VIRTUAL METEOROLOGIST'S EXAM: BENCHMARKING METEOROLOGICAL EXPERTISE IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the growing demand for virtual meteorologist assistants in operational scenarios has driven the incorporation of meteorological corpora into the training of large language models (LLMs). However, due to the scale and diversity, existing benchmarks fail to provide a comprehensive assessment of LLMs' understanding and reasoning capabilities in meteorology. To this end, we introduce a benchmark, namely Virtual Meteorologist's Exam (VME), to evaluate whether existing large language models have professional-level meteorological expertise equivalent to that of humans. To be concrete, our VME comprises nearly 6,500 question–answer pairs across four formats, making it four times the size of comparable datasets. It covers both academic knowledge and operational scenarios. Meanwhile, it is also organized into five fist-level categories and 25 fine-grained subcategories, enabling a detailed analysis of existing models' strengths and weaknesses across meteorological subdomains. Based on that, we employ numerous state-of-the-art large language models, either general-purpose or meteorologically specialized, to examine their capability ceilings and preferences. Our results reveal critical weaknesses in current LLMs across specific meteorological subdomains, particularly meteorological observation, indicating key directions for further improvement. We hope that VME will serve as a standardized testbed for tracking progress and guiding the development of meteorological LLMs.

1 INTRODUCTION

Meteorology is an interdisciplinary and application-oriented field that integrates atmospheric science, observations, weather forecasting, and risk management. The complexity of meteorology and its strong reliance on specialized expertise make it difficult to train a sufficient number of expert meteorologists in operational scenarios. In this context, developing large language models (LLMs) with meteorological expertise to assist meteorologists in operational settings has become increasingly important. Based on that, many general-purpose (Qwen Team, 2026; GLM-5-Team et al., 2026; Moonshot AI, 2026) and meteorologically specialized large language models (Thulke et al., 2024; Huang et al., 2025; PMSCCMA, 2026) have incorporated meteorological corpora into their training processes, to enhance their meteorological knowledge and domain-specific capabilities.

Alongside the development of these LLMs, many benchmarks(Chen et al., 2025; Ma et al., 2024; Manivannan et al., 2025; Li et al., 2025; Kim et al., 2026) have been proposed to evaluate their capabilities. However, existing benchmarks suffer from two major limitations. First, some benchmarks (Ma et al., 2024; Manivannan et al., 2025; Kim et al., 2026) primarily focus on visual understanding tasks, which narrows their evaluation scope and consequently limits dataset scale. These constraints hinder a comprehensive assessment of LLMs' foundational meteorological reasoning capabilities. Second, many existing datasets (Manivannan et al., 2025) are derived primarily from textbooks and academic resources, making them less representative of the complexity and practical requirements of real-world meteorological operational scenarios. A comprehensive benchmark should evaluate models' capabilities in both academic meteorological knowledge and real-world operational scenarios. Therefore, there is an urgent need for a large-scale and comprehensive benchmark to systematically assess LLMs' meteorological understanding and reasoning abilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To this end, we propose a large-scale and comprehensive benchmark to assess LLMs' meteorological expertise and reasoning capabilities. We term it the Virtual Meteorologist's Exam (VME), a benchmark for evaluating whether large language models have professional-level meteorological expertise equivalent to that of humans. To be concrete, VME comprises nearly 6,500 question–answer pairs across four formats: single-choice questions (SCQ), multiple-choice questions (MCQ), true-or-false questions (TF), and cloze questions. To the best of our knowledge, it is the largest benchmark of its kind, containing approximately four times as many questions as comparable datasets. Beyond its scale, VME consists of five first-level and 25 fine-grained second-level categories, covering a broad range of common operational meteorological scenarios that combine scientific principles with practical operational knowledge. On top of that, we conduct extensive evaluations on our VME with numerous state-of-the-art large language models, either general-purpose or scientifically specialized. Our results show that current LLMs at best achieve a good level of performance, i.e., a score of approximately 60. However, it still fall far short of the expertise demonstrated by outstanding operational forecasters, i.e., a score of approximately 80, with particularly pronounced weaknesses in specific meteorological subdomains, such as meteorological observation. The findings presented in our paper highlight key directions for future improvement.

In summary, the main contributions of this paper are as follows:

- We introduce a large-scale and comprehensive benchmark for evaluating the meteorological knowledge and reasoning capabilities of current large language models.
- We systematically analyze the benchmark's data distribution, demonstrating its broad and balanced coverage of both academic knowledge and operational scenarios.
- We benchmark a wide range of advanced large language models on the proposed dataset, offering a comprehensive assessment of their current meteorological capabilities and valuable insights for future research.

2 Related Work

Large language models for meteorology. In recent years, advances in large language models have driven their increasing adoption across scientific disciplines, including meteorology and climate science. For example, ClimateGPT (Thulke et al., 2024) introduces a family of climate-specialized LLMs that combine domain-specific pre-training and expert-guided instruction tuning to synthesize interdisciplinary climate knowledge. Similarly, GeoGPT (Huang et al., 2025) leverages a large corpus of geoscience-related texts and academic publications to train a series of domain-specialized language models. More recently, an official meteorological model, namely Fenghe (PMSCCMA, 2026), was introduced and trained on a high-quality meteorological service corpus incorporating authoritative meteorological datasets. Beyond training models, several studies have incorporated external knowledge sources to enhance the capabilities of existing LLMs. For instance, ChatClimate (Vaghefi et al., 2023) grounds an LLM in the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC AR6) (Masson-Delmotte et al., 2021), improving the accuracy and reliability of its responses. Similarly, ClimSight (Kuznetsov et al., 2025) integrates LLMs with multi-source geographical and climate data to generate location- and activity-specific climate assessments. Despite these advances, existing studies primarily focus on developing advanced models, while leaving their professional meteorological capabilities under-evaluated in a unified setting. Our work addresses this gap by introducing VME, which enables consistent and systematic comparisons of meteorological knowledge and reasoning capabilities across those diverse large language models.

Meteorological datasets and evaluation benchmarks. Alongside the increasing adoption of large language models in meteorology, various benchmarks have been developed to assess their capabilities. Based on real-world meteorological data, ClimateIQA (Chen et al., 2025) proposes the Sparse Position and Outline Tracking (SPOT) algorithm to identify visual anomalies and accordingly introduces a multimodal dataset to evaluate these capabilities. Similarly, WeatherQA (Ma et al., 2024) introduces a dataset containing various meteorological variables and human experts' analyses as inputs to assess models' reasoning capabilities regarding severe weather events. In contrast, moving toward knowledge-driven evaluation, ClimaQA (Manivannan et al., 2025) constructs a Climate QA Generator pipeline (ClimGen) to generate questions from graduate-level textbooks and presents two sets of benchmarks with different standards, i.e, ClimaQA-Gold and ClimaQA-Silver. Moreover, K-MetBench (Kim et al., 2026) advances professional-level assessment through 1,774 questions

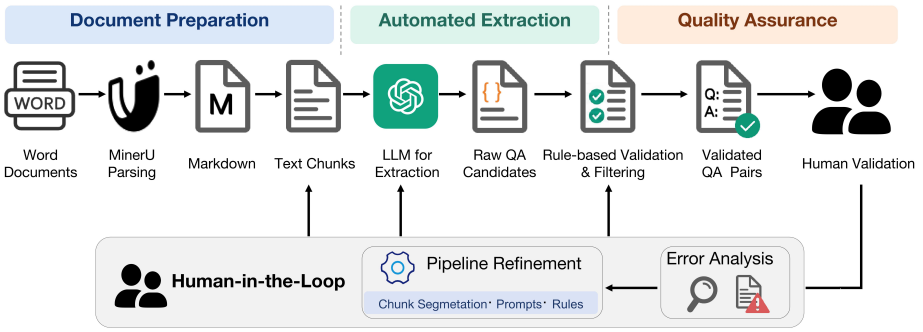

Figure 1: **Benchmark Construction Workflow.** First, we convert 123 exam papers from Chinese meteorological operational skill competitions into Markdown format and split those texts into textual chunks. Then, we feed those text chunks into a large language model, i.e., GPT-5.5, to extract candidate question–answer pairs. After that, we employ pre-defined rules to filter out invalid question-answer pairs. Throughout the process, the human review mechanism is introduced to improve the procedure of text segmentation, LLM prompts, and filtering rules.

from the Korean National Meteorological Engineer Examination, evaluating expert reasoning, visual understanding, geo-cultural knowledge, and meteorological expertise across five subject areas. Unlike existing benchmarks, VME is designed to cover both academic and operational meteorological scenarios, achieving large-scale coverage while maintaining data diversity, which should enable a comprehensive assessment of LLMs' meteorological understanding and reasoning capabilities.

3 Virtual Meteorologist's Exam

In this section, we detail how the benchmark is constructed and illustrate the properties of this benchmark. For benchmark construction, we split the whole procedure into three parts, i.e., Document Preparation, Automated Extraction and Quality Assurance. For the properties discussion, we compare our benchmark with various existing counterparts and present detailed statistics on our benchmark.

3.1 Benchmark Construction

Document Preparation. In terms of data sources, we collect 123 exam papers from Chinese meteorological operational skill competitions, spanning the period from 2016 to 2024. Each paper is accompanied by official ground-truth answers, ensuring the accuracy and reliability of the answers. In addition, the collection covers both national-level and regional-level competitions. The original files are in Word format, and we convert them into PDF format for subsequent PDF parsing. For PDF parsing, we employ MinerU (Wang et al., 2024), a document parsing engine, to extract the document structures and contents. These are further converted into plain-text Markdown format. After that, we split each document into fixed-size chunks to accommodate the context windows of downstream large language models, with each chunk containing approximately 16,000 characters.

Automated Extraction. For text extraction, we employ GPT-5.5 as the extraction model and design prompts to guide it in extracting the questions and their corresponding answers. We store the outputs generated by the LLMs in JSON files for subsequent evaluation and analysis. Based on these raw question-answer (QA) candidates, we use a set of validation rules to assess their quality and remove candidates that violate predefined criteria. We provide those rules and criteria in the Appendix.

Quality Assurance. To ensure the quality of the generated questions, we incorporate a human-in-the-loop quality control process with manual verification. Based on the feedback collected during this process, we iteratively refine the entire pipeline, including adjusting chunk sizes to better align with the effective context window of the LLM, refining prompts based on observed extraction behaviors to better control LLM outputs, and progressively incorporating extensive general-purpose validation rules. We adopt a two-level quality control strategy. At the document level, we check

Dataset	Domain	Modality	Size	Source	Academic	Operational
ClimaQA-Gold	Climate	Text	566	Grad Text	✓	
ClimaQA-Silver	Climate	Text	3k	Grad Text	✓	
ClimateIQA	Climate	Image + Text	152k	Template		✓
WeatherQA	Forecast	Image + Text	600	Experts' Analysis		✓
K-MetBench	Meteorology	Image + Text	1.7k	Exam Papers	✓	✓
VME (Ours)	Meteorology	Text	6.5k	Exam Papers	✓	✓

Table 1: Comparison with existing meteorological benchmarks.

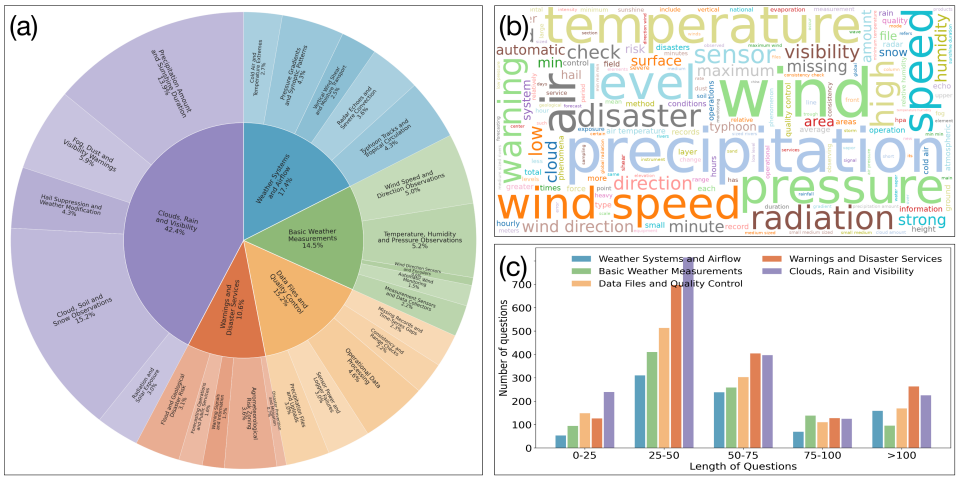

Figure 2: **Properties of VME.** (a) First-level and second-level taxonomy categories. We collect the TF-IDF features of VME and cluster them into five first-level and 25 fine-grained second-level categories via K-means algorithm. (b) Word cloud of keywords. (c) Question length distribution.

samples of the extracted questions from each document. At the dataset level, we randomly sample questions from the entire question pool for further verification. For each sampled question, we trace it back to its original source, identify potential issues, and continuously add corresponding rules to the pipeline to improve the data construction process.

The initial LLM-based extraction yields approximately 25,910 raw question-answer candidates. After applying a set of fundamental filtering rules, including question merging, deduplication, and validity checking, we retain 16,828 candidates. We further perform domain-specific filtering based on the question content, removing questions related to policy regulations, administrative procedures, and other topics beyond meteorological knowledge. As a result, we obtain a final collection of 6,508 high-quality meteorological QA samples, consisting of 2,641 single-choice questions (SCQ), 2,202 multiple-choice questions (MCQ), 1,531 true-or-false questions (TF), and 134 cloze questions.

3.2 Properties of Benchmark

Dataset Comparison. Here, we compare our benchmark with existing benchmarks in climate and meteorology. Through this comparison, we have two observations. First, most existing benchmarks are relatively small in scale, due to the demands of multimodal alignment or constraints imposed by their data sources, except for ClimateIQA (Chen et al., 2025). Such limited scale may hinder comprehensive and robust evaluation of LLMs across diverse meteorological knowledge and reasoning tasks. Second, existing benchmarks often focus exclusively on either academic knowledge or operational scenarios, which may limit their ability to assess whether models can bridge theoretical understanding and practical meteorological applications. Among these benchmarks, K-MetBench (Kim et al., 2026) is the closest counterpart to VME in covering both settings. Nevertheless, VME is

Model type	Model	SCQ	MCQ	TF	Cloze	Avg.
Specialized	GeoGPT-Llama3.1-70B	44.19	22.12	49.05	17.91	37.32
	ClimateGPT-3-8B	45.40	21.84	49.05	11.94	37.59
	GeoGPT-Qwen2.5-72B	49.26	20.07	47.42	22.39	38.40
	Fenghe-100B	50.89	26.84	52.97	23.88	42.68
	Intern-s2-preview-35B	55.51	32.11	56.76	28.36	47.32
General-purpose	Glm-5.2	64.79	38.78	64.60	43.28	55.50
	Qwen-3.8-max	63.08	43.51	63.95	41.79	56.22
	Gpt-5.5	68.04	42.33	62.05	48.51	57.53
	Deepseek-v4-pro	65.85	46.19	62.12	47.76	57.94
	Claude-opus-4.8	66.38	45.46	67.73	46.27	59.08
	Kimi-2.6	72.47	50.59	68.65	55.97	63.82

Table 2: Evaluation on VME across different large language models, including both general-purpose and specialized variants. For the average metric, we multiply the accuracy of each category by its proportion of the total number of questions to obtain a normalized value.

nearly four times larger than K-MetBench, providing broader coverage of meteorological topics and supporting a more comprehensive assessment of model capabilities.

Dataset Statistics To better understand the composition of the dataset, we organize it into five first-level categories and 25 fine-grained second-level categories. Specifically, we first extract their TF-IDF features (Salton & Buckley, 1988) from the questions and answers, and reduce them to 80 dimensions via the truncated singular value decomposition (SVD) algorithm. Then, we apply K-means clustering algorithm to partition these items into five clusters. After clustering, we characterize each cluster using its cluster-level TF-IDF keywords and the questions nearest to its centroid, and assign concise category names based on this evidence. As shown in Figure 2-(a), the five first-level topics are Clouds, Rain and Visibility (CRV), Weather Systems and Airflow (WSA), Data Files and Quality Control (DFQ), Basic Weather Measurements (BWM), and Warnings and Disaster Services (WDS). Among them, CRV and WSA are primarily associated with academic scenarios, whereas DFQ, BWM, and WDS are more closely related to operational scenarios. Similarly, we apply the same procedure to generate five second-level categories based on the questions within each first-level category. Overall, we divide the benchmark into 25 second-level categories.

In addition, in order to provide an intuitive overview of the benchmark's keywords and their frequencies, we use the word cloud to visualize its lexical distribution. As shown in Figure 2-(b), core meteorological concepts such as temperature, precipitation, wind, pressure, and radiation appear frequently, while terms such as sensor, visibility, warning, and disaster reflect its coverage of both scientific principles and operational scenarios, aligning with our original goal in constructing the dataset. Finally, we also present the distribution of question lengths across the entire dataset in Figure 2-(c). As shown, a considerable fraction of the questions are relatively lengthy, with many exceeding 50 words. This characteristic further demonstrates the challenging nature of our benchmark, as longer questions require models to process more complex linguistic structures and place greater demands on their comprehension and reasoning capabilities.

4 Experiments

4.1 Implementation Details

To comprehensively evaluate the capabilities of existing large language models in meteorology, we select a range of representative mainstream large language models, including general-purpose variants trained on diverse data sources, as well as specialized models fine-tuned on meteorological or climate-related data. The specialized models include GeoGPT (Huang et al., 2025), ClimateGPT (Thulke et al., 2024), Intern-S2 (Bai et al., 2026) and Fenghe (PMSCCMA, 2026), while the general-purpose models consists of Qwen-3.8-max (Qwen Team, 2026), Gpt-5.5 (OpenAI, 2026), Deepseek-v4-pro (DeepSeek-AI, 2026), Glm-5.2 (GLM-5-Team et al., 2026), Claude-opus-4.8 (Anthropic, 2026), Kimi-2.6 (Moonshot AI, 2026).

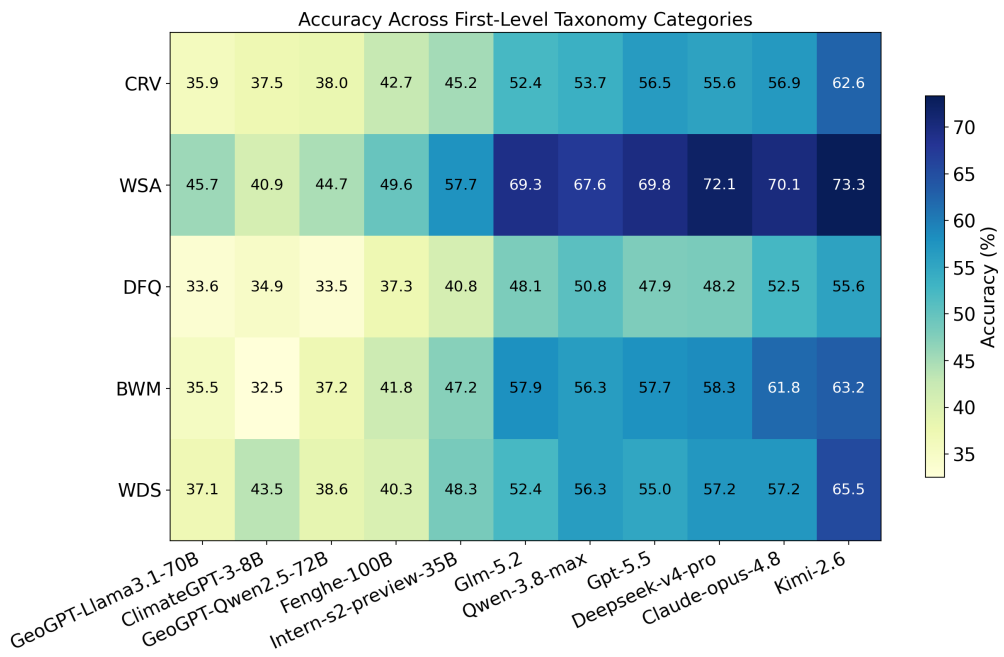

Figure 3: **Category-wise Performance Analysis.** We assess the performance of various large language models across different first-level categories in spite of question type.

During the evaluation process, we design corresponding prompts to encourage models to perform logical reasoning through a step-by-step thinking process, thereby obtaining the final answers. For objective questions, including single-choice, multiple-choice, and true-or-false questions, we assess the correctness of each response by comparing it with the reference answer and then compute the overall accuracy. For subjective questions such as fill-in-the-blank questions, we adopt an LLM-as-a-judge approach, allowing benign variations between the predicted response and the reference answer in spacing, wording, or abbreviations, as long as they are semantically equivalent. We adopt Gpt-5.5 as the judge. For reproducibility, we set the temperature to 0 and top_p to 1 for most models to minimize random sampling during inference. However, for the Kimi-2.6 model, top_p can only be set to 0.95. In addition, the random seed is set as 42.

4.2 Results

Unless stated otherwise, we conduct the experiments with the aforementioned configurations. We first evaluate a range of large language models on our benchmark. Then, we analyze their performance across multiple dimensions to identify the common limitations shared by these LLMs.

Performance Comparison. As shown in Table 2, we present the evaluation results of various large language models on our dataset, including both general-purpose and specialized variants. For the average metric, we multiply the accuracy of each category by its proportion of the total number of questions to obtain a normalized value. Based on that, we have the following observations. Although specialized models (Huang et al., 2025; Thulke et al., 2024; PMSCCMA, 2026) are developed specifically for meteorology or climate science, current general-purpose models generally outperform them. This may be attributable to several factors, including differences in model versions, model sizes, and the training data used. Among these models, Kimi-2.6 achieves the highest reported accuracy across multiple question types, particularly SCQ and TF, on which it scores 72.47% and 68.65%, respectively. Besides, we further observe that accuracy is higher on TF questions than on Cloze questions. TF questions generally assess the capacity for contextual understanding, whereas Cloze questions may require additional reasoning capacities based on the given context. This pattern may suggest that current large language models may be better at understanding meteorological knowledge than at using that knowledge for reasoning. Finally, we collect scores from meteorological professionals on these questions. Non-expert meteorological professionals typically

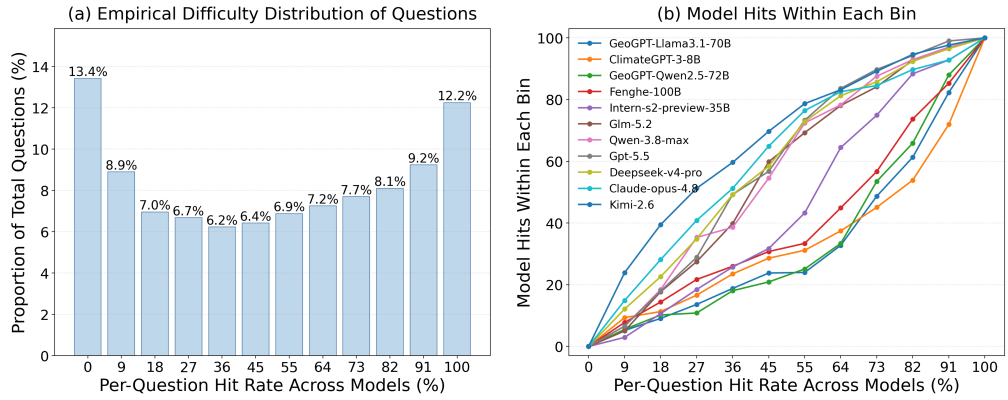

Figure 4: **Dataset Complexity Analysis.** (a) We categorize all questions based on their hit rates across models to showcase the distribution of difficulty levels. (b) We compare model performance across different difficulty levels to identify the distribution of capability gaps among the models.

scored around 60, whereas experts achieved scores of up to 80. Based on this criterion, current large language models at best achieve a good level of performance, while still falling far short of the expertise demonstrated by outstanding operational forecasters. These results establish a clear performance baseline for understanding the current level of meteorological capacities demonstrated by large language models.

Category-wise Performance Analysis. Beyond comparing the performance of large language models across question types, we also assess their capabilities across different first-level categories. Specifically, we compute and visualize each model's accuracy across the five previously defined first-level taxonomy categories, grouping questions by their respective subdomains without distinguishing between question types. As shown in Figure 3, those models perform relatively well in the WSA categories, where they perform the worst in DFQ. The relatively weak performance in DFQ may be attributed to its reliance on specialized operational knowledge, including data formats, transmission procedures, quality-control rules, missing-record handling, and equipment-related failures. Those factors may be underrepresented in general-purpose pretraining corpora. Through this evaluation, our benchmark highlights a clear weakness of current large language models in this domain, which requires further improvement in future model development.

Dataset Complexity Analysis. To assess the overall difficulty of the dataset, we categorize questions based on the per-question hit rate across models, defined as the proportion of evaluated models that answer each question correctly. Given 11 models, we classify the questions into 12 difficulty levels, ranging from questions answered correctly by none of the models (0% hit rate) to those answered correctly by all models (100% hit rate). In Figure 4-(a), from left to right, question difficulty decreases from hardest to easiest, while the hit rate increases from lowest to highest. We observe that approximately 12.2% of the questions are answered correctly by all evaluated models, whereas approximately 13.4% are answered correctly by none of them. This suggests that substantial limitations still remain to be addressed in current large language models. In addition, we plot each model's hit rate across the difficulty levels to examine how its performance varies with question difficulty. As shown in Figure 4-(b), most general-purpose models perform better than specialized models across different difficulty levels. The largest performance gap among different types of models occurs in the difficulty range corresponding to a hit rate of 55%. In addition, Kimi-2.6 achieves the best overall performance, particularly on the most difficult questions, where it leads by a substantial margin in the difficulty range corresponding to hit rates between 18% and 27%.

Illustration Difficulty Distribution To further explore how questions in different first-level categories are distributed across difficulty levels, we construct a heatmap to examine the fine-grained relationship between question category and difficulty. We normalize each row to show the distribution of questions across difficulty levels within each first-level category. As shown in Figure 5, we observe that approximately 16% of the DFQ, BWM, and WDS questions fall into the first bin, corresponding to a 0% hit rate, indicating that none of the evaluated models answered these questions correctly. In contrast, 18.8% of WSA questions fall into the last bin (100% hit rates), suggesting

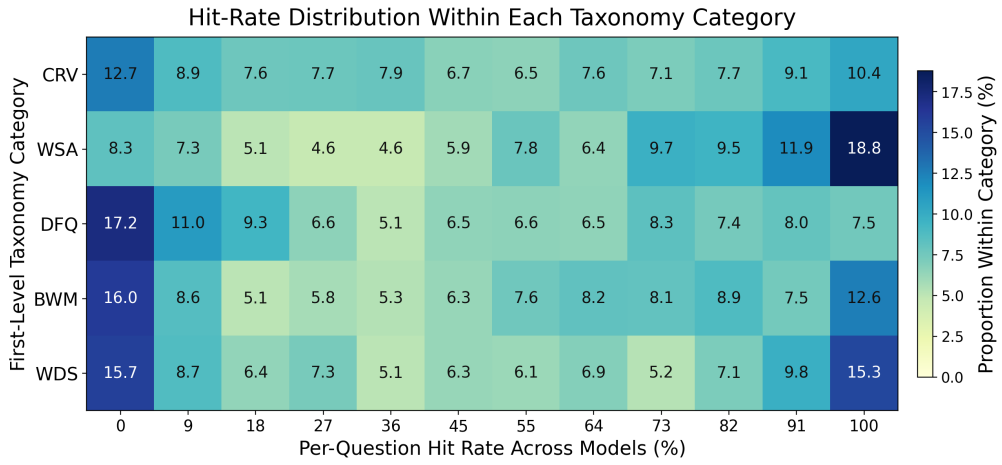

Figure 5: **Illustration of Difficulty Distribution.** We construct a heatmap to examine the fine-grained relationship between first-level categories and various difficulty levels. We normalize each row to show the distribution of questions across difficulty levels within each first-level category.

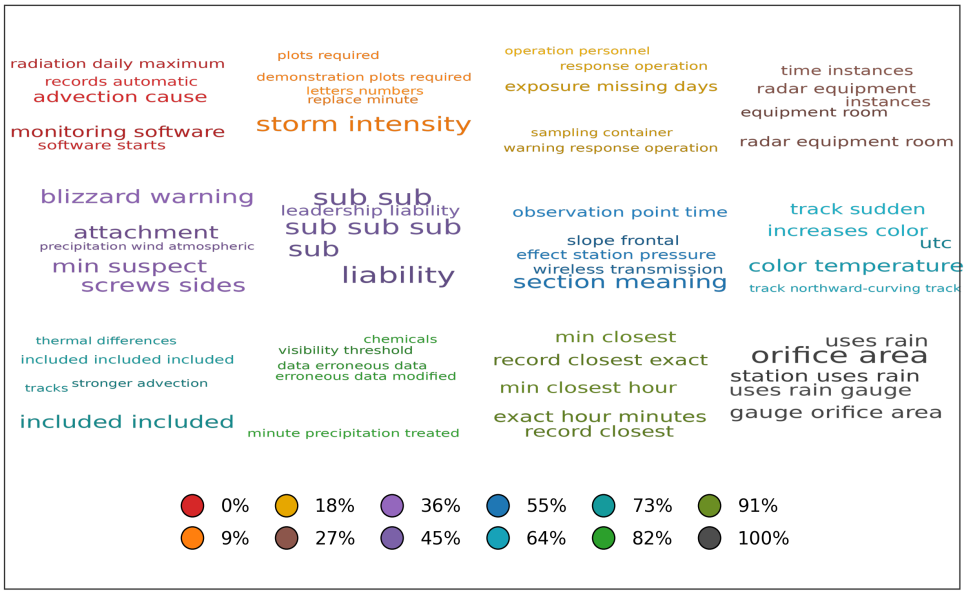

Figure 6: **Word Cloud of Highest-scoreing items.** To characterize the most challenging difficulty levels, we group questions by difficulty, compute the TF–IDF scores of individual words and phrases, and visualize the most salient terms using word clouds.

that the models have relatively broad coverage of the WSA knowledge assessed in this dataset. In addition, to provide an intuitive overview of the keywords associated with each difficulty level, we group questions by difficulty and compute their TF-IDF scores for individual words and phrases. We then visualize the highest-scoring terms using word clouds. As shown in Figure 6, we find that the keywords in the first bin (0% hit rate) mainly include "record automatic" and "monitoring software", which are primarily related to observation and monitoring. This suggests potential gaps in the evaluated models' knowledge of these topics and highlights the need to improve their capabilities in these areas.

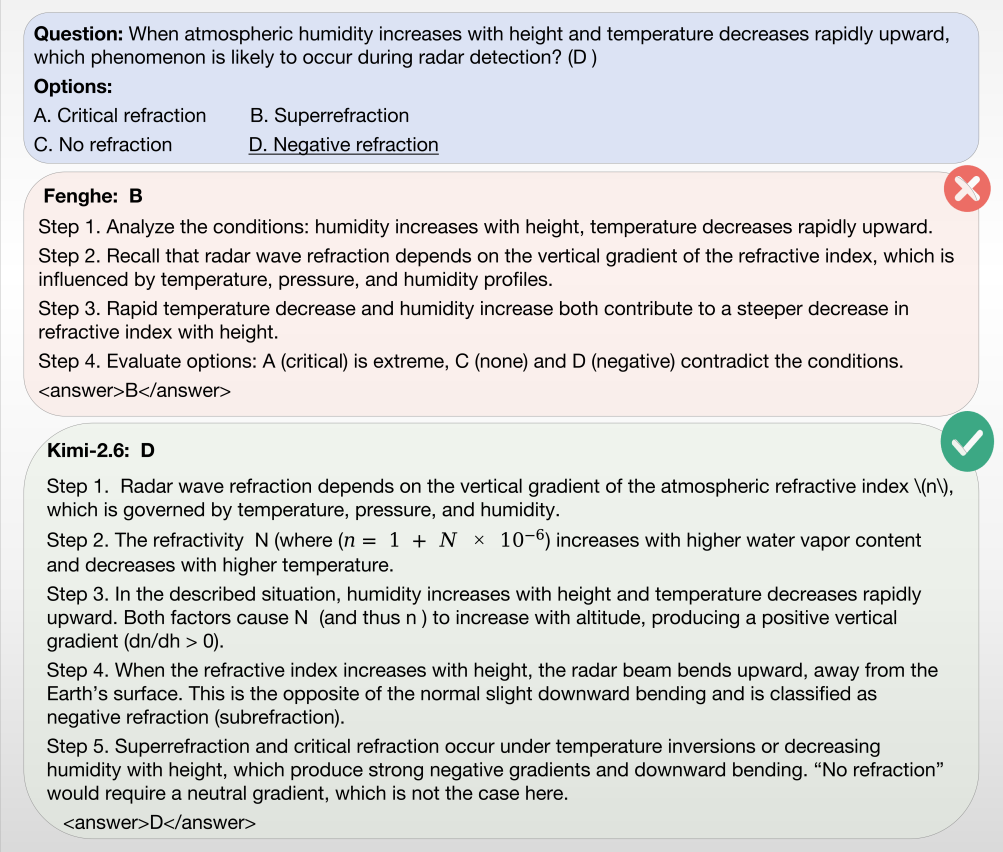

Figure 7: **Case Analysis.** To intuitively understand the differences in models' prediction behavior, we select several questions and their corresponding model responses for case analysis.

4.3 Case Analysis

To gain an intuitive understanding of differences in models' prediction behavior, we select several questions and their corresponding model responses for case analysis. We focus on cases where the specialized models answer incorrectly but the general-purpose models answer correctly. As shown in Figure 7, the specialized models' responses tend to be brief and consist merely of factual statements, often stating a conclusion without deep reasoning. In contrast, the general-purpose models typically provide longer responses with more detailed reasoning steps. General-purpose models are more likely to rely on and explicitly present precise formulas to frame problems during their reasoning process, rather than relying on simple verbal descriptions. Here, we present only a subset of the cases and model responses. Please refer to the Appendix for more detail.

4.4 Conclusion

In this paper, we introduce the Virtual Meteorologist's Exam (VME), a benchmark designed to assess whether existing LLMs possess professional meteorological expertise comparable to that of human experts. VME comprises nearly 6,500 question–answer pairs across four formats, covering both academic knowledge and operational scenarios. We further organize it into five first-level categories and 25 fine-grained second-level categories, enabling a detailed analysis of model strengths and weaknesses across meteorological subdomains. Through extensive evaluations of various advanced LLMs, we find that even the best-performing models achieve only a good level of proficiency and still fall substantially short of the expertise demonstrated by outstanding operational forecasters. We hope that VME will serve as a standardized testbed for tracking progress, identifying critical capability gaps, and guiding the development of more reliable and capable meteorological LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This benchmark is intended to advance research on large language models in meteorology. We aim to evaluate whether existing large language models have professional-level meteorological expertise equivalent to that of humans. We will release this benchmark once our paper is accepted. Finally, the authors in this work declare no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We will release the full dataset and accompanying evaluation scripts to facilitate its use. We hope that our benchmark will serve as a standardized testbed for tracking progress and guiding the development of meteorological large language models.

REFERENCES

Anthropic. Claude Opus 4.8 System Card. Technical report, Anthropic, May 2026. URL `https://www.anthropic.com/claude-opus-4-8-system-card`.

Lei Bai, Jiaqi Cao, Chiyu Chen, Guanzhou Chen, Kai Chen, Guangran Cheng, Erfei Cui, Xuanlang Dai, Shengyuan Ding, Shangheng Du, et al. Intern-s2-preview: Scientific agentic foundation model. *arXiv preprint arXiv:2608.13505*, 2026.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5322–5333, 2025.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang,

Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Fei Huang, Fan Wu, Zeqing Zhang, Qihao Wang, Long Zhang, Grant Michael Boquet, and Hongyang Chen. Geogpt-rag technical report. *arXiv preprint arXiv:2509.09686*, 2025.

Soyeon Kim, Cheongwoong Kang, Myeongjin Lee, Eun-Chul Chang, Lee Jaedeok, and Jaesik Choi. K-metbench: A multi-dimensional benchmark for fine-grained evaluation of expert reasoning, locality, and multimodality in meteorology. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5574–5612, 2026.

Ivan Kuznetsov, Antonia Anna Jost, Dmitrii Pantiukhin, Boris Shapkin, Thomas Jung, and Nikolay Koldunov. Transforming climate services with llms and multi-source data integration. *npj Climate Action*, 4(1):97, 2025.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Haobo Li, Zhaowei Wang, Jiachen Wang, YueYa Wang, Alexis Kai Hon Lau, and Huamin Qu. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17547–17573, 2025.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. Climaqa: An automated evaluation framework for climate question answering models. In *International Conference on Learning Representations*, volume 2025, pp. 25357–25382, 2025.

Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Yang Chen, Leah Goldfarb, Melissa I Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021.

Moonshot AI. Kimi K2.6. Hugging Face model card, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6.

OpenAI. GPT-5.5 System Card. Technical report, OpenAI, April 2026. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf.

PMSCCMA. Fenghe: A large language model for meteorological services, 2026. URL https://huggingface.co/PMSCCMA/FengHe. Accessed: 2026-09-17.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian Van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4 (1):480, 2023.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

A APPENDIX

A.1 BENCHMARK CONSTRUCTION CRITERIA

We first employ the large language model (LLM) to extract approximately 25,910 candidate question-answer records from the original examination materials. Then, we apply a series of rule-based cleaning procedures to the initially extracted records:

Filtering rules

- Cross-chunk merging: We merge records from different chunks of the same source.
- Duplicate removal: We identify duplicates by question type, number, and normalized text, retaining the record with stronger answer evidence.
- Question completeness checking: We remove questions with incomplete stems, missing context, or placeholder text.
- Figure and table dependency checking: We exclude questions that depend on unavailable figures, tables, or diagrams.
- Option structure validation: We check for missing, invalid, duplicate, or empty options.
- Answer format normalization: We use uppercase option letters for choice questions and binary labels for true/false questions.
- Answer-option consistency checking: We ensure that answer labels match available options and remove invalid cases.
- Answer evidence verification: We require explicit answer evidence from the original questions, answer documents, or answer tables.
- Missing-information repair: When reliable evidence exists, we recover missing options, true/false markers, and overwritten answers.
- Abnormal-record filtering: We remove records with conflicting answers, unrecoverable errors, or out-of-scope content.

After merging, deduplication, validation, repair, and filtering, 16,828 structurally valid and evidence-supported candidate questions remain. We then conduct a first-stage domain-filtering process based on question content. This stage retains questions that directly concern meteorological knowledge and technical meteorological operations and excludes content that primarily concerns:

Excluded categories

- Meteorological laws and regulatory provisions.
- Administrative licensing and approval procedures.
- Organizational structures and departmental responsibilities.
- Personnel training, assessment, and qualification management.
- General operational management procedures.
- Policy documents and development plans.
- General administrative knowledge with limited relevance to meteorological science or technology.

This stage removes 3,000 records and leaves 13,828 meteorological knowledge candidates. We subsequently apply a stricter core-domain retention procedure. Only questions that fall into one of the following four meteorological domains enter the final dataset:

Core domains

1. Meteorological Observation, Measurement, and Data Reporting: This category covers air temperature, atmospheric pressure, humidity, wind direction, wind speed, precipitation, visibility, radiation, automatic weather stations, sensors, missing-value processing, BUFR

messages, Z files, and meteorological reports.
2. Weather Forecasting, Hazard Warning, and Climate: This category includes weather forecasting, nowcasting, typhoons, heavy rainfall, hail, strong winds, cold waves, flash floods, meteorological hazard warnings, and climate analysis.
3. Weather Modification Technologies: This category covers artificial precipitation enhancement, hail suppression, rockets, anti-aircraft guns, cloud seeding, and other weather modification technologies.
4. Agricultural Meteorology and Ecological Services: This category includes agricultural meteorology, crop-weather relationships, soil moisture, agrometeorological hazards, and ecological meteorological services.

Questions that the keyword-based rules cannot explicitly assign to one of these four domains enter a review set and do not appear in the current version of the dataset. After this strict domain-retention stage, the final dataset contains 6,508 high-quality meteorological QA records

A.2 PROMPT EXAMPLE

Prompt for Text Extraction. During dataset construction, we use an LLM to extract structured question–answer pairs from the collected examination papers. We design the prompt as follows:

System prompt for text extration

You are a strict evidence extractor for Chinese meteorology exam papers. You are not a test solver. Never answer from meteorology knowledge, common sense, or option semantics.

Your job has two parts only:
1. Extract question structure from {SOURCE_PAPER_UNIT}.
2. Extract answers only when the answer is explicitly visible in {SOURCE_PAPER_UNIT} or {ANSWER_UNIT}.

Hard rules:
- verified=true is allowed only when answer_evidence is an exact substring copied from {SOURCE_PAPER_UNIT} or {ANSWER_UNIT}.
- Do not invent answer_evidence. Do not write synthetic evidence such as "answer: ABCD" unless that exact text exists in the provided input.
- If {ANSWER_UNIT} has an answer table, use the visible question number and answer position in that table. Do not guess.
- If the question number is visible in the question text, preserve it in local_question_number.
- If evidence is ambiguous, set verified=false and answer="".
- For single_choice answer, return one uppercase letter only.
- For multi_choice answer, return uppercase letters only, e.g. AB or ABCD.
- For true_false answer, normalize to true or false only.
- For fill_blank answer, return a string or list of strings extracted from the text.
- Return valid JSON only.

User prompt for text extration

Parse exam questions from {SOURCE_PAPER_UNIT} and extract answers only from explicit evidence.

The {QUESTION_FILE} and {ANSWER_FILE} are manually paired by the user; do not search for another answer file.

Return JSON with exactly one key: items.
Each item must contain:

- local_question_number: integer or null. Use the visible whole-paper question number when present.
- question_type: one of single_choice, multi_choice, fill_blank, true_false, unsupported
- question: cleaned question text, with answers removed; keep options for choice questions
- options: object mapping A/B/C/D... to option text, or empty object
- answer: string or list; empty if unverified
- verified: boolean
- answer_evidence: exact substring copied from {SOURCE_PAPER_UNIT} or {ANSWER_UNIT}; empty if unverified
- answer_source: source_unit_inline, source_unit_answer_marker, answer_unit_same_unit, answer_file_table, answer_file_highlight, or unverified
- evidence_extraction_note: briefly explain how answer is read, including suite/paper section if the answer file has multiple sets
- exclude_reason: empty if verified and usable; otherwise a short reason

Important rules:
- If {ANSWER_UNIT} contains several papers/sets, first match the suite/title/context of {QUESTION_FILE} or {SOURCE_PAPER_UNIT}, then use the visible question number.
- If {SOURCE_PAPER_UNIT} includes embedded answers, remove them from question and quote the embedded answer text as evidence.
- If blanks are already filled in the source, replace the filled answer with a blank marker in question.
- If answers are shown by table, quote the table row/block that contains the question number and answer.
- If answers are shown by highlight, font color, or markup text, quote the highlighted/colored/marked fragment as evidence.
- The question field must contain the full visible question text. Never summarize, abbreviate, use ellipses, or write placeholders such as "see SOURCE_PAPER_UNIT".
- For choice questions, include all visible options in options. If the full question or options are not visible, set verified=false and explain why.
- Determine question_type from the surrounding section heading in {SOURCE_PAPER_UNIT}. A question inside multi_choice must be multi_choice even if it only shows A-D options or one blank.
- Prefer visible local_question_number plus the source section heading over qtype ordinal/counting. Do not pair a question in one section with an answer table from another section.
- If the same normalized question appears more than once in one pair, keep the version whose section, local question number, options, and answer evidence form a closed chain; do not keep a wrong duplicate.
- Drop questions that require missing external context such as a figure/table/number reference not included in the question text, empty options, duplicated option labels, or incomplete option extraction.
- Never answer from knowledge. verified=true requires exact evidence in {SOURCE_PAPER_UNIT} or {ANSWER_UNIT}.

Prompt for Answer Generation. During the evaluation process, we apply the following prompts to guide vairous large language models to generate answers.

Single-choice prompt

You are an expert in meteorology, hydrology, and Chinese technical standards.

Answer the single-choice question below. The question has exactly one correct option.

Question: {question}

Rules:
- Choose exactly one correct option from A, B, C, and D.
- Reason step by step before giving the final answer.
- The final answer must contain one option letter only.
- Return one letter, such as ”A”.
- Do not use the option content as the final answer.
- If unsure, choose the most likely option based on domain knowledge.
- The final answer should be enclosed by <answer >and </answer >.

Example:
Step 1. ...
Step 2. ... ... <answer >A </answer >

Multi-choice prompt

You are an expert in meteorology, hydrology, and Chinese technical standards.

Answer the multiple-choice question below. The question may have one or more correct options.

Question: {`question`}

Rules:
- Choose all correct options from A, B, C, and D.
- Reason step by step before giving the final answer.
- The final answer must contain option letters only.
- If there is only one correct option, return one letter, such as ”A”.
- If there are multiple correct options, return the letters in alphabetical order with no spaces, such as ”AC” or ”BCD”.
- Do not use the option content as the final answer.
- If unsure, choose the most likely option or combination based on domain knowledge.
- The final answer should be enclosed by <answer >and </answer >.

Example:
Step 1. ...
Step 2. ...
... <answer >AC </answer >

True/false prompt

You are an expert in meteorology, hydrology, and Chinese technical standards.

Answer the true-or-false question below. The question has exactly one correct answer.

Question: {`question`}

Rules:
- Choose exactly one correct answer from True and False.
- Reason step by step before giving the final answer.
- The final answer must contain one answer only.
- Return ”True” or ”False”.
- Do not use any other words as the final answer.
- If unsure, choose the most likely answer based on domain knowledge.

- The final answer should be enclosed by <answer >and </answer >.

Example:
Step 1. ...
Step 2. ...
... <answer >True </answer >

Cloze prompt

You are an expert in meteorology, hydrology, and Chinese technical standards.

Answer the fill-in-the-blank question below.

Question: {question}

Rules:

- Identify and fill every blank. A blank may be written as an empty pair of parentheses or as underscores, for example (), ____, (____).
- Give the answers in the same order as the blanks appear in the question.
- Reason step by step before giving the final answer.
- The final answer must be a JSON array of strings, with one string for each blank.
- Return only the words or phrases that fill the blanks; do not repeat the question.
- If unsure, choose the most likely option based on domain knowledge.
- The final answer should be enclosed by <answer >and </answer >.

Example:
Step 1. ...
Step 2. ...
... <answer >["message", "Emergency Warning"] </answer >

Prompt for Cloze Evaluation. As mentioned in the implementation details, we adopt an LLM-as-a-judge approach (Li et al., 2024; Gu et al., 2026) to evaluate subjective questions, such as cloze questions. Such evaluation allows benign variations between the predicted response and the reference answer in spacing, wording, or abbreviations, as long as they are semantically equivalent.

Llm-as-a-judge prompt

You are a strict grader for fill-in-the-blank questions. Compare the reference answers and predicted answers in order. Return true only when every predicted answer has the same meaning as its corresponding reference answer. Allow harmless differences in spacing, units, wording, abbreviations, and equivalent terminology. Do not accept answers that change the meaning, omit required information, or put answers in the wrong order. If the numbers of answers differ, return false. Output exactly one word: true or false.

A.3 Details of taxonomy categories

As described in the benchmark properties, we organize the entire dataset into five first-level categories and 25 fine-grained second-level categories. Here, we present these categories in detail, along with their respective proportions.

A.4 Case Analysis

To gain an intuitive understanding of differences in models' prediction behavior, we select several questions and their corresponding model responses for case analysis. We focus on cases where the specialized models answer incorrectly but the general-purpose models answer correctly.

First-level categories	Second-level categories	Proportion
Weather Systems and Airflow (WSA)	Cold Air and Temperature Extremes	2.7%
	Pressure Gradients and Synoptic Patterns	4.3%
	Vertical Wind Shear and Moisture Transport	2.5%
	Radar Echoes and Severe Convection	3.6%
	Typhoon Tracks and Tropical Circulation	4.3%
Basic Weather Measurements (BWM)	Wind Speed and Direction Observations	5.0%
	Temperature, Humidity and Pressure Observations	5.2%
	Wind Direction Sensors and Encoders	0.6%
	Automatic Wind Monitoring	1.5%
	Measurement Sensors and Data Collectors	2.2%
Data Files and Quality Control (DFQ)	Missing Records and Time-Series Gaps	2.3%
	Consistency and Range Checks	2.2%
	Operational Data Processing	4.6%
	Sensor Power and Logger Failures	3.0%
	Precipitation Files and Uploads	3.0%
Warnings and Disaster Services (WDS)	Disaster Prevention and Mitigation	0.7%
	Agrometeorological Risk Zoning	3.6%
	Warning Signals and Information	1.5%
	Forecasting Operations and Public Services	1.6%
	Flood and Geological Disaster Risk	3.1%
Clouds, Rain and Visibility (CRV)	Radiation and Solar Exposure	3.0%
	Cloud, Soil and Snow Observations	15.2%
	Hail Suppression and Weather Modification	4.3%
	Fog, Dust and Visibility Warnings	5.9%
	Precipitation Amount and Sunshine Duration	13.9%

Table 3: Details of taxonomy categories. The entire dataset can be grouped into five first-level categories and 25 fine-grained second-level categories.

Case-1

Question: When atmospheric humidity increases with height and temperature decreases rapidly upward, which phenomenon is likely to occur during radar detection? (D)
Options:
A. Critical refraction
B. Superrefraction
C. No refraction
D. Negative refraction

GeoGPT-Llama3.1-70B Response:
When atmospheric humidity increases with height and temperature decreases rapidly upward, the refractive index of the atmosphere increases with height. This condition can lead to superrefraction, where radar waves are bent more than normal, potentially causing them to follow the curvature of the Earth and travel further than they would under normal conditions.
<answer >B </answer>

ClimateGPT-3-8B Response:
Step 1. When atmospheric humidity increases with height and temperature decreases rapidly upward, the refractive index of the atmosphere decreases with height.
Step 2. This decrease in refractive index causes radar waves to bend towards the ground, resulting in a phenomenon known as superrefraction.
Step 3. Superrefraction is a common phenomenon in radar detection, particularly in areas with high humidity and temperature gradients.
Step 4. Based on the given information, the correct answer is B. Superrefraction.
<answer >B </answer >

GeoGPT-Qwen2.5-72B Response:
When atmospheric humidity increases with height and temperature decreases rapidly upward, it creates conditions where the refractive index gradient becomes very steep. This can lead to superrefraction, where radar waves bend downward more than they would under normal conditions, potentially causing them to follow the curvature of the Earth's surface for longer distances. The correct option is B.
<answer >B </answer>

Fenghe-100B Response:
Step 1. Analyze the conditions: humidity increases with height, temperature decreases rapidly upward.
Step 2. Recall that radar refraction depends on the vertical gradient of the refractive index.
Step 3. Rapid temperature decrease and humidity increase both contribute to a steeper decrease in refractive index with height.
Step 4. This results in a more pronounced downward bending of radar waves, known as superrefraction.
Step 5. Evaluate options: A (critical) is extreme, C (none) and D (negative) contradict the conditions.
<answer >B </answer >

Intern-s2-preview-35B Response:
Step 1: Analyze the atmospheric conditions described - humidity increases with height and temperature decreases rapidly upward.
Step 2: Recall that radar wave refraction depends on the vertical gradient of the refractive index, which is influenced by temperature, pressure, and humidity profiles.
Step 3: When temperature decreases rapidly with height (strong lapse rate) and humidity increases with height, the refractive index decreases more rapidly with height than in the standard atmosphere.
Step 4: A rapid decrease in refractive index with height causes radar waves to bend more strongly downward toward the Earth's surface than under normal conditions.
Step 5: This enhanced downward bending is defined as superrefraction. Critical refraction is a specific extreme case of superrefraction where the wave follows Earth's curvature exactly, while negative refraction involves upward bending (opposite of what occurs here).
Step 6: Therefore, the described conditions most likely produce superrefraction during radar detection.
<answer >B </answer >

Glm-5.2 Response:
Step 1. Analyze the atmospheric conditions given in the question: humidity increases with height and temperature decreases rapidly upward.
Step 2. Recall the relationship between atmospheric conditions and radar beam refraction. The refractive index of air depends on pressure, temperature, and humidity.
Step 3. When temperature decreases rapidly with height (a steep lapse rate) and humidity increases with height, the vertical gradient of the refractive index becomes positive (i.e., the refractive index increases with height).
Step 4. A positive vertical gradient of the refractive index causes the radar beam to bend upward, away from the surface of the Earth. In radar meteorology, this phenomenon is known as negative refraction (or subrefraction).
Step 5. Compare this with the options. Critical refraction and superrefraction occur when the refractive index decreases rapidly with height (e.g., temperature inversions and rapid humidity decrease). Therefore, the correct option is D.
<answer >D </answer >

Qwen-3.8-max Response:
Analyzing Radar Refraction Under Given Atmospheric Conditions
Step 1: Understanding the atmospheric conditions

The question describes two conditions: - Humidity increases with height (moisture content is greater aloft) - Temperature decreases rapidly with height (large lapse rate)
Step 2: Effect on atmospheric refractivity The atmospheric refractivity N is given by: $N = \frac{77.6}{T}\left(P + \frac{4810e}{T}\right)$
- Increasing humidity with height → water vapor pressure e increases with height → N increases with height - Rapid temperature decrease with height → since $N \propto \frac{1}{T}$, lower T aloft means higher N aloft → N increases with height
Both effects cause the refractivity gradient dN/dh to become positive (or at least much less negative than the normal value of about -39 N-units/km).
Step 3: Determining the type of refraction
- Normal/positive refraction: dN/dh is negative (beam bends toward Earth)
- Superrefraction: dN/dh is very negative (¡ -157 N-units/km), typically from temperature inversions and rapid humidity decrease
- Critical refraction: dN/dh = -157 N-units/km
- Negative refraction (sub-refraction): dN/dh is positive or near zero → the radar beam curves upward, away from the Earth's surface
Since both stated conditions make dN/dh positive, the radar beam bends upward — this is negative refraction.
Step 4: Conclusion
The described atmospheric profile (humidity increasing and temperature decreasing rapidly with height) leads to negative refraction during radar detection.
<answer >D </answer >

Gpt-5.5 Response:
Step 1. Radar refraction depends on the vertical gradient of atmospheric refractive index.
Step 2. Increasing humidity with height and rapidly decreasing temperature upward tend to make refractive index increase with height.
Step 3. This causes the radar beam to bend upward, known as negative refraction.
<answer >D </answer >

Deepseek-v4-pro:
Step 1. Radar wave refraction depends on the vertical gradient of atmospheric refractive index.
Step 2. When humidity increases with height and temperature decreases rapidly upward, the refractive index tends to increase with height or decrease less than normal, causing radar waves to bend upward rather than downward.
Step 3. This upward bending relative to normal radar propagation is referred to as negative refraction.
<answer >D </answer >

Claude-opus-4.8 Response:
Step 1. Radar refraction is determined by the vertical gradient of atmospheric refractivity N, which depends on temperature, pressure, and water vapor pressure: N = (77.6/T)(P + 4810e/T).
Step 2. Normally N decreases with height, bending the beam gently downward (normal refraction).
Step 3. Superrefraction requires N to decrease rapidly with height, typically caused by temperature inversion (warming aloft) and humidity decreasing with height.
Step 4. Here humidity increases with height (e larger aloft → raises N aloft) and temperature decreases rapidly (lower T aloft → also raises N aloft). Both effects make N decrease slowly with height or even increase, reversing the normal gradient.
Step 5. When the refractivity gradient becomes weak or positive, the beam bends upward instead of downward — this is subrefraction / negative refraction.
<answer >D </answer >

Kimi-2.6 Response:
Step 1. Radar wave refraction depends on the vertical gradient of the atmospheric refractive

index n, which is governed by temperature, pressure, and humidity.
Step 2. The refractivity N (where $n = 1 + N \times 10^{-6}$)) increases with higher water vapor content and decreases with higher temperature.
Step 3. In the described situation, humidity increases with height and temperature decreases rapidly upward. Both factors cause N (and thus n) to increase with altitude, producing a positive vertical gradient (dn/dh ¿ 0).
Step 4. When the refractive index increases with height, the radar beam bends upward, away from the Earth's surface. This is the opposite of the normal slight downward bending and is classified as negative refraction (subrefraction).
Step 5. Superrefraction and critical refraction occur under temperature inversions or decreasing humidity with height, which produce strong negative gradients and downward bending. "No refraction" would require a neutral gradient, which is not the case here.
<answer >D </answer >

Case-2

Question: Which of the following situations is favorable for frontogenesis? (C)
Options:
A. Wind speed increases along the direction of the potential temperature gradient
B. In stable stratification, the warm air mass rises and the cold air mass sinks
C. Latent heat release during the lifting of warm, moist air
D. Cold air moves southward over a warm underlying surface

GeoGPT-Llama3.1-70B Response:
Frontogenesis refers to the process of the formation or intensification of a front, which is a boundary between two air masses of different densities. This process is favored by conditions that enhance the temperature gradient and convergence of air masses.
A. Wind speed increases along the direction of the potential temperature gradient: This situation can lead to frontogenesis because it can cause the air masses to converge, enhancing the temperature gradient and leading to the formation or intensification of a front.
B. In stable stratification, the warm air mass rises and the cold air mass sinks: This situation is not favorable for frontogenesis because it describes a stable atmosphere where the warm air rises and the cold air sinks, which does not typically lead to the formation of a front.
C. Latent heat release during the lifting of warm, moist air: This situation is more related to the development of convection and precipitation rather than frontogenesis. While it can influence the dynamics of a front, it is not a direct cause of frontogenesis.
D. Cold air moves southward over a warm underlying surface: This situation can lead to frontogenesis because the cold air moving over a warm surface can create a temperature gradient and potentially lead to the formation of a front.
Based on the options, the most favorable situation for frontogenesis is when the wind speed increases along the direction of the potential temperature gradient, as it directly contributes to the convergence and intensification of the temperature gradient.
<answer >A </answer >

ClimateGPT-3-8B Response:
Step 1. Frontogenesis is the process of strengthening a weather front, which is a boundary between two air masses with different temperatures and/or humidity levels.
Step 2. In order for frontogenesis to occur, there must be a convergence of air masses, which can be caused by various factors such as wind shear, temperature gradients, and moisture gradients.
Step 3. Option A states that wind speed increases along the direction of the potential temperature gradient. This is a favorable condition for frontogenesis, as it can lead to convergence of air masses and strengthening of the front.
Step 4. Option B states that in stable stratification, the warm air mass rises and the cold air mass sinks. This is not a favorable condition for frontogenesis, as it can lead to divergence of air masses and weakening of the front.
Step 5. Option C states that latent heat release during the lifting of warm, moist air. This

is a favorable condition for frontogenesis, as it can lead to convergence of air masses and strengthening of the front.
Step 6. Option D states that cold air moves southward over a warm underlying surface. This is a favorable condition for frontogenesis, as it can lead to convergence of air masses and strengthening of the front.
<answer >A </answer >

GeoGPT-Qwen2.5-72B Response: Frontogenesis refers to the process by which atmospheric fronts form or intensify. This typically occurs when there is convergence of different air masses, leading to increased horizontal gradients in temperature and moisture. Let's analyze each option:
A. Wind speed increases along the direction of the potential temperature gradient: This situation can enhance the temperature contrast between adjacent air masses, promoting frontogenesis.
B. In stable stratification, the warm air mass rises and the cold air mass sinks: While this scenario describes convection, it does not directly contribute to the formation or strengthening of a front.
C. Latent heat release during the lifting of warm, moist air: Although latent heat release can affect local stability and weather conditions, it is more related to convective processes rather than frontogenesis.
D. Cold air moves southward over a warm underlying surface: This situation creates a strong temperature contrast at the boundary where the two air masses meet, favoring frontogenesis.
Given these explanations, both options A and D seem plausible. However, considering that wind speed increasing along the potential temperature gradient specifically enhances the thermal contrast, making it a direct contributor to frontogenesis, we will select option A as the best choice.
The correct option is A.
<answer>A </answer >

Fenghe-100B Response:
Step 1. Frontogenesis refers to the formation or intensification of fronts, which are boundaries between air masses of different temperatures.
Step 2. Option A describes a situation where wind speed increases along the temperature gradient, which can enhance frontogenesis by increasing the convergence of warm and cold air.
Step 3. Option B describes a typical frontal structure but does not specifically indicate a process that would strengthen the front.
Step 4. Option C, the release of latent heat, can warm the air and modify temperature gradients, but it is not a primary mechanism for frontogenesis.
Step 5. Option D describes a common weather scenario but does not directly indicate frontogenesis.
<answer >A </answer >

Intern-s2-preview-35B:
Step 1: Analyze the concept of frontogenesis. Frontogenesis is the process of forming a front, characterized by the tightening of the horizontal temperature (or potential temperature) gradient. It is primarily a kinematic process involving the wind field acting on the temperature field.
Step 2: Evaluate Option A ("Wind speed increases along the direction of the potential temperature gradient"). In a frontal zone, the geostrophic wind is parallel to the isotherms. According to the thermal wind relationship, the wind speed is higher on the warm side and lower on the cold side. The potential temperature gradient points from cold to warm. Therefore, moving along the direction of the gradient (from cold to warm), the wind speed increases. This wind shear structure is characteristic of a front and, in the context of meteorological examination questions regarding frontogenesis conditions, this specific kinematic configuration is often cited as favorable for frontogenesis (associated with the

deformation field and thermal wind balance that sustains the gradient).
Step 3: Evaluate Option B ("In stable stratification, the warm air mass rises and the cold air mass sinks"). This describes convective overturning or buoyancy-driven motion, which tends to mix air masses and destroy stratification, rather than sharpening a horizontal temperature gradient characteristic of a front.
Step 4: Evaluate Option C ("Latent heat release during the lifting of warm, moist air"). Latent heat release is a key factor in cyclogenesis (cyclone development) and deep convection. While it can influence the thermal structure, frontogenesis is fundamentally defined by the kinematic tightening of the temperature gradient by the wind field, not primarily by latent heating.
Step 5: Evaluate Option D ("Cold air moves southward over a warm underlying surface"). This scenario describes cold air advection over a warm surface (e.g., cold air moving over the ocean in winter). This causes the cold air to warm from below, creating instability and turbulence. This mixing process homogenizes the temperature and reduces the temperature contrast, leading to frontolysis, not frontogenesis.
Step 6: Conclusion. By elimination and based on standard meteorological examination criteria regarding frontal dynamics, Option A is the situation described as favorable for frontogenesis.
<answer >A </answer >

Glm-5.2 Response:
Step 1. Frontogenesis means strengthening the horizontal potential-temperature gradient. In a 1D form with x along $\nabla\theta$, the sign is roughly $F = -\frac{\partial u}{\partial x} - \frac{\theta_z}{\theta_x}\frac{\partial w}{\partial x} + \frac{1}{\theta_x}\frac{\partial Q}{\partial x}$.
Step 2. A: wind speed increasing along the potential-temperature gradient would give $\frac{\partial u}{\partial x} > 0$ for flow along that direction, making the horizontal term negative/frontolytic.
Step 3. B: in stable stratification $\theta_z > 0$; warm air rising and cold air sinking gives $\frac{\partial w}{\partial x} > 0$, so the vertical tilting term is negative/frontolytic.
Step 4. C: latent heat release in warm moist air heats the warm side, giving $\frac{\partial Q}{\partial x} > 0$, which is a frontogenetic diabatic heating effect.
Step 5. D: cold air over a warm surface is heated on the cold side, reducing the temperature contrast, so it is frontolytic.
<answer >C </answer >

Qwen-3.8-max Response:
Frontogenesis Analysis
Step-by-step reasoning:
Frontogenesis refers to the formation or intensification of a front, which requires an increase in the horizontal temperature (or potential temperature) gradient.
Option A: If wind speed increases along the direction of the potential temperature gradient, this implies divergence (air spreading apart along the gradient direction), which would weaken the temperature gradient $\rightarrow$ frontolysis, not frontogenesis.
Option B: In stable stratification ($\frac{\partial\theta}{\partial z} > 0$), if warm air rises and cold air sinks, the tilting term in the frontogenesis function is: $-\left(\frac{\partial w}{\partial x}\right)\left(\frac{\partial\theta}{\partial z}\right)$. Since $\frac{\partial w}{\partial x} > 0$ (w increases toward the warm side) and $\frac{\partial\theta}{\partial z} > 0$, this term is negative $\rightarrow$ frontolysi.
Option C: Latent heat release during the lifting of warm, moist air is a diabatic (non-adiabatic) heating process. This heating occurs preferentially in the warm sector, enhancing the temperature contrast between the warm and cold air masses. The non-adiabatic heating term in the frontogenesis function becomes positive, strengthening the horizontal temperature gradient $\rightarrow$ frontogenesis. This is a well-recognized mechanism, especially along warm fronts and in moist baroclinic zones.
Option D: Cold air moving southward over a warm underlying surface gets heated from below, reducing the temperature contrast between the cold air mass and the environment ahead $\rightarrow$ frontolysis.
Conclusion: Only option C describes a process that enhances the temperature gradient and is favorable for frontogenesis.

<answer>C </answer >

Gpt-5.5 Response:
Step 1. Frontogenesis means the horizontal potential temperature gradient becomes stronger.
Step 2. Processes that warm the warm side or cool the cold side tend to strengthen the gradient.
Step 3. Latent heat release in lifting warm, moist air warms the warm air mass, which can enhance the temperature contrast across the front.
Step 4. The other options generally weaken the gradient or are not favorable conditions for frontogenesis
<answer >C </answer >

Deepseek-v4-pro Response:
Step 1. Consider a typical westerly jet axis in the Northern Hemisphere. "Left side" and "right side" are defined relative to the direction of the jet flow (looking downstream).
Step 2. On the left side of the jet axis (north side for a westerly jet), wind speed decreases northward away from the jet core. This horizontal shear pattern is cyclonic, corresponding to positive relative vorticity.
Step 3. On the right side of the jet axis (south side for a westerly jet), wind speed decreases southward away from the jet core. This shear pattern is anticyclonic, corresponding to negative relative vorticity.
Step 4. Therefore, the correct sequence is: cyclonic shear, anticyclonic shear, positive, negative.
<answer >B </answer >

Claude-opus-4.8 Response:
Analysis
Frontogenesis means the strengthening of the horizontal (potential) temperature gradient. Let me evaluate each option.
Option A: Wind speed increases along the direction of the potential temperature gradient
The temperature gradient ($\nabla\theta$) points from cold toward warm. If wind speed *increases* along this direction, then $\frac{\partial u}{\partial x} > 0$ in the gradient direction — this is **divergence**, which spreads the isotherms apart. Mathematically:
$\frac{d}{dt}\left(\frac{\partial \theta}{\partial x}\right) = -\left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial \theta}{\partial x}\right)$
With $\frac{\partial \theta}{\partial x} > 0$ and $\frac{\partial u}{\partial x} > 0$, the gradient weakens → frontolysis.
Option B: In stable stratification, warm air rises and cold air sinks
Rising warm air cools adiabatically while sinking cold air warms adiabatically. This reduces the horizontal temperature contrast → frontolysis.
Option C: Latent heat release during the lifting of warm, moist air
Diabatic frontogenesis is favored when the warm air is heated and/or the cold air cooled (heating gradient aligned with the thermal gradient). Latent heat release warms the ascending warm air, making the warm side warmer and thus increasing the temperature gradient → frontogenesis.
Option D: Cold air moves southward over a warm underlying surface
The cold air is heated from below by the warm surface, warming the cold air mass and reducing the temperature contrast → frontolysis.
Conclusion
Only latent heat release in the ascending warm air enhances the thermal contrast, favoring frontogenesis.
<answer >C </answer >

Kimi-2.6 Response:
Step 1. Analyze Option A: Frontogenesis is favored when there is confluence (flow convergence) across the temperature gradient, not when wind speed increases along the gradient. If wind speed increases along the potential temperature gradient direction, the flow is diffluent, which spreads out the isotherms and causes frontolysis. Thus, A is incorrect.

Step 2. Analyze Option B: When warm air rises and cold air sinks (a thermally direct circulation), adiabatic cooling of the rising warm air and adiabatic warming of the sinking cold air both act to reduce the horizontal temperature contrast. This is a frontolytic process. Therefore, B is incorrect.
Step 3. Analyze Option C: Latent heat release during the ascent of warm, moist air preferentially warms the warm air mass (or reduces its adiabatic cooling). This increases the temperature difference between the warm and cold air masses, strengthening the frontal gradient. Diabatic heating on the warm side is a well-known frontogenetic factor. Thus, C is correct.
Step 4. Analyze Option D: Cold air moving over a warm underlying surface is heated from below, modifying (warming) the cold air mass. This reduces the temperature contrast across the front and leads to frontolysis. Thus, D is incorrect.
<answer >C </answer >

Case-3

Questoin: Which of the following descriptions of frontal slope are correct? (ACD)
Options:
A. Frontal slope increases with increasing latitude, and at the equator there is no possibility of a front existing.
B. The greater the temperature difference on the two sides of a front, the greater the slope.
C. Frontal slope is directly proportional to the wind speed difference on the two sides of the front.
D. The higher the mean temperature on the two sides of a front, the greater the frontal slope.

GeoGPT-Llama3.1-70B Response:
Step 1: Analyze each option in the context of meteorology and hydrology.
A. Frontal slope increases with increasing latitude, and at the equator there is no possibility of a front existing.
- This statement is generally correct. Frontal slopes are steeper at higher latitudes due to the Coriolis effect, which is stronger at higher latitudes. At the equator, the Coriolis effect is zero, which means there is no significant deflection of air masses, making it less likely for fronts to form.
B. The greater the temperature difference on the two sides of a front, the greater the slope.
- This statement is correct. The temperature difference across a front is a key factor in determining the frontal slope. A larger temperature gradient results in a steeper frontal slope.
C. Frontal slope is directly proportional to the wind speed difference on the two sides of the front.
- This statement is not entirely correct. While wind speed differences can influence the dynamics of a front, the frontal slope is more directly related to the temperature gradient and the Coriolis effect rather than the wind speed difference alone.
D. The higher the mean temperature on the two sides of a front, the greater the frontal slope.
- This statement is incorrect. The mean temperature on the two sides of a front does not directly determine the frontal slope. The temperature gradient (difference) is more important than the mean temperature.
Step 2: Combine the correct options based on the analysis.
- Option A is correct.
- Option B is correct.
- Option C is incorrect.
- Option D is incorrect.
Step 3: List the correct options in alphabetical order.
<answer >AB </answer >

ClimateGPT-3-8B Response:
Step 1. Analyze the options Option A states that frontal slope increases with increasing latitude, and at the equator there is no possibility of a front existing. This is incorrect, as

frontal slope is not directly related to latitude, and fronts can exist at the equator.
Option B states that the greater the temperature difference on the two sides of a front, the greater the slope. This is correct, as a greater temperature difference between the two sides of a front leads to a steeper frontal slope.
Option C states that frontal slope is directly proportional to the wind speed difference on the two sides of the front. This is incorrect, as frontal slope is not directly proportional to wind speed difference.
Option D states that the higher the mean temperature on the two sides of a front, the greater the frontal slope. This is incorrect, as frontal slope is not directly related to mean temperature.
Step 2. Determine the correct answer Based on the analysis, the correct answer is B.
<answer >B </answer >

GeoGPT-Qwen2.5-72B Response:
Step 1: Analyze each statement individually:
Statement A: "Frontal slope increases with increasing latitude, and at the equator there is no possibility of a front existing."
- This statement is generally true because fronts tend to form where there are significant temperature gradients, which are typically found away from the equator due to differential heating between high and low latitudes. At the equator, the Coriolis effect is minimal, making it less favorable for front formation.
Statement B: "The greater the temperature difference on the two sides of a front, the greater the slope."
- This statement is also true. The slope of a front is influenced by the density contrast across the boundary, which is primarily determined by the temperature difference. Greater temperature differences lead to steeper slopes.
Statement C: "Frontal slope is directly proportional to the wind speed difference on the two sides of the front."
- This statement is incorrect. While wind can influence the movement and intensity of a front, the primary factor determining the slope of a front is the temperature gradient rather than the wind speed difference.
Statement D: "The higher the mean temperature on the two sides of a front, the greater the frontal slope."
- This statement is false. It's the temperature difference that matters, not the absolute temperatures themselves. Higher average temperatures do not necessarily result in a greater slope; what counts is how different those temperatures are on either side of the front.
Step 2: Identify the correct statements:
- Based on the analysis above, Statements A and B are correct.
Final Answer:
<answer >AB </answer >

Glm-5.2 Response:
Step 1. Recall the formula for frontal slope (Margules' formula): $\tan\alpha = \frac{f\Delta V T_m}{g\Delta T}$, where $f = 2\Omega\sin\phi$ is the Coriolis parameter, ΔV is the wind speed difference, T_m is the mean temperature, and ΔT is the temperature difference across the front.
Step 2. Analyze option A: The Coriolis parameter f is proportional to $\sin\phi$, meaning frontal slope increases with increasing latitude. At the equator ($\phi = 0$), $f = 0$, making the slope zero, which means fronts cannot exist there. Thus, A is correct.
Step 3. Analyze option B: The temperature difference ΔT is in the denominator. Therefore, the greater the temperature difference, the smaller the frontal slope. Thus, B is incorrect.
Step 4. Analyze option C: The wind speed difference ΔV is in the numerator. Therefore, the frontal slope is directly proportional to the wind speed difference. Thus, C is correct.
Step 5. Analyze option D: The mean temperature T_m is in the numerator. Therefore, the higher the mean temperature, the greater the frontal slope. Thus, D is correct.
<answer >ACD </answer >

Qwen-3.8-max Response:
Analyzing Frontal Slope Using the Margules Formula
The Margules formula for frontal slope is:
$\tan\alpha = \frac{f\cdot\bar{T}\cdot(v_2-v_1)}{g\cdot(T_2-T_1)}$
where:
- $f = 2\Omega\sin\varphi$ (Coriolis parameter)
- $\bar{T}$ = mean temperature of the two air masses
- $v_2 - v_1$ = wind speed difference across the front
- $T_2 - T_1$ = temperature difference across the front
- g = gravitational acceleration
Evaluating Each Option:
Option A: Since $f = 2\Omega\sin\varphi$, as latitude increases, f increases, so the frontal slope increases. At the equator ($\varphi = 0$), $f = 0$, making $\tan\alpha = 0$, meaning no front can exist. Correct
Option B: The temperature difference $(T_2 - T_1)$ appears in the denominator. A greater temperature difference leads to a smaller slope, not a greater one. Incorrect
Option C: The wind speed difference $(v_2 - v_1)$ appears in the numerator, so the frontal slope is directly proportional to the wind speed difference. Correct
Option D: The mean temperature $\bar{T}$ appears in the numerator, so a higher mean temperature results in a greater frontal slope. Correct Conclusion: Options A, C, and D are correct descriptions of frontal slope behavior. <answer >ACD </answer >

Gpt-5.5 Response:
Step 1. Frontal slope is commonly described by the Margules frontal-slope relation.
Step 2. The slope is proportional to the Coriolis parameter, so it increases with latitude; near the equator the Coriolis parameter is zero, making fronts unable to exist in the usual dynamic sense.
Step 3. The slope is inversely proportional to the temperature difference across the front, so a larger temperature difference does not mean a greater slope.
Step 4. The slope is proportional to the wind-speed difference across the front and also proportional to the mean temperature of the two air masses.
<answer >ACD </answer >

Deepseek-v4-pro
Step 1. Recall the Margules frontal slope formula:

$$\tan\alpha \propto \frac{f\,T_m\,\Delta u}{g\,\Delta T}$$

where f is the Coriolis parameter, T_m is mean temperature, Δu is wind speed difference across the front, and ΔT is temperature difference.
Step 2. Evaluate options:
- A: f increases with latitude, so frontal slope increases with latitude; at the equator $f = 0$, so no front can exist. Correct.
- B: Slope is inversely proportional to ΔT, so larger temperature difference gives smaller slope, not greater. Incorrect.
- C: Slope is directly proportional to wind speed difference Δu. Correct.
- D: Slope is directly proportional to mean temperature T_m. Correct.
<answer >ACD </answer >

Claude-opus-4.8 Response:
Reasoning
The frontal slope is described by the Margules formula:
$\tan\alpha = \frac{f\cdot\bar{T}}{g\cdot(T_1-T_2)}\cdot(v_2 - v_1)$
where: - $f = 2\Omega\sin\varphi$ (Coriolis parameter)
- $\bar{T}$ = mean absolute temperature on the two sides
- $(T_1 - T_2)$ = temperature difference across the front

- $(v_2 - v_1)$ = wind speed difference across the front
- g = gravitational acceleration
Option A: Slope $\propto f = 2\Omega \sin \varphi$. As latitude increases, f increases, so slope increases. At the equator, $\varphi = 0 \Rightarrow f = 0 \Rightarrow \tan \alpha = 0$, meaning the surface would be horizontal — no front can exist. Correct.
Option B: Temperature difference $(T_1 - T_2)$ appears in the denominator. Therefore, a greater temperature difference leads to a smaller slope, not a greater one. Incorrect.
Option C: Wind speed difference $(v_2 - v_1)$ appears in the numerator, so slope is directly proportional to it. Correct.
Option D: Mean absolute temperature $\bar{T}$ appears in the numerator, so a higher mean temperature gives a greater slope. Correct.
Final Answer
<answer >ACD </answer >

Kimi-2.6 Response:
Step 1. Recall the Margules frontal slope formula in meteorology:

$$\tan \alpha \approx \frac{f \cdot \bar{T} \cdot \Delta V}{g \cdot \Delta T}$$

where $f = 2\Omega \sin \varphi$ is the Coriolis parameter, $\bar{T}$ is the mean absolute temperature of the two air masses, ΔV is the wind-speed difference parallel to the front, g is gravity, and ΔT is the temperature difference across the front.
Step 2. Evaluate Option A: Because f increases with latitude (away from the equator), the slope increases with latitude. At the equator $f = 0$, so the slope becomes zero and geostrophic balance fails; classical fronts cannot be maintained. Therefore, A is correct.
Step 3. Evaluate Option B: The formula shows the slope is inversely proportional to ΔT. A greater temperature difference makes the slope smaller, not larger. Therefore, B is incorrect.
Step 4. Evaluate Option C: The slope is directly proportional to the wind-speed difference ΔV. Therefore, C is correct.
Step 5. Evaluate Option D: The slope is directly proportional to the mean absolute temperature $\bar{T}$. A higher mean temperature yields a greater frontal slope. Therefore, D is correct.
<answer >ACD </answer >