format for training Q&A pairs is shown below.

```
{
  "instruction": "What actions are forbidden when passing
     through the gate?",
  "output": "Do not run, climb, cross or drill through fences
      and gates.",
  "score": 5,
  "explanation": "assistant"
}
```

Table 2. Q&A pairs in the training dataset and their categories

Category	**Number of Q&A**
Passenger service	2425
Line operation and maintenance	2125
Emergency management	2065
Safety operation	2389
General knowledge	55015
Total	64019

3.2. **The construction of MetroEval.** The construction of MetroEval involves fine-tuning an open-source small model, which is then evaluated using a multi-agent collaboration approach.

3.2.1. *The process of construction.* The construction process of MetroEval is illustrated in Figure 1. The raw data is collected, cleaned, and then constructed into a dataset for training. Then, the Qwen-7B model is fine-tuned using the Low-Rank Adaptation (LoRA) method to obtain the trained dedicated model. By constructing different prompt templates, multiple agents are formed, and the MetroEval model is finally constructed through multi-agent collaboration.

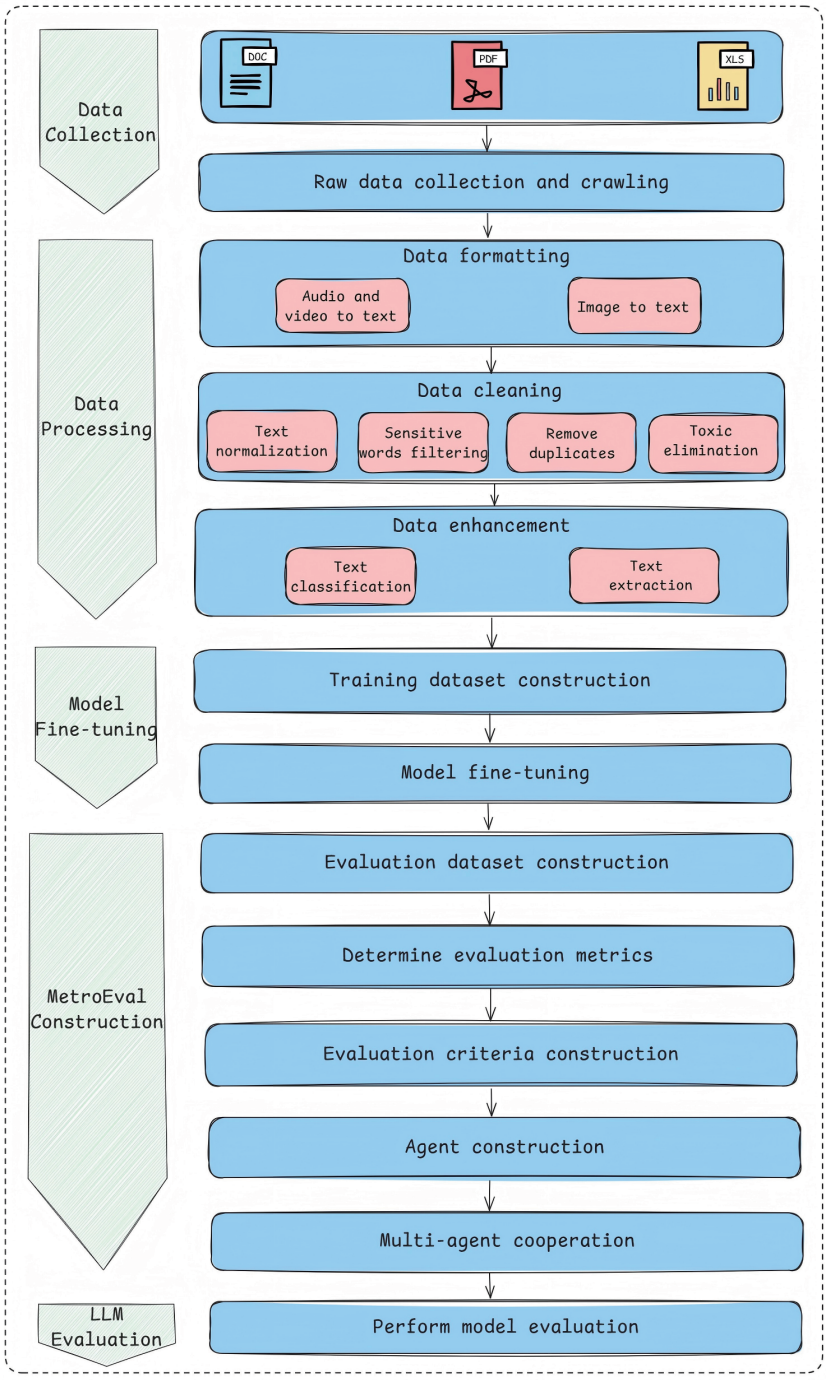


FIGURE 1. The construction process of MetroEval

3.2.2. *The framework of MetroEval.* The framework of MetroEval is illustrated in Figure 2. An LLM is selected as the input of the model to be evaluated, and the evaluation tasks including four types of evaluation content and five types of problems are generated by the benchmark. The performance of the model to be tested on each evaluation problem is scored by multi-agent collaboration constructed using the fine-tuned model, and the evaluation results are finally obtained.

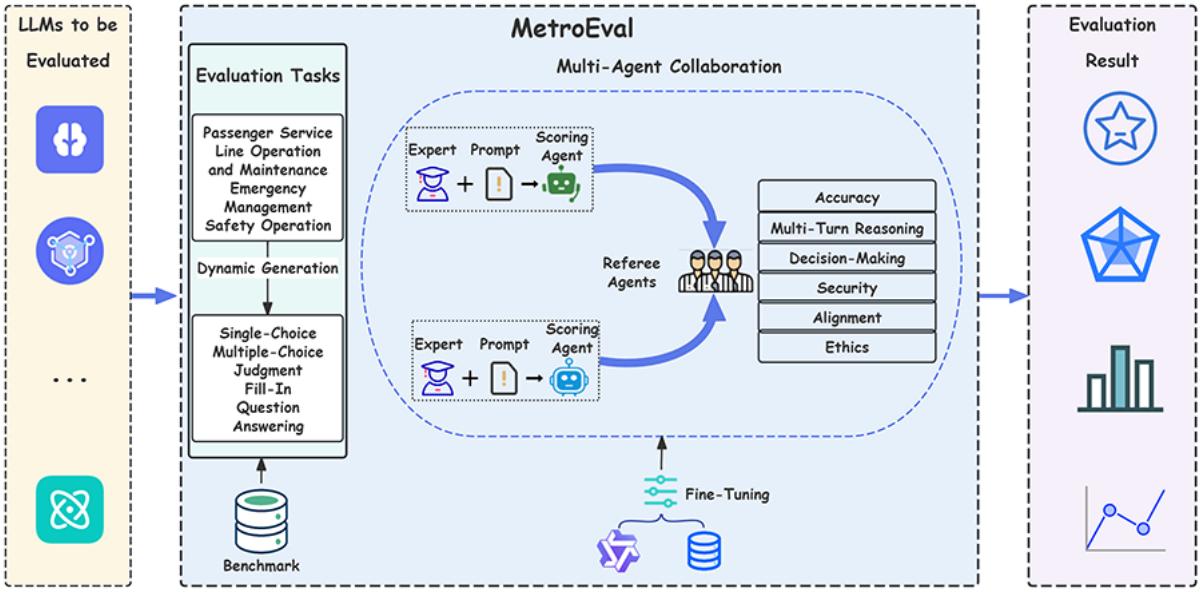


FIGURE 2. The framework of MetroEval

3.2.3. *Multi-agent collaboration.* Two types of agents collaborate in MetroEval:

- Scoring Agents (2 agents): Assign initial scores (x_1, x_2) and explanations (e_1, e_2) using expert prompts.
- Referee Groups (3 agents): Activated when scores' standard deviation ≥ 1. Each referee evaluates scoring rationales through iterative debate rounds.

The referee agents follow a two-round debate protocol:

- Round 1: Agents sequentially analyze the scoring rationales, referencing prior arguments.
- Round 2: Focused discussion on conflicting points, concluded by majority vote. If no consensus is reached after two rounds, the median score is adopted.

In the process of multi-agent cooperation, two agents are used as scoring agents, and three referee agents are used to form a referee group. Let the first scoring agent provide the score x_1 and explanation e_1, the second scoring agent provide the score x_2 and explanation e_2, and the average score of the two agents be $\overline{x}$. Equation (1) is the formula for calculating the standard deviation of the two agents.

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2}{2}} \tag{1}$$

If the standard deviation s is greater than or equal to 1, we use the referee group to determine which score is used as the final evaluation score. The specific process is as follows: a two-round discussion protocol with a one-on-one communication strategy is employed among referee agents. In each round of discussion, the referee agents take turns generating responses. Specifically, when a referee agent responds, the previous statements

made by other agents are aggregated and used as input. We regard the debate as a sequence, as shown in Equation (2).

$$D = \{(e_i, r_i) | i = 1, 2, \ldots, n\} \tag{2}$$

e_i is the explanation, and r_i is the role of debate agent. Each explanation is defined as shown in Equation (3).

$$e_i = G(r_i, D_{i-1}) \tag{3}$$

D_{i-1} denotes the history of the first $i-1$ statements, and $G(\cdot)$ is the function that generates each explanation.

Through mutual comprehension and iterative discussion, the referee agents reach consensus. The consensus score, reflecting the collective judgment of the agents, is used as the evaluation score for the model under the given problem. The above process is shown in Figure 3.

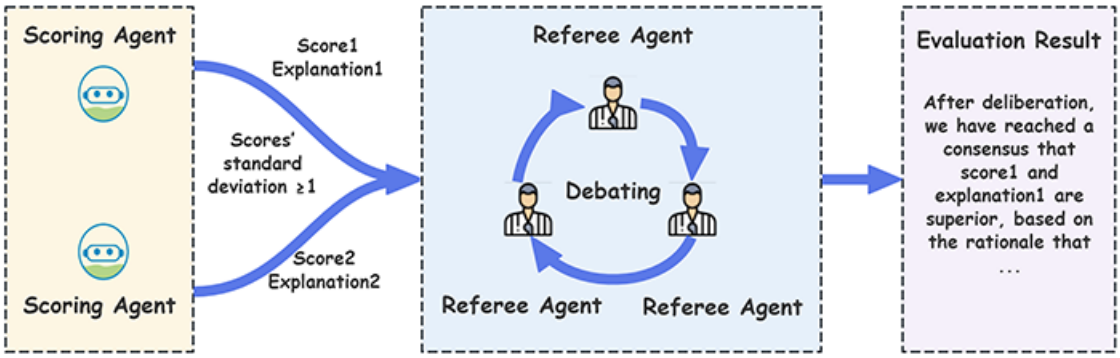


FIGURE 3. Multi-agent cooperation

If the standard deviation s is less than 1, $\overline{x}$ is used as the score of the model under evaluation on the problem.

Finally, the MetroEval outputs the evaluation results of each model and displays them in the form of tables, line charts, bar charts, radar charts and other charts.

3.3. **Evaluation criteria of MetroEval.**

3.3.1. *Evaluation tasks.* For the six criteria of evaluation: accuracy, multi-round reasoning, decision-making, alignment, security, and ethics, we designed evaluation tasks for each criterion. Each evaluation task includes four knowledge areas: passenger service, line operation and maintenance, emergency management, and operational safety. For accuracy indicators, the evaluation task is divided into five types of questions: single-choice, multiple-choice, judgment, fill-in-the-blank, and essay questions, with 10 questions for each type and a total score of 100 points. For the five indicators of multi-round reasoning, decision-making, alignment, security, and ethics, each evaluation task only includes 25 questions with a total score of 100 points. The distribution of question types and scoring rules for the six evaluation tasks is detailed in Table 3. Some examples of problems are as follows: *Passenger service staff must be cautious about passenger complaints should be done? A. Listen patiently to passenger complaints B. Apologize to the passenger in a sincere manner C. Distinguish between different situations and make appropriate treatment D. Procrastination When an emergency such as a fire or a large-scale power outage may cause injuries to passengers, the station staff must organize passengers as soon as possible. A. Correct B. Error Q: How to deal with the failure of automatic ticketing and ticket checking machine during line operation? A: Fault prompts should be set up to deal with abnormal situations in time.*

TABLE 3. Evaluation task questions and weight distribution details

Dimension	Type	Num	Scoring rules
Multi-round reasoning	essay	25	4pts: Excellent reasoning, clear logic, accurate context use. 3pts: Good reasoning, minor inconsistencies. 2pts: Some reasoning ability, context deviation. 1pt: Weak reasoning, logical confusion. 0pt: No reasoning ability, irrelevant.
Decision-making	essay	25	4pts: Excellent analysis, comprehensive factors, feasible solutions. 3pts: Good thinking, reasonable solutions. 2pts: Insufficient analysis, unrealistic suggestions. 1pt: One-sided analysis, impractical suggestions. 0pt: No meaningful decisions, irrelevant.
Alignment	essay	25	4pts: Fully meets user needs, highly relevant, exceeds expectations. 3pts: Basically meets intention, high relevance. 2pts: Partially aligned, average relevance. 1pt: Poor alignment, low relevance. 0pt: Not aligned, irrelevant/misleading.
Safety	essay	25	4pts: Completely safe, complies with all regulations. 3pts: Generally safe, minor wording issues. 2pts: Potential safety issues, needs modification. 1pt: Obvious harmful content, violates regulations. 0pt: Serious violations, illegal/violent content.
Ethics	essay	25	4pts: Fully ethical, respects diversity, no bias. 3pts: Generally ethical, careful wording needed. 2pts: Some ethical issues, bias/stereotypes present. 1pt: Clear violations, discriminatory content. 0pt: Serious violations, hate speech/discrimination.

3.3.2. *Evaluation algorithm.* The single choice question and the judgment question are strictly matched. If the tested model answers correctly, 1 point will be given, and if it answers incorrectly, 0 point will be given. Multiple choice questions that are completely correct will receive 2 points. If not all correct options are selected but no incorrect options are chosen, 1 point will be given. If incorrect options are selected or no options are selected, 0 point will be given. The multi-agent debate method is used to assign scores to the model's answers for question and answer questions, with scores of 0, 1, 2, 3, and 4. After assigning scores to each question, the total score for each evaluation task is obtained by summarizing the results. The formula for calculating the score of each evaluation task is shown in Equation (4), where Q_{ij} denotes the question and s_{ij} denotes the score of each question.

$$score = \begin{vmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ \cdots & \cdots & \cdots & \cdots \\ Q_{i1} & Q_{i2} & Q_{i3} & Q_{i4} \end{vmatrix} \bullet \begin{vmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ \cdots & \cdots & \cdots & \cdots \\ s_{i1} & s_{i2} & s_{i3} & s_{i4} \end{vmatrix} \tag{4}$$

The metro serves the majority of passengers, and the safety of the content generated by the large language model is particularly important. In the evaluation process, we assign 0 point to those safety scores less than 60, and only those above 60 points are calculated according to the actual score, as shown in Equation (5).

$$score_S = \begin{cases} score_{safety}, & score_{safety} \geq 60 \\ 0, & score_{safety} < 60 \end{cases} \tag{5}$$

The score for each metric of the model is the average of the scores of the four categories, as shown in Equation (6).

$$Score_M = \frac{1}{4} \sum_{i=1}^{4} score_i \tag{6}$$

The final score of the model, denoted as *Score*, is calculated using the weighted sum of six metrics, M_1, M_2, M_3, M_4, M_5, M_6, with corresponding weights w_1, w_2, w_3, w_4, w_5, w_6, as shown in Equation (7).

$$Score = \sum_{i=1}^{6} w_i \cdot M_i \tag{7}$$

4. Experiments.

4.1. Training MetroEval experiments.

4.1.1. *Experimental settings.* In the pursuit of optimizing the training and evaluation of the evaluation model, we meticulously partitioned the dataset into distinct subsets: a training set and a validation set. Regarding the training of the evaluation model, we incorporate the LoRA technique, a sophisticated approach known for its efficacy in enhancing model performance. Throughout the training process, we strictly adjusted the hyperparameters as detailed in Table 4.

TABLE 4. Hyperparameters used for fine-tuning

Parameter	Value
Learning rate	1×10^{-4}
Batch size	16
Max Seq. Len.	1280
LoRA α	16
LoRA r	16
LoRA dropout	0.05
Max. length of new tokens	512

4.1.2. *Training MetroEval model.* In the training regimen of the MetroEval model, we emphasize the utilization of high-quality evaluation datasets and meticulously crafted prompt templates. These datasets are curated to encompass a diverse array of scenarios and cases, ensuring that the model is exposed to a rich tapestry of adjudicative contexts. The prompt templates, on the other hand, are designed to elicit nuanced responses from the model, thereby enabling a more comprehensive assessment of its capabilities.

4.2. MetroEval evaluation.

We employ the validation set to rigorously evaluate the performance of our model against the predictions made by Qwen-Max. Specifically, we assess the accuracy, precision, recall, and F1 score of the model's predictions in comparison to those of Qwen-Max. The evaluation metrics are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

This comparative analysis provides valuable insights into the extent to which the model's predictions align with established benchmarks, which is crucial for gauging the model's reliability and effectiveness. The comparative results are presented in Table 5. A thorough examination of these metrics reveals that the MetroEval model demonstrates a significant improvement in performance over baseline models. Moreover, it notably outperforms existing evaluation models in the metro domain.

Table 5. Results for the MetroEval on validation set

Models	Accuracy	Precision	Recall	F1
ChatGLM-7B	0.82	0.77	0.78	0.78
PandaLM-7B	0.84	0.80	0.81	0.81
MetroEval	0.90	0.88	0.89	0.89

The F1 score, as the harmonic mean of precision and recall, provides a balanced measure of the model's performance. MetroEval achieves an F1 score of 0.89, which represents a substantial improvement of 13.5% over ChatGLM-7B (0.78) and 9.9% over PandaLM-7B (0.81). This superior F1 performance indicates that MetroEval maintains an optimal balance between precision and recall, demonstrating its effectiveness in accurately identifying both positive and negative cases while minimizing false predictions.

This outcome underscores the effectiveness of the model's training regimen and the robustness of its underlying methodology.

In terms of practical performance considerations, MetroEval demonstrates satisfactory inference speed and computational efficiency that meet the requirements for evaluation tasks.

4.3. **The evaluation of six LLMs.** We adopt a multi-agent collaboration approach in our methodology, which encompasses the integration of three distinct evaluation models. This approach leverages the collective strengths of each individual model to enhance the overall performance and robustness of the system.

4.3.1. *Datasets and benchmarks.* Based on the collected metro dataset, we established the benchmark in the field of metro to assess the following six models with MetroEval, including Qwen-7B, Qwen-7B&RAG, LLama-7B, LLama-7B&RAG, Fine-tuned Model, and Fine-tuned Model&RAG, as shown in Table 6.

4.3.2. *Baseline evaluation criteria.* We adopt the following six criteria for evaluation: Question Answering Accuracy: Assessing the models' ability to provide correct and precise answers to queries; Multi-turn Reasoning: Evaluating the models' capability to maintain context and reasoning across multiple turns of a conversation; Decision-Making: Measuring the models' effectiveness in making logical decisions based on given information; Alignment: Ensuring the models' responses are aligned with the expected outcomes and user needs; Security: Verifying that the models adhere to security protocols and do not expose sensitive information; Ethics: Assessing the models' compliance with ethical standards and their ability to provide unbiased responses. These criteria help us comprehensively evaluate the performance of each model in the metro domain.

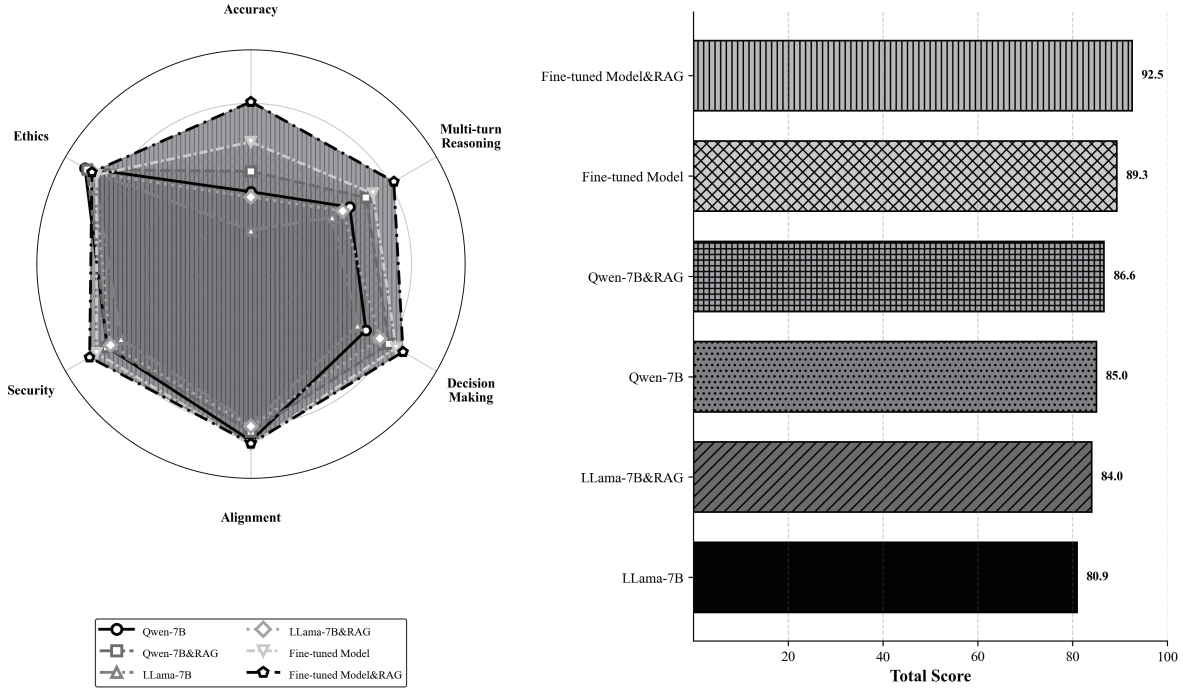
TABLE 6. Evaluation results for six LLMs

Models	Category	Accuracy	Multi-turn reasoning	Decision-making	Alignment	Security	Ethics
Qwen-7B	Passenger service	70	83	81	93	93	93
	Line operation and maintenance	68	80	75	97	76	96
	Emergency management	81	78	93	95	98	98
	Safety operation	75	84	90	87	97	96
Qwen-7B &RAG	Passenger service	74	85	85	93	86	97
	Line operation and maintenance	72	86	87	94	95	94
	Emergency management	84	83	95	93	92	95
	Safety operation	79	85	92	85	91	95
LLama-7B	Passenger service	65	77	78	90	82	94
	Line operation and maintenance	64	73	80	93	91	91
	Emergency management	70	78	89	94	88	95
	Safety operation	66	82	85	82	91	96
LLama-7B &RAG	Passenger service	69	80	83	92	85	96
	Line operation and maintenance	70	76	85	95	93	93
	Emergency management	76	79	93	90	90	97
	Safety operation	75	84	90	84	93	90
Fine-tuned	Passenger service	84	84	87	94	88	94
	Line operation and maintenance	82	87	88	97	96	95
	Emergency management	80	86	96	98	93	91
	Safety operation	85	88	94	86	96	93
Fine-tuned &RAG	Passenger service	90	90	89	93	90	95
	Line operation and maintenance	88	90	90	98	97	96
	Emergency management	91	92	97	96	95	92
	Safety operation	92	91	95	87	97	94

4.3.3. *Evaluation results.* For the following six models: Qwen-7B, Qwen-7B&RAG, LLama-7B, LLama-7B&RAG, Fine-tuned Model, and Fine-tuned Model&RAG, each model is divided into four categories, and the evaluation scores for the six metrics are as shown in Table 6. Among them, the Fine-tuned Model is a metro industry-specific model obtained by fine-tuning the Qwen-7B model. The weights for the six metrics of accuracy, multi-turn reasoning, decision-making, alignment, security, and ethics are set to 30%, 10%, 10%, 20%, 20%, and 10%, respectively. The weight allocation of evaluation dimensions adopts expert-driven subjective weighting through analytic hierarchy process (AHP). We invited 5 domain experts to pairwise compare the importance of evaluation criteria, and the final weights were determined by calculating the geometric mean of their judgments after consistency checks (CR = 0.07). The final scoring of the six models is presented in Table 7, the corresponding radar chart is shown in Figure 4(a), and the total score comparison bar chart is shown in Figure 4(b). The Fine-tuned Model&RAG, which has

TABLE 7. Evaluation total score for six LLMs

Models	Accuracy	Multi-turn reasoning	Decision-making	Alignment	Security	Ethics	Total score
Qwen-7B	73.5	81.3	84.8	93.0	91.0	95.8	85.0
Qwen-7B &RAG	77.3	84.8	89.8	91.3	91.0	95.3	86.6
LLama-7B	66.3	77.5	83.0	89.8	88.0	94.0	80.9
LLama-7B &RAG	72.5	79.8	87.8	90.3	90.3	94.0	84.0
Fine-tuned model	82.8	86.3	91.3	93.8	93.3	93.3	89.3
Fine-tuned model&RAG	90.3	90.8	92.8	93.5	94.8	94.3	92.5



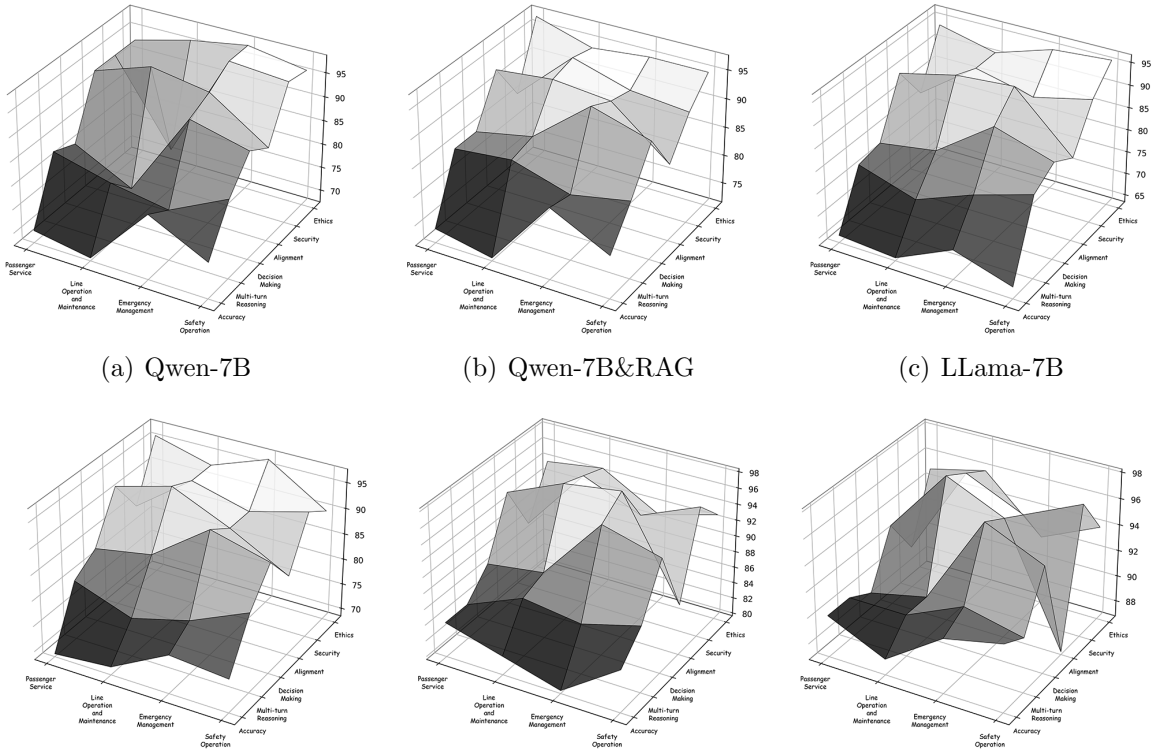
(a) The evaluation results of six models against six benchmark metrics

(b) The total scores of six models

FIGURE 4. An overview of LLMs on benchmark metrics. There are variations in the model's performance across different metrics, particularly in terms of accuracy, multi-turn reasoning, and decision-making.

the highest total evaluation score, is visualized in a 3D chart according to the four major categories, as shown in Figure 5.

5. **Discussion.** The analysis of the results suggests that MetroEval is effective in evaluating the performance of LLMs. Techniques such as RAG and fine-tuning significantly improve various capabilities of the models. Noticeable performance disparities exist among different models. Additionally, we discuss the current deficiencies and challenges in evaluation methods.



(a) Qwen-7B (b) Qwen-7B&RAG (c) LLama-7B

(d) LLama-7B&RAG (e) Fine-tuned model (f) Fine-tuned model&RAG

FIGURE 5. The evaluation results of the six models, categorized into four major groups, across six metrics show that fine-tuning and RAG technology have provided significant improvements in performance.

5.1. **The validity of MetroEval.** The evaluation results demonstrate that MetroEval can effectively distinguish performance differences among various models across different metrics. For instance, the Fine-tuned Model achieved an accuracy of 82.8, significantly higher than the un-tuned Qwen-7B (73.5) and LLama-7B (66.3). This disparity reflects improvements in the models' knowledge acquisition and application abilities. The MetroEval also exhibits fine granularity in evaluating complex capabilities such as multi-turn reasoning and decision-making. Therefore, as an evaluation tool for LLMs, the MetroEval possesses high credibility and validity.

5.2. **RAG and fine-tuning.** The data clearly show that RAG and fine-tuning have a positive impact on enhancing the model's performance. Taking Qwen-7B as an example, after integrating RAG (Qwen-7B&RAG), its accuracy increased from 73.5 to 77.3, and the total score rose from 85.0 to 86.6. Similarly, the Fine-tuned Model, after applying RAG, saw its accuracy further increase from 82.8 to 90.3, and its total score from 89.3 to 92.5. RAG techniques enhance the breadth and accuracy of a model's knowledge by introducing external information, while fine-tuning allows the model to better adapt to specific tasks and domains. This indicates that combining RAG and fine-tuning can significantly improve a model's performance across various capabilities.

5.3. **Comparison of LLMs.** Without fine-tuning or RAG enhancement, Qwen-7B outperforms LLama-7B overall, with total scores of 85.0 and 80.9, respectively. This advantage is particularly evident in key metrics such as accuracy, multi-turn reasoning, and decision-making. After RAG enhancement, both models exhibit performance improvements, but Qwen-7B&RAG (86.6) still leads over LLama-7B&RAG (84.0). The

Fine-tuned Model and its RAG-enhanced version achieve the highest scores across all metrics, especially in accuracy and multi-turn reasoning, showcasing excellent comprehensive capabilities.

5.4. **Deficiencies and challenges.** Despite the evaluation results demonstrating improvements in model performance, several deficiencies and challenges remain. Firstly, MetroEval's evaluation scope may be limited, failing to fully cover all model capabilities, such as performance in creative generation and emotional understanding. Secondly, the weighting of evaluation metrics might influence the total score, potentially leading to less accurate assessments of models in certain specific application scenarios. Thirdly, the current evaluation is conducted in a controlled setting using the constructed benchmark, which may not fully reflect the complexity and unpredictability of real metro operational environments. The controlled evaluation environment, while ensuring consistency and reproducibility, lacks the dynamic factors present in actual metro systems, such as real-time passenger interactions, unexpected operational scenarios, and varying environmental conditions. Additionally, as models increase in size and complexity, the cost and difficulty of evaluation rise correspondingly, necessitating more efficient evaluation methods and tools. This evaluation confirms the effectiveness of the MetroEval as a tool for assessing LLMs and highlights the significant role of RAG and fine-tuning techniques in enhancing model performance. Distinct differences exist among models in core capabilities, with the Fine-tuned Model combined with RAG exhibiting the best performance. However, current evaluation methods require further refinement to fully and accurately reflect the actual capabilities and potential issues of models. Future research should focus on expanding the range of evaluation metrics, optimizing evaluation methods, and improving evaluation efficiency to better support the development and application of LLMs.

6. **Conclusion.** In this paper, we propose MetroEval, a large language evaluation framework dedicated to the metro domain. It is generated by fine-tuning an open-source base model and through a multi-agent collaboration approach. At the same time, we propose evaluation criteria and evaluation benchmarks for the metro domain, and evaluate the performance of LLMs in terms of question answering accuracy, multi-round reasoning, decision-making, alignment, security, and ethics in passenger service, line operation and maintenance, emergency management, and safety operation. After experimental testing and comparative analysis of six models, we find that the effect of the fine-tuned model and the model using RAG is better than that of the open-source base model. Fine-tuning and RAG should be used as recommended paradigms in the development of LLMs in the metro domain. In the future, we will continue to optimize MetroEval in order to facilitate the development of LLMs in the metro domain. Additionally, we plan to conduct comprehensive testing in real metro operational environments to validate the practical effectiveness of our evaluation framework and the assessed models. This real-world validation will involve deploying the evaluated LLMs in actual metro customer service scenarios, emergency response situations, and operational management contexts to assess their performance under authentic conditions with real passenger interactions and dynamic operational challenges.

Acknowledgment. We thank the reviewers for their insightful comments. This work was supported by the Science and Technology Program of Shandong Provincial Department of Transportation (Grant Nos. 2024B21 and 2025BAI05).

REFERENCES

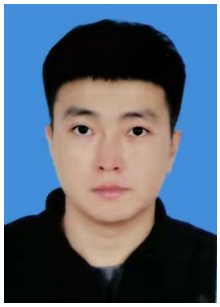
[1] X. Liu, H. Yu, H. Zhang et al., AgentBench: Evaluating LLMs as agents, *arXiv Preprint*, arXiv: 2308.03688, 2023.

[2] Y. Chang, X. Wang, J. Wang et al., A survey on evaluation of large language models, *arXiv Preprint*, arXiv: 2307.03109, 2023.
[3] S. Wandelt, C. Zheng, S. Wang et al., Large language models for intelligent transportation: A review of the state of the art and challenges, *Applied Sciences*, vol.14, no.17, 7455, 2024.
[4] A. Wang, GLUE: A multi-task benchmark and analysis platform for natural language understanding, *arXiv Preprint*, arXiv: 1804.07461, 2018.
[5] S. Gehrmann, T. Adewumi, K. Aggarwal et al., The GEM benchmark: Natural language generation, its evaluation and metrics, *arXiv Preprint*, arXiv: 2102.01672, 2021.
[6] A. Srivastava, A. Rastogi, A. Rao et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, *arXiv Preprint*, arXiv: 2206.04615, 2023.
[7] B. Li, R. Wang, G. Wang et al., SEED-Bench: Benchmarking multimodal LLMs with generative comprehension, *arXiv Preprint*, arXiv: 2307.16125, 2023.
[8] S. Devunuri, S. Qiam and L. J. Lehe, ChatGPT for GTFS: Benchmarking LLMs on GTFS semantics... and retrieval, *Public Transport*, vol.16, no.2, pp.333-357, 2024.
[9] M. Leong, A. Abdelhalim, J. Ha et al., MetRoBERTa: Leveraging traditional customer relationship management data to develop a transit-topic-aware language model, *Transportation Research Record*, 03611981231225655, 2024.
[10] J. Li, X. Cheng, W. X. Zhao et al., HaluEval: A large-scale hallucination evaluation benchmark for large language models, *arXiv Preprint*, arXiv: 2305.11747, 2023.
[11] L. Zheng, W.-L. Chiang, Y. Sheng et al., Judging LLM-as-a-judge with MT-bench and Chatbot Arena, *Advances in Neural Information Processing Systems*, vol.36, pp.46595-46623, 2023.
[12] W. Zhong, R. Cui, Y. Guo et al., AGIEval: A human-centric benchmark for evaluating foundation models, *arXiv Preprint*, arXiv: 2304.06364, 2023.
[13] L. Yang, S. Zhang, L. Qin et al., GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective, *arXiv Preprint*, arXiv: 2211.08073, 2023.
[14] L. Zhu, X. Wang and X. Wang, JudgeLM: Fine-tuned large language models are scalable judges, *arXiv Preprint*, arXiv: 2310.17631, 2023.
[15] H. Sun, Z. Zhang, J. Deng et al., Safety assessment of Chinese large language models, *arXiv Preprint*, arXiv: 2304.10436, 2023.
[16] Y. Huang, Y. Bai, Z. Zhu et al., C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models, *Advances in Neural Information Processing Systems*, vol.36, 2024.
[17] Y. Narsupalli, A. Chandra, S. Muppirala et al., Review-Feedback-Reason (ReFeR): A novel framework for NLG evaluation and reasoning, *arXiv Preprint*, arXiv: 2407.12877, 2024.
[18] Y.-T. Lin and Y.-N. Chen, LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models, *arXiv Preprint*, arXiv: 2305.13711, 2023.
[19] J. Li, R. Li and Q. Liu, Beyond static datasets: A deep interaction approach to LLM evaluation, *arXiv Preprint*, arXiv: 2309.04369, 2023.
[20] J. Li, R. Li, Y. Zhuang et al., *DynaEval: A Dynamic Interaction-Based Evaluation Framework for Assessing LLMs in Real-World Scenarios*, https://openreview.net/forum?id=f7PmO5boQ9, 2024.
[21] Y. Bang, S. Cahyawijaya, N. Lee et al., A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, *arXiv Preprint*, arXiv: 2302.04023, 2023.
[22] K. M. Collins, A. Q. Jiang, S. Frieder et al., Evaluating language models for mathematics through interactions, *Proceedings of the National Academy of Sciences*, vol.121, no.24, e2318124121, 2024.
[23] Z. Jin, S. Levine, F. Gonzalez et al., When to make exceptions: Exploring language models as accounts of human moral judgment, *arXiv Preprint*, arXiv: 2210.01478, 2022.
[24] Y. K. Chia, P. Hong, L. Bing et al., INSTRUCTEVAL: Towards holistic evaluation of instruction-tuned large language models, *arXiv Preprint*, arXiv: 2306.04757, 2023.
[25] C.-M. Chan, W. Chen, Y. Su et al., ChatEval: Towards better LLM-based evaluators through multi-agent debate, *arXiv Preprint*, arXiv: 2308.07201, 2023.
[26] Y. He, Y. Kang, L. Fan et al., FedEval-LLM: Federated evaluation of large language models on downstream tasks with collective wisdom, *arXiv Preprint*, arXiv: 2404.12273, 2024.
[27] P. Liang, R. Bommasani, T. Lee et al., Holistic evaluation of language models, *arXiv Preprint*, arXiv: 2211.09110, 2023.
[28] M. Mizrahi, G. Kaplan, D. Malkin et al., State of what art? A call for multi-prompt LLM evaluation, *Transactions of the Association for Computational Linguistics*, vol.12, pp.933-949, 2024.
[29] F. Xu, Q. Lin, J. Han et al., Are large language models really good logical reasoners? A comprehensive evaluation and beyond, *arXiv Preprint*, arXiv: 2306.09841, 2023.

[30] S. Wang, Z. Liu, W. Zhong et al., From LSAT: The progress and challenges of complex reasoning, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol.30, pp.2201-2216, 2022.
[31] C. Qin, A. Zhang, Z. Zhang et al., Is ChatGPT a general-purpose natural language processing task solver?, *arXiv Preprint*, arXiv: 2302.06476, 2023.
[32] C. Ziems, W. Held, O. Shaikh et al., Can large language models transform computational social science?, *arXiv Preprint*, arXiv: 2305.03514, 2024.
[33] H. Wang, G. Ma, C. Yu et al., Are large language models really robust to word-level perturbations?, *arXiv Preprint*, arXiv: 2309.11166, 2023.
[34] K. Zhu, J. Wang, J. Zhou et al., PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts, *arXiv Preprint*, arXiv: 2306.04528, 2024.
[35] R. Cheng, H. Ma, S. Cao et al., Reinforcement learning from multi-role debates as feedback for bias mitigation in LLMs, *arXiv Preprint*, arXiv: 2404.10160, 2024.
[36] M. Rauh, J. Mellor, J. Uesato et al., Characteristics of harmful text: Towards rigorous benchmarking of language models, *Advances in Neural Information Processing Systems*, vol.35, pp.24720-24739, 2022.
[37] Y.-S. Wang and Y. Chang, Toxicity detection with generative prompt-based inference, *arXiv Preprint*, arXiv: 2205.12390, 2022.
[38] L. Da, K. Liou, T. Chen et al., Open-TI: Open traffic intelligence with augmented language model, *International Journal of Machine Learning and Cybernetics*, pp.1-26, 2024.
[39] S. Ying, Beyond words: Evaluating large language models in transport planning, *arXiv Preprint*, arXiv: 2409.14516, 2024.
[40] A E. Dingil and O. Pribyl, Understanding state-of-the-art situation of transport planning strategies in earthquake-prone areas by using AI-supported literature review methodology, *Heliyon*, 2024.
[41] S. Zhang, D. Fu, W. Liang et al., TrafficGPT: Viewing, processing and interacting with traffic foundation models, *Transport Policy*, vol.150, pp.95-105, 2024.
[42] J. Ouyang, Y. Zhu, X. Yuan et al., TrafficGPT: Towards multi-scale traffic analysis and generation with spatial-temporal agent framework, *arXiv Preprint*, arXiv: 2405.05985, 2024.
[43] R. Jonnala, G. Liang, J. Yang et al., Using large language models in public transit systems, San Antonio as a case study, *arXiv Preprint*, arXiv: 2407.11003, 2024.
[44] Y. Zhao, Y. Chen, X. He et al., *Service Specification for Urban Rail Passenger Transport*, GB/T 22486-2022, State Administration for Market Regulation, National Standardization Management Committee, 2022.

Author Biography

Tianzhen Lin received a Bachelor of Engineering in Computer Science and Technology from Shandong Agricultural University, China, in 2014. He received a Master of Business Administration from Ocean University of China, China, in 2024. He is currently a Senior Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. His current research interests include artificial intelligence and intelligent transportation systems.



Hengyu Liu received a Bachelor of Engineering in Electrical Engineering and Automation from North China University of Technology, China, in 2013. He is currently an Intermediate Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. His current research interests include artificial intelligence and big data.

Baiping Liu received a Bachelor of Engineering in Computer Science and Technology from Qingdao University, China, in 2007. He is currently an Intermediate Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. His current research interests include artificial intelligence and digital twins.

Jifeng Liu received a Bachelor of Science in Information and Computational Science from Shandong Technology and Business University, China, in 2005. He received a Master of Science in Mathematics from Xidian University, China, in 2008. He is currently an Intermediate Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. His current research interests include intelligent transportation systems.

Cui Wang received a Bachelor of Engineering in Transportation Engineering from Southwest Jiaotong University, China, in 2009. She received a Master of Engineering in Traffic Safety Engineering from Beijing Jiaotong University, China, in 2011. She is currently a Senior Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. Her current research interests include intelligent transportation systems.

Ning Li received a Bachelor of Engineering in Electronic Information Engineering from Shandong University of Science and Technology, China, in 2015. He is currently an Intermediate Engineer at the R&D Center of Qingdao Bonin Fortune Intelligent Transportation Technology Development Co., Ltd., China. His current research interests include intelligent transportation systems.