K-MetBench: A Multi-Dimensional Benchmark for Fine-Grained Evaluation of Expert Reasoning, Locality, and Multimodality in Meteorology

Anonymous ACL submission

Abstract

The development of practical (multimodal) large language model assistants for Korean weather forecasters is hindered by the absence of a multidimensional, expert-level evaluation framework grounded in authoritative sources. To address this, we introduce K-MetBench, a diagnostic benchmark grounded in national qualification exams. It exposes critical gaps across four dimensions: expert visual reasoning of charts, logical validity via expert-verified rationales, Korean-specific geo-cultural comprehension, and fine-grained domain analysis. Our evaluation of 55 models reveals a profound *modality gap* in interpreting specialized diagrams and a *reasoning gap* where models hallucinate logic despite correct predictions. Crucially, Korean models outperform significantly larger global models in local contexts, demonstrating that parameter scaling alone cannot resolve cultural dependencies. K-MetBench serves as a roadmap for developing reliable, culturally aware expert AI agents.

1 Introduction

Large language models (LLMs) and multimodal large language models (MLLMs) have shown growing promise in scientific domains (Taylor et al., 2022; Team et al., 2023; OpenAI, 2025), achieving performance matching passing thresholds on professional certification exams (Singhal et al., 2023; Katz et al., 2024). As these models are increasingly positioned as assistants for domain-specific tasks, there is a growing need for evaluation frameworks that go beyond surface-level correctness and more precisely characterize domain-relevant competencies (Liang et al., 2022). However, existing benchmarks for vertical domains often summarize performance using a single aggregate score, making it difficult to understand why a model succeeds or fails in practice. In complex applied fields such as meteorology, this coarse evaluation obscures **critical limitations** for real-world deployment. We identify four recurring limitations in current evaluation settings for meteorological reasoning.

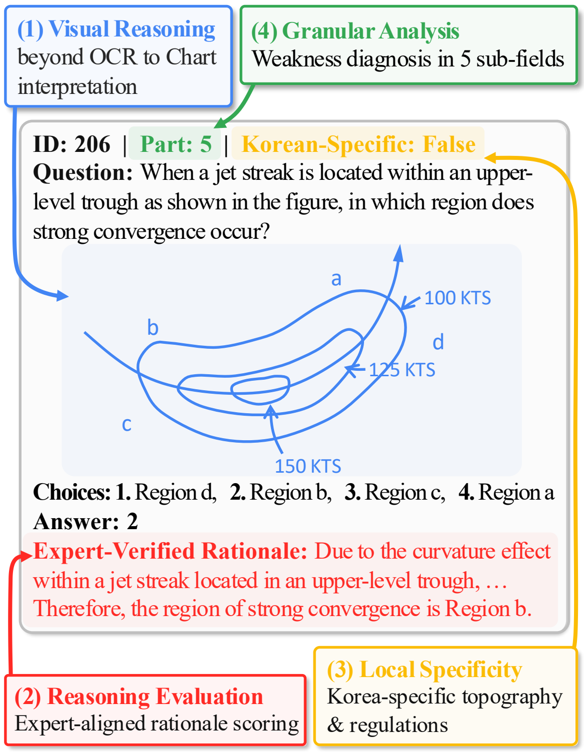


Figure 1: **An example of the K-MetBench dataset (translated into English).** K-MetBench provides evaluation across four critical dimensions: (1) multimodal understanding, (2) expert-level reasoning, (3) geo-cultural context sensitivity, and (4) fine-grained domain knowledge across five meteorological sub-fields.

First, **the modality gap**. Meteorological analysis, inherently multimodal, requires the synthesis of numerical data, textual descriptions, and specialized visual charts (e.g., weather maps, skew-T log-P diagrams). However, most scientific benchmarks remain predominantly text-based and provide limited assessment of a model' s ability to interpret domain-specific charts and spatial patterns. As a result, visual understanding capabilities central to operational forecasting remain under-evaluated.

Second, **the reasoning gap**. Conventional benchmarks primarily rely on answer accuracy, without explicitly evaluating the validity or structure of the underlying reasoning. In high-stakes domains like weather forecasting, a correct prediction reached through shallow heuristics or incomplete logic may still lead to brittle or unreliable behavior (Turpin et al., 2023). Without access to expert-aligned rationales, it is difficult to distinguish genuine understanding from shortcut learning.

Third, **the geo-cultural gap**. Many existing datasets emphasize global or universal physical principles while abstracting away local geographic and institutional context. In meteorology, however, local topography, climatological conventions, and region-specific regulations play a substantial role in interpretation and decision-making. Models trained and evaluated solely on decontextualized data may therefore fail to generalize reliably to region-specific applications.

Fourth, **the granularity gap**. Aggregate performance scores often mask uneven competence across sub-domains. A model may perform well on factual recall or chart interpretation while struggling with quantitative reasoning or applied dynamics. Without fine-grained analysis, such disparities remain difficult to diagnose.

To address these limitations, we introduce **K-MetBench**, a Korean meteorological benchmark designed for multi-dimensional evaluation of LLMs and MLLMs. Rather than treating meteorological expertise as a monolithic capability, K-MetBench decomposes evaluation along four complementary axes: **(1) multimodal understanding** of meteorological charts and symbols, **(2) reasoning quality** assessed using expert-verified rationales, **(3) sensitivity to geo-cultural and regional context**, and **(4) fine-grained coverage** across five officially defined meteorological sub-domains. Through this structured design, K-MetBench is intended as a diagnostic tool that helps reveal which aspects of meteorological reasoning remain challenging for current models, and why.

2 Related Work

Existing benchmarks for meteorological and climate reasoning reflect diverse assumptions about knowledge sources, modalities, and evaluation objectives. Rather than treating them as competitors, we situate them along complementary axes that highlight different aspects of domain expertise.

ClimaQA (Manivannan et al., 2024) evaluates climate question answering using textbooks as the knowledge source. By grounding questions in established instructional materials, it emphasizes conceptual understanding and theoretical reasoning characteristic of graduate-level climate science. While this approach provides scientific rigor, it remains purely text-based and does not assess visual interpretation or operational reasoning grounded in real-world artifacts. ClimateIQA (Chen et al., 2025) constructs instruction-style QA data from numerical weather prediction (NWP) heatmaps and associated geospatial metadata. This enables evaluation of visual pattern recognition and structured data interpretation. WeatherQA (Ma et al., 2024) further targets operational forecasting scenarios by combining multiple meteorological images with expert-written mesoscale discussions. These datasets advance multimodal evaluation, but their emphasis remains on task-level performance rather than fine-grained diagnosis across sub-fields or distinct reasoning failures.

In the Korean-language evaluation landscape, KMMLU (Son et al., 2025) is derived from official national examinations, measuring expert-level linguistic competence across a wide range of professions. Since it is based on official Korean exams, KMMLU captures linguistic and cultural aspects of the Korean language. KMMLU-Redux (Hong et al., 2025) is a reconstructed version of KMMLU that removes erroneous, ambiguous, or contaminated items to improve reliability. While these benchmarks offer high reliability and clear passing criteria, they are primarily text-based and treat meteorological knowledge as a small subset within a broader evaluation suite, limiting their ability to analyze domain-specific competencies in depth.

3 K-MetBench Construction

K-MetBench is designed to complement the existing benchmarks by explicitly separating and jointly examining four dimensions that are often conflated in prior benchmarks. Rather than introducing new task formats, K-MetBench focuses on providing diagnostic visibility into how and where current models succeed or fail when approaching expert-level meteorological reasoning.

Table 1: **Comparison with existing benchmarks.** K-MetBench distinguishes itself by covering four key axes: visual understanding, rationale reliability, geo-cultural alignment, and fine-grained diagnosis in sub-domains.

Dataset	Lang.	Domain	Modality	Reasoning	Geo-Cultural	Granularity
KMMLU (Son et al., 2025)	Kor	General(Expert-level)	Text	×	Korea	45 Domains
KMMLU-Redux (Hong et al., 2025)	Kor	General(Expert-level)	Text	×	Korea	14 Domains
ClimaQA (Manivannan et al., 2024)	Eng	Meteorology	Text	×	Global	3 Tasks
ClimateIQA (Chen et al., 2025)	Eng	Meteorology	Image + Text	×	Global	4 Tasks
WeatherQA (Ma et al., 2024)	Eng	Meteorology	Image + Text	×	United States	2 Tasks
K-MetBench (Ours)	**Kor**	**Meteorology**	**Image + Text**	**Expert-Verified**	**Korea**	**5 Sub-Domains**

Note for K-MetBench. *Modality:* Includes multimodal questions that evaluate interpreting weather charts and diagrams. *Reasoning:* Provides rationale verified by domain experts. *Geo-Cultural:* Includes questions requiring knowledge of local geography and regulations that are specific to Korea (e.g., Yeongdong region and the Korea Meteorological Administration (KMA) protocols). *Granularity:* Supports fine-grained diagnosis across the five sub-domains officially defined in Korea Engineer Meteorology certification exam.

Table 2: **Detailed statistics of K-MetBench.** The dataset is structured into four key dimensions to enable structured evaluation: Modality, Reasoning, Geo-cultural, and Granularity.

Diagnostic Axis	Statistic	Value
1. Overview	Total Questions	1,774
2. Modality	Image + Text Questions	**82 (4.62%)**
(*Visual Understanding*)	(*Charts, Diagrams*)	
3. Reasoning	Avg. Rationale Length	**93.72 tokens**
(*Expert Rationales*)	Text-Only Reasoning	121 (6.82%)
	Multimodal Reasoning	20 (1.13%)
4. Geo-Cultural	Korean-Specific	**73 (4.11%)**
(*Local Knowledge*)	Questions	
5. Granularity	Part 1: Forecast Theory	373 (21.03%)
(*5 Subject Areas*)	Part 2: Observation	332 (18.71%)
	Part 3: Atmos. Dynamics	359 (20.24%)
	Part 4: Climatology	376 (21.20%)
	Part 5: Atmos. Physics	334 (18.83%)

Note: The number of tokens is calculated using the `gemini-2.5-flash` tokenizer. (Atmos.: Atmospheric)

3.1 Data Collection and Processing

K-MetBench is constructed from raw data drawn from the National Meteorological Engineer certification examinations, covering 25 exam sessions between March 16, 2003 and March 5, 2022. The initial pool comprised 2,500 multiple-choice questions. Because these examinations are generated from a shared question bank, substantial overlap exists across years. To construct a balanced and non-redundant benchmark, we applied a multi-stage filtering and augmentation pipeline.

We first performed automated deduplication using `difflib.SequenceMatcher` with a similarity threshold of 0.6, removing exact duplicates as well as items with trivially permuted answer options. Importantly, questions with inverted logic (e.g., 'highest' vs. 'lowest', 'saturated' vs. 'unsaturated') were manually reviewed and retained, as they probe distinct reasoning behaviors despite surface similarity. This process yielded a refined set of 1,774 questions.

To reduce memorization and contamination effects, we applied two transformations. First, we randomized answer option orders for all questions. Second, we paraphrased question stems using `Gemini-2.5-Pro`, with strict constraints to preserve technical terminology and domain-specific meaning. The system prompt used for paraphrasing is provided in Appendix C.1. To maintain quality, a human researcher reviewed and refined 14.88% (264/1,774) of the paraphrased items.

For multimodal questions, both text and visual elements were extracted from the original examination PDFs. Three researchers reviewed and cross-checked all extracted images to correct parsing artifacts such as missing axis labels, distorted symbols, or incomplete annotations. As a design choice to separate perceptual challenges from reasoning difficulty, mathematical formulas embedded as images were transcribed into LaTeX code to prevent OCR bottlenecks, while meteorological charts and diagrams were preserved in their original format.

3.2 Subset 1: Multimodal Diagnosis

The multimodal subset of K-MetBench consists of 82 questions (4.62% of the dataset) that require interpretation of meteorological visuals. Unlike general-purpose multimodal benchmarks that focus on object recognition or scene description, this subset targets domain-specific charts and symbolic representations. The included materials span surface weather maps, upper-level charts (e.g., 200 and 500 hPa), and thermodynamic diagrams such

Table 3: **Distribution of K-MetBench across five sub-domains.** The number of questions for each subject area is reported, with the number of reasoning questions featuring expert-verified rationales in parentheses (Reas. stands for Reasoning).

Part	Subject Area	Overall Volume	Modality		Geo-Cultural
		Total (Reas.)	Text (Reas.)	Image + Text (Reas.)	Korean (Reas.)
1	Weather Analysis & Forecast Theory	373 (28)	364 (24)	9 (4)	6 (0)
2	Meteorological Observation Methods	332 (28)	318 (24)	14 (4)	0 (0)
3	Atmospheric Dynamics	359 (29)	340 (25)	19 (4)	0 (0)
4	Climatology	376 (28)	363 (24)	13 (4)	50 (7)
5	Atmospheric Physics	334 (28)	307 (24)	27 (4)	17 (0)
Sum	**Total Coverage**	**1,774 (141)**	**1,692 (121)**	**82 (20)**	**73 (7)**

as Skew-T Log-P plots and emagrams derived from radiosonde measurements. Solving these questions requires extracting structured information, including pressure gradients, wind vectors, and thermodynamic indices—from dense visual fields that cannot be resolved through OCR alone. Consequently, this subset assesses the ability of MLLMs to integrate textual meteorological knowledge with the interpretation of domain-specific visual cues. Representative examples are provided in Appendix Table 6.

3.3 Subset 2: Reasoning-Aware Evaluation

To evaluate reasoning quality beyond final answer correctness, K-MetBench includes a reasoning-aware subset consisting of 141 questions paired with expert-verified rationales. These rationales serve as reference explanations for assessing the validity, coherence, and depth of model-generated reasoning. Rationale construction followed a two-stage process. First, `GPT-5` was used to generate initial reasoning drafts, guided by prompts that emphasized logical flow, factual consistency, clarity, and completeness. Second, two meteorology professors reviewed these drafts, correcting factual errors, refining physical explanations, and resolving ambiguities. We employ an LLM-as-a-Judge framework (Zheng et al., 2023) to score model-generated rationales against the expert-verified rationales as reference standard. The system prompts used for reasoning generation and evaluation are detailed in Appendix C.5 and C.6. To validate the reliability of this framework in a specialized domain, we conduct a meta-evaluation (Li et al., 2024) comparing LLM judgments with human expert scores. The experimental and survey protocols are provided in Appendix D.1 and C.9.

3.4 Subset 3: Geo-Cultural Sensitivity

Meteorological reasoning is strongly influenced by local geography, climate patterns, and institutional conventions. To capture this dependency, we annotate a *Korean-Specific* subset comprising 73 questions that involve implicit, speaker-centric, or high-context expressions specific to the Korean Peninsula. Candidate items were identified using prompt-enhanced LLMs (`GPT-4.1` and `Gemini-2.5-Pro`) designed to detect references to localized phenomena, such as regional topography (e.g., the Yeongdong region) or regulations issued by the Korea Meteorological Administration (KMA). These candidates were subsequently reviewed and validated by two researchers to ensure relevance and correctness. Rather than testing translation ability, this subset probes whether models can appropriately ground meteorological knowledge in region-specific context. As such, it provides a controlled setting for analyzing geo-cultural alignment in domain-specific reasoning.

3.5 Subset 4: Domain Specificity

To enable fine-grained analysis of meteorological expertise, K-MetBench is organized into five official subject areas defined in the Korean Meteorological Engineer certification exam. These include: Part 1 (Weather Analysis and Forecast Theory), Part 2 (Meteorological Observation Methods), Part 3 (Atmospheric Dynamics), Part 4 (Climatology), and Part 5 (Atmospheric Physics). Each subject area targets a distinct aspect of professional competence, ranging from chart interpretation and numerical weather prediction principles to instrumentation, large-scale atmospheric motion, climate systems, and thermodynamic calculations. This structure allows model performance to be examined at a level of granularity that is not visible from aggregate scores alone. By aligning evaluation with established subject boundaries, this design facilitates diagnosis of domain-specific strengths and weaknesses, for example, distinguishing models that perform well on descriptive climatology but struggle with quantitative dynamics or thermodynamics.

4 Experiments

4.1 Experimental Setup

Evaluated Models. To ensure a comprehensive benchmark, we evaluated a diverse array of models categorized by scale, training language, and modality support. The selection includes proprietary state-of-the-art models renowned for superior reasoning capabilities, such as `GPT-5.2` (evaluated with and without reasoning modules enabled) (OpenAI, 2025) and `Gemini-3-Pro-Preview` (Team et al., 2023). We also incorporated open-source models ranging from 0.6B to 235B parameters, exemplified by `InternVL3.5` (Wang et al., 2025) and `Qwen3-VL` (Yang et al., 2025), alongside large-scale foundation models such as `gpt-oss-120b` (Agarwal et al., 2025), `command-a-reasoning-08-2025` (Cohere et al., 2025), and `Llama-3.2-90B-Vision-Instruct` (Meta, 2024). To investigate the impact of geo-cultural knowledge, we specifically included Korean-centric models, including `EXAONE-4.0` (Research et al., 2025), `A.X-4.0` (Lab, 2025), `VARCO-Vision-2.0` (Cha et al., 2025), and `HyperCLOVA X` (Yoo et al., 2024). Finally, strictly text-based baselines were established by evaluating non-multimodal models solely on the textual components of questions to quantify text dependency.

Geo-Cultural Disambiguation Protocol. To establish a fair evaluation protocol for global models, we designed four experimental configurations that cross-reference question formulation with prompting conditions. This setup ensures that models are assessed on their meteorological competence rather than their ability to decode localized linguistic ambiguities. For question formulation, we compared an *Implicit* condition, using original speaker-centric terms like 'Our country,' against an *Explicit* condition, which replaces these with proper nouns (e.g., 'South Korea') to isolate and evaluate pure domain knowledge. Regarding prompting conditions, beyond a *Standard* prompt that injects an expert persona, we introduced an *Advanced* prompt providing explicit disambiguation (e.g., " 'Our country' refers to South Korea"). This advanced protocol serves as a specialized support layer, mitigating performance degradation caused by implicit geo-cultural references and enabling global models to compete on an equal footing.

Comparison with Existing Benchmarks. We evaluated models using the official test sets of all datasets, employing the Chain-of-Thought (CoT) (Wei et al., 2022) protocol for WeatherQA. Task orthogonality was analyzed using Kendall's Tau-b rank correlation coefficient. To align the distance-based Haversine metric of ClimaIQA (where lower is better) with standard accuracy metrics, we inverted the sign of ClimaIQA scores prior to calculating correlations.

Meta-Evaluation Setup: Validating LLM-as-a-Judge. Given the specialized nature of meteorology, validating the reliability of commercial LLMs as judges is crucial. We conducted a meta-evaluation comparing human expert judgments with LLM judgments. We selected ten representative questions varying in difficulty and type, and collected reasoning outputs from ten open-source LLMs. Two human experts provided gold standard scores, while `Gemini-2.5-Pro` served as the AI evaluator. Both parties utilized identical expert-verified references and a scoring rubric across four axes: Factuality, Logicality, Depth, and Clarity. We calculated Kendall's Tau-b (τ_b) correlation between human and AI scores, confirming the alignment of the automated judge ($\tau_b > 0.8$). We also computed Krippendorff's α (interval) and Intraclass Correlation Coefficient (ICC, 2-way mixed, absolute) to assess inter-rater reliability, which indicated acceptable agreement ($\alpha > 0.7$). To investigate whether incorporating human expert rationales improves the alignment between the LLM evaluator and human judgment, we compared the correlations of their scores under conditions with and without rationale availability.

Implementation Details. To ensure a fair comparison, we utilized *Standard* prompts across all models. We applied a zero-shot setting to all text, multimodal, and reasoning questions to evaluate intrinsic capabilities. We computed accuracy by extracting final answers via regular expressions. To rigorously assess instruction-following capabilities, we counted any output that violated

Table 4: **K-MetBench performance scores across diverse models.** Models are sorted by accuracy. Accuracy score ranges from 0 to 100, while the reasoning score (**Reas.**) ranges from 4 to 20. The highest scores in each column are shown in **bold** for proprietary and open-source models, respectively. (Acc.: Accuracy, K: Korean model, V: Vision language model, R: Reasoning model, `Inst.`: Instruct, `Vis.`: Vision.)

Type	**Model**	**Flags**			**Acc.**	**Reas.**	**Geo-Cult.**	**Modality**		**Granularity (P1–P5)**				
		K	V	R			Korean	Text	Multi	P1	P2	P3	P4	P5
Proprietary	`gemini-3-pro-preview` (Thinking)		V	R	**93.7**	**18.01**	**90.4**	**94.6**	**75.6**	**92.5**	**97.9**	**94.2**	**92.8**	**91.6**
	`gpt-5.2` (Thinking)		V	R	87.8	17.33	80.8	90.6	29.3	86.3	93.4	88.0	86.2	85.3
	`gpt-5.2`		V		77.6	17.39	75.3	79.0	50.0	77.2	81.3	71.9	81.4	76.3
	Multilingual Models													
Open-source	`Qwen3-VL-235B-A22B-Thinking`		V	R	**84.4**	**17.22**	72.6	**86.2**	48.8	**81.5**	**88.6**	**87.2**	**83.2**	**82.0**
	`Qwen3-VL-32B-Thinking`		V	R	78.6	16.17	60.3	79.9	**51.2**	74.3	85.2	78.8	78.7	76.3
	`command-a-reasoning-08-2025`			R	77.8	14.12	74.6	77.8	-	73.4	85.2	73.8	78.8	78.5
	`gpt-oss-120b`			R	77.3	16.12	62.0	77.3	-	72.5	85.8	76.5	77.4	74.9
	`Qwen3-30B-A3B-Thinking-2507`			R	76.7	15.76	67.6	76.7	-	75.5	82.1	75.6	74.9	75.9
	`InternVL3.5-38B-Inst.`		V	R	57.3	11.38	47.9	58.1	40.2	56.0	64.8	48.7	61.4	55.7
	`Llama-3.2-90B-Vis.-Inst.`		V		56.9	9.72	52.1	58.2	30.5	57.1	59.3	52.4	62.2	53.3
	`Phi-4`				51.5	11.75	40.8	51.5	-	52.5	53.8	50.0	55.1	45.3
	Korean Models													
	`A.X-4.0`	K			76.1	15.46	**78.9**	76.1	-	76.6	77.7	68.2	81.3	76.5
	`EXAONE-4.0-32B`	K		R	59.9	13.57	59.2	59.9	-	58.2	64.8	52.4	63.1	61.2
	`VARCO-Vis.-2.0-14B`	K	V		58.7	11.24	57.5	59.5	42.7	59.0	62.3	54.3	61.7	56.0
	`A.X-4.0-Light`	K			55.7	11.45	60.6	55.7	-	55.8	54.4	50.9	61.4	55.7
	`A.X-4.0-VL-Light`	K	V		52.5	9.76	54.8	53.0	42.7	51.5	50.6	50.1	58.0	52.1
	`HyperCLOVAX-SEED-Think-14B`	K		R	50.8	11.29	52.1	50.8	-	51.6	53.8	41.8	55.6	51.1

the required format as a failure case. We employed the vLLM library (Kwon et al., 2023) with its default configurations, except for `A.X-4.0-VL-Light` and `Llama-3.2-90B-Vision-Instruct`, which were run using Hugging Face Transformers. The random seed was fixed at 42, and sampling temperatures were set to 0.1 by default, while a temperature of 1.0 was employed for reasoning models.

5 Results

Beyond simple leaderboards, we dissect the performance of models across four dimensions to reveal their true capabilities and limitations.

The Modality Gap: Text-Only vs. Multimodal. Figure 2 reveals a distinct *dented* shape along the *Multimodal* axis, confirming that visual reasoning is the primary bottleneck for current MLLMs. Specifically, models exhibited a sharp accuracy decline (avg. -16.39%p) on multimodal questions compared to text-only ones. This deficit is most pronounced in professional tasks involving Skew-T Log-P diagrams and surface weather maps, where models failed to extract key data despite their general vision capabilities.

The Reasoning Gap: Knowledge vs. Reasoning. Table 13 and Figure 2 highlight a distinct gap between answer accuracy and reasoning quality. Although Kendall's τ_b (0.78) indicates a general correlation (Appendix Figure 25), qualitative analysis reveals that models frequently provide correct answers accompanied by insufficient rationales, including the use of improper or hallucinated terminology (Appendix Table 7). Additionally, while models achieve high accuracy on simple retrieval tasks, performance significantly degrades on calculation and multi-step reasoning tasks, even when CoT prompting—explicitly guiding the model to use a <scratchpad>—is employed.

The Geo-Cultural Gap. Table 13 reveals that large multilingual models struggle with the *Korean-Specific* subset (e.g., Changma, topography) despite their scale. The Korean-centric `A.X-4.0` (72B) scored 78.9, outperforming the larger `Qwen3-VL-235B-Thinking` (72.6). This confirms that parameter scaling does not automatically grant proficiency in local domains.

Granular Domain Analysis. Finally, decomposing performance across the five official subject areas reveals fine-grained disparities masked by aggregated scores. As shown in Table 13, models generally exhibit robust performance in Part 2 (*Meteorological Observation*), which focuses on instrumentation and factual knowledge

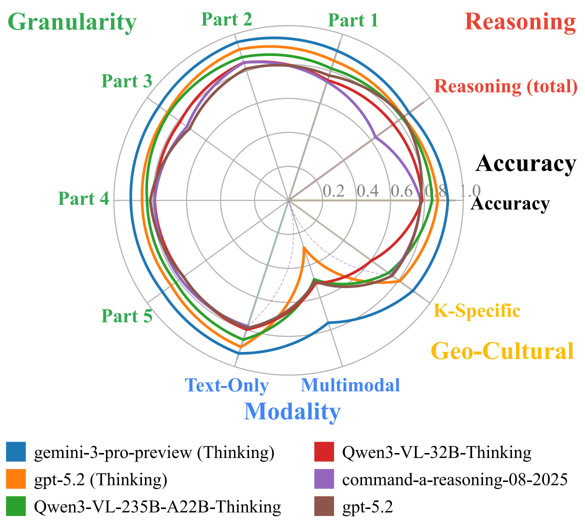


Figure 2: **Holistic performance analysis of top-6 models across five dimensions.** The radar chart visualizes model capabilities in Accuracy, Reasoning, Geo-Cultural alignment (K-Specific), Modality (Text-only vs. Multimodal), and Granularity (Subject Parts 1–5). While models show balanced performance across theoretical subjects, a sharp decline is observed in the *Multimodal* axis, highlighting the modality gap.

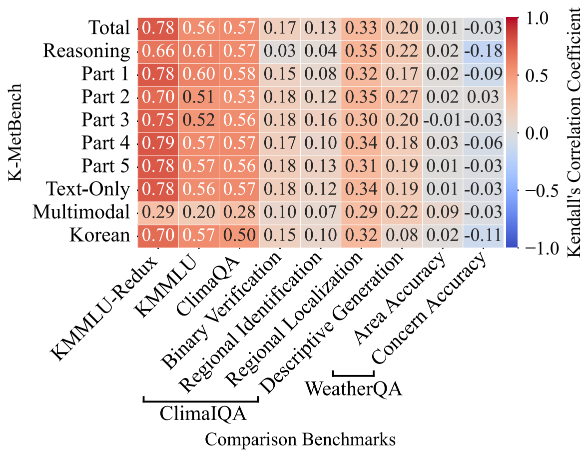


Figure 3: **Correlation analysis with existing benchmarks.** The heatmap visualizes Kendall's τ_b correlation coefficients between K-MetBench metrics and existing benchmarks.

(e.g., `Gemini-3-Pro` reaching 97.9). However, significant performance drops are observed in calculation-intensive and abstract domains like Part 3 (*Atmospheric Dynamics*) and Part 5 (*Atmospheric Physics*). A striking example is the Korean model `A.X-4.0`, which achieves its highest accuracy in Part 4 (*Climatology*) (81.3)—likely benefiting from training on local meteorological laws—but struggles disproportionately in Part 3 (68.2), where understanding synoptic motions is required. This granular diagnosis identifies specific domain weaknesses: while models may possess sufficient regulatory knowledge (Part 4), they require targeted fine-tuning to enhance quantitative reasoning in thermodynamics and dynamics (Part 3, 5).

Orthogonality between Existing Baselines. As shown in Figure 3, we analyzed Kendall' s τ_b correlations to assess the independence of K-MetBench. While the *Text-Only* subset correlates strongly with general Korean benchmarks (KMMLU-Redux, $\tau_b = 0.78$), we observe a distinct decoupling in complex capabilities. Notably, the correlation weakens for the *Reasoning* subset ($\tau_b = 0.66$) and drops sharply for the *Multimodal* subset ($\tau_b = 0.29$). Furthermore, correlations with external weather baselines (e.g., ClimaQA, ClimaIQA, and WeatherQA) remain consistently low across both reasoning and multimodal dimensions (avg. $\tau_b < 0.14$). This quantitative gap demonstrates that K-MetBench evaluates specialized domain logic and visual interpretation skills that are orthogonal to general linguistic proficiency and existing meteorological tasks.

Meta-Evaluation: Human-LLM Agreement. We validated our reasoning evaluation framework by measuring inter-rater agreement on 100 sampled responses (Table 5). All axes surpassed the reliability threshold ($\alpha > 0.7$), with *Reasoning Total* achieving a robust α of 0.838. Additionally, Figure 4 illustrates a strong correlation between human and LLM scores. The *w/ rationale* setting yielded a Kendall's τ_b of 0.99 with low variance, slightly outperforming the *w/o rationale* setting ($\tau_b = 0.96$).

Table 5: **Inter-rater agreement analysis.** The agreement between the average scores of two human experts and the LLM evaluator. We report Krippendorff's α (interval) and Intraclass Correlation Coefficient (ICC, two-way mixed, absolute agreement).

Evaluation Axis	**Krippendorff's α**	**ICC**	N
Factuality	0.827	0.829	100
Logicality	0.827	0.830	100
Depth	0.742	0.747	100
Clarity	0.825	0.827	100
Reasoning Total	0.838	0.841	100

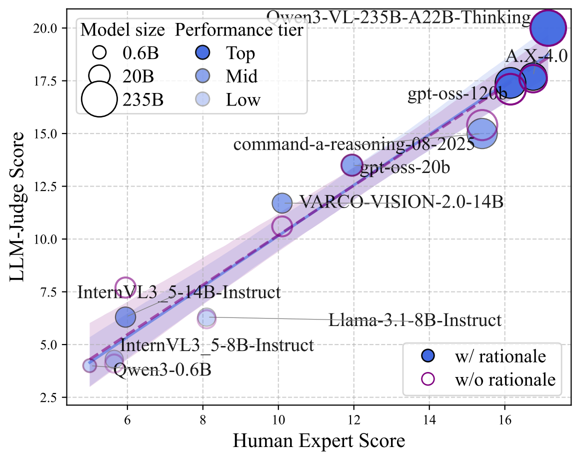


Figure 4: **Scatter plot comparing human expert vs. LLM-judge scores.** The *w/ rationale* condition ($\tau_b = 0.99$) shows slightly higher precision and lower variance than the *w/o rationale* condition ($\tau_b = 0.96$), while both maintain a strong correlation.

6 Discussion

6.1 The Challenge of Visual Reasoning in Specialized Domains

The observed modality gap in Table 13 and Figure 2 underscores a fundamental limitation: current MLLMs lack the *domain-specific visual literacy* necessary for forecasting. Although proficient in general recognition, models struggle to ground specialized visual patterns—such as isobars, fronts, and wind barbs—in physical principles. This indicates that training on general image-text pairs is insufficient for mastering the fine-grained visual reasoning required in specialized scientific domains.

6.2 Geo-Cultural Alignment in Meteorology

Meteorology requires applying universal laws to localized contexts. The observed performance gap indicates a critical lack of *geo-cultural alignment* in global models. Despite linguistic fluency, multilingual models frequently hallucinate on specific Korean geographic and terminological nuances. Consequently, effective deployment in vertical domains demands more than mere scaling; it requires rigorous alignment with local topographic and legal contexts to bridge the gap between general capability and expert-level application.

6.3 Superficial Reasoning vs. Causal Deduction

The observation that models output correct answers with shallow or erroneous explanations points to *shortcut learning* (Geirhos et al., 2020)—a reliance on surface-level associations rather than genuine understanding. Furthermore, the inability to reach expert-level performance on formula-based problems (e.g., calculating geostrophic wind speed) highlights a critical deficiency in applying physical laws. Addressing this requires shifting from general instruction tuning to training on high-quality reasoning trace data grounded in rigorous physical principles.

6.4 Reliability of Automated Evaluation in Specialized Domains

Our results confirm that `Gemini-2.5-Pro` is a reliable proxy for human experts in meteorology. The high agreement in *Factuality* and *Logicality* in Table 5 demonstrates objective evaluation of logic and evidence. While *Depth* showed slightly more subjectivity, the overall consistency supports the framework's robustness. Furthermore, the tight correlation observed in the scatter plot in Figure 4 indicates that expert rationales effectively minimize variance. However, the model's high intrinsic knowledge ensures reliable grading even in their absence. These findings demonstrate that, when guided by high-quality rubrics, modern LLMs are cost-effective and reliable judges even in fine-grained domains like meteorology. This validates adopting the LLM-as-a-Judge framework for the reasoning evaluation in this study.

7 Conclusion

We present K-MetBench, a multi-dimensional benchmark for fine-grained evaluation of large language models in meteorological reasoning. By decomposing performance across modality, reasoning quality, geo-cultural context, and domain-specific sub-fields, K-MetBench provides diagnostic insights that are not observable from aggregate accuracy alone. Our evaluation reveals persistent challenges in interpreting domain-specific visual artifacts, producing coherent expert-level rationales, and grounding meteorological knowledge in local context. In addition, analysis across official subject areas exposes uneven performance that is obscured by holistic scores. Overall, K-MetBench is intended as a diagnostic complement to existing benchmarks, helping identify where current models succeed and where targeted improvements are needed for reliable deployment in specialized scientific domains.

8 Limitations

While K-MetBench serves as a rigorous diagnostic tool for meteorological AI, we acknowledge several limitations. First, regarding modality, the benchmark focuses on static visual reasoning (e.g., snapshot weather charts). While interpreting these charts is fundamental to forecasting, the current dataset does not evaluate the temporal reasoning required to interpret atmospheric evolution, such as sequential radar imagery or satellite loops. Third, the dataset is geo-specifically rooted in the Korean context. Although this design effectively evaluates geo-cultural alignment—a key contribution of our work—it inherently limits direct generalizability to other climatic regions without adaptation. Finally, we utilized the official examination passing criteria (60%) as a proxy for human competency. While this provides a validated baseline for qualification, a fine-grained human expert ceiling (e.g., the upper-bound score of top-tier meteorologists) was not explicitly measured in this study. Future work will focus on establishing this upper bound to quantify the 'super-human' gap precisely.

9 Ethical Considerations

We strictly adhere to copyright laws and ethical guidelines throughout the dataset construction process. The questions in K-MetBench are derived from the National Meteorological Engineer examinations administered between March 16, 2003, and March 5, 2022. Among the 43 total sessions during this period, we utilized data exclusively from the 18 sessions that were officially released to the public. We have obtained explicit permission from the Human Resources Development Service of Korea (HRDK) to use these materials for research purposes and to publish the refined dataset in an open repository.

Furthermore, we ensured data safety and fair labor practices. The dataset has been thoroughly reviewed to ensure it contains no personally identifiable information or harmful content. Regarding human involvement in the annotation process, the two professor-level professional meteorologists who participated in the rationale verification and cross-validation were compensated with fair wages commensurate with their expertise and the time they invested.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Young-rok Cha, Jeongho Ju, SunYoung Park, Jong-Hyeon Lee, Younghyun Yu, and Youngjune Kim. 2025. Varco-vision-2.0 technical report. *arXiv preprint arXiv:2509.10105*.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. 2025. 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*, pages 5322–5333.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. 2025. From KMMLU-redux to pro: A professional Korean benchmark suite for LLM evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19067–19096, Suzhou, China. Association for Computational Linguistics.

Human Resources Development Service of Korea. 2024. Examination standards for meteorological engineer (2023.1.1–2026.12.31). `https://www.q-net.or.kr/pageLink.do?link=cst/cstReport`. Accessed: 2026-01-06. Available at Q-Net.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

SKT AI Model Lab. 2025. A.X 4.0.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2024. Climaqa: An automated evaluation framework for climate question answering models. *arXiv preprint arXiv:2410.16701*.

Meta. 2024. Llama 3.2 model card. Accessed: 2024-01-04.

OpenAI. 2025. Update to gpt-5 system card: Gpt-5.2. Accessed: 2026-01-04.

A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint*.

LG Research, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Kyubeen Han, Seokhee Hong, Junwon Hwang, Taewan Hwang, and 1 others. 2025. Exaone 4.0: Unified large language models integrating non-reasoning and reasoning modes. *arXiv preprint arXiv:2507.11407*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring massive multitask language understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (*Volume 1: Long Papers*), pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, and 1 others. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

Table of Contents

A Dataset Examples **14**

B Case Study of Reasoning Answer **14**

C Prompts and Questionnaires for Benchmark Construction **14**

C.1 Question Paraphrasing . . . 14
C.2 Identification of Korean-Specific Subset . . . 14
C.3 Implicit vs. Explicit Dataset Design for Korean-Specific Subset . . . 14
C.4 Evaluation Prompts for Korean-Specific Subset . . . 15
C.5 Prompt for Reference Rationale Generation . . . 15
C.6 Questionnaire for Expert Verification on Reference Rationale . . . 15
C.7 Reasoning Prompt for Open-Source LLMs . . . 16
C.8 Prompt for LLM-as-a-Judge Evaluation . . . 16
C.9 Questionnaire for Expert Scoring of LLM Reasoning . . . 16

D Experimental Setups **16**

D.1 Meta-Evaluation for LLM-as-a-Judge . . . 16

E Additional Results and Discussion **29**

E.1 Detailed Orthogonality Analysis of K-MetBench . . . 29
E.2 Detailed Analysis of K-MetBench Performance . . . 29
E.3 Results of Meta Evaluation of LLM-as-a-Judge . . . 30
E.4 MCQA Accuracy vs. Reasoning Score . . . 30
E.5 Computational Cost and Efficiency Analysis . . . 31

F Compute Resources **31**

G Hierarchical Topic Distribution **31**

Table 6: **Representative examples of K-MetBench tasks.** The examples are organized by modality: **Text-only** (Top) and **Multimodal** (Bottom). We showcase three task types within each modality: **Standard** (fundamental knowledge), **K-Specific** (geo-cultural context), and **Reasoning** (complex deduction). **Part** denotes the corresponding subject from the five official fields. *Gray text* indicates English translations.

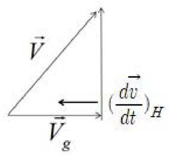
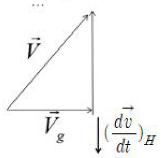
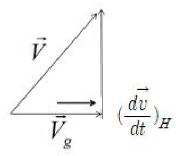
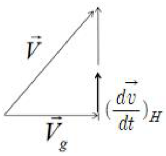
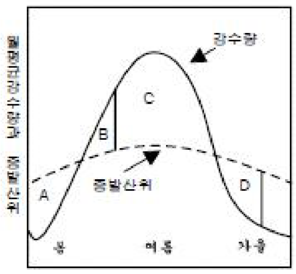
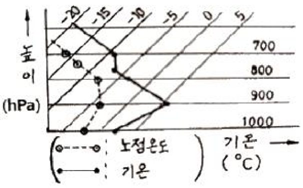
Modality	Standard MCQA	K-Specific MCQA	Reasoning MCQA
Text-Only	**ID:** 1535, **Part:** 5 **질문:** 상층 일기도의 활용에 대해 올바르게 설명한 것은? *Question: Which of the following is a correct description regarding the utilization of upper-level weather charts?* **1.** 500 hPa 일기도의 한랭기압골에서 등온선의 진폭이 등고선의 진폭보다 클 경우에는 그 기압골의 후방에 약한 상승기류가 있고, 전방에 약한 하강기류가 있다. **2.** 300 hPa 면에서는 온도가 지형, 복사의 영향을 받으므로 전선분석이 용이하다. **3.** 300 hPa 제트기류 출구의 좌측에 하강기류, 우측에 상승기류가 있으며, 입구에서는 좌측에 상승기류, 우측에 하강기류가 있다. **4.** 500 hPa 기류가 지상 한랭전선에 수직으로 불면 이 전선은 활성으로서 악천이 나타난다. *1. In a cold trough on a 500 hPa chart, if the amplitude of the isotherms is larger than the amplitude of the contours (geopotential height), there is a weak updraft behind the trough and a weak downdraft ahead of it.* *2. On the 300 hPa surface, temperature is affected by topography and radiation, making frontal analysis easy.* *3. At the exit of a 300 hPa jet stream, there is a downdraft on the left and an updraft on the right; at the entrance, there is an updraft on the left and a downdraft on the right.* *4. If the 500 hPa airflow blows perpendicular to a surface cold front, the front becomes active and severe weather occurs.* **정답:** 1 *Ground Truth: 1*	**ID:** 65, **Part:** 5 **질문:** 다음은 한국 지역에 영향을 주는 고기압의 특성을 설명한 것이다. 내용이 옳지 않은 것은? *Question: The following describes the characteristics of high-pressure systems affecting the **Korean region**. Which statement is incorrect?* **1.** 시베리아 고기압은 겨울철의 춥고 건조한 날씨를 만든다. **2.** 오호츠크해 고기압은 동해안 지방의 고온현상을 일으킨다. **3.** 북태평양 고기압은 고온다습하며, 여름철의 무더운 날씨를 만든다. **4.** 이동성 고기압의 영향을 받으면 봄에는 따뜻한 날씨, 가을에는 맑은 날씨가 된다. *1. The Siberian High creates cold and dry weather during the winter.* *2. The Okhotsk Sea High causes high-temperature phenomena in the east coastal regions.* *3. The North Pacific High is hot and humid, creating sweltering weather during the summer.* *4. Under the influence of migratory highs, the weather becomes warm in spring and clear in autumn.* **정답:** 2 *Ground Truth: 2*	**ID:** 18, **Part:** 2 **질문:** 비열의 차원을 올바르게 나타낸 것은 무엇입니까? *Question: What is the correct dimensional representation of specific heat?* **1.** [$L^2T^2\theta^{-1}$] **2.** [$L^2T^{-2}\theta^{-1}$] **3.** [$ML^{-1}T^{-2}$] **4.** [ML^2T^{-2}] **전문가 검증 참조 자료: 비열은 단위 질량당 단위 온도 상승에 필요한 에너지로서 차원은 (에너지)/(질량·온도) = $(ML^2T^{-2})/(M\theta) = L^2T^{-2}\theta^{-1}$ 이므로 2번이 맞고, 4번은 에너지 자체의 차원, 3번은 압력의 차원, 1번은 시간 지수가 부호가 반대라 틀립니다.** *Expert-Verified Rationale: Specific heat is the energy required to raise the temperature of a unit mass by one unit. Its dimension is (Energy)/(Mass · Temperature) = $(ML^2T^{-2})/(M\theta) = L^2T^{-2}\theta^{-1}$. Therefore, option 2 is correct. Option 4 represents the dimension of energy itself, option 3 represents the dimension of pressure, and option 1 is incorrect because the sign of the time exponent is reversed.* **정답:** 2 *Ground Truth: 2*
Multimodal	**ID:** 460, **Part:** 3 **질문:** 북반구에서 나타나는 지균풍 ($\vec{V_{g}}$), 실제풍 ($\vec{V}$), 수평가속도 ($(\frac{d\vec{v}}{dt})_{H}$) 사이의 관계를 올바르게 표현한 그림은 어느 것인가? *Question: Which figure correctly represents the relationship between the geostrophic wind ($\vec{V_{g}}$), the actual wind ($\vec{V}$), and the horizontal acceleration ($(\frac{d\vec{v}}{dt})_{H}$) in the Northern Hemisphere?*  **1.**  **2.**  **3.**  **4.** **정답:** 1 *Ground Truth: 1*	**ID:** 687, **Part:** 4 **질문:** 제시된 그림은 한국의 어떤 지점의 연평균 물수지를 보여준다. 이 그림에서 D 부분이 의미하는 것은 무엇인가? *Question: The presented figure shows the annual average water balance of a certain location in **Korea**. What does section D in this figure represent?*  1. 토양수분의 과잉 2. 토양수분의 보충 3. 토양수분의 이용 4. 토양수분의 결핍 *1. Soil moisture surplus* *2. Soil moisture recharge* *3. Soil moisture utilization* *4. Soil moisture deficit* **정답:** 3 *Ground Truth: 3*	**ID:** 460, **Part:** 3 **질문:** 다음 그림이 보여주는 역전층의 종류로 옳은 것은? *Question: Which of the following is the correct type of inversion layer shown in the figure below?*  1. 복사역전 2. 난류역전 3. 전선역전 4. 침강역전 *1. Radiation inversion* *2. Turbulence inversion* *3. Frontal inversion* *4. Subsidence inversion* **전문가 검증 참조 자료: 그림처럼 지표에서 바로 시작하는 얕은 역전층이 위로 갈수록 약화되는 형태는 야간 지표 복사냉각으로 생기는 복사역전의 전형이며, 침강역전은 고기압 하 하강류로 상층에 분리되어 나타나고 전선역전은 전선면을 따라 경사져 있으며 난류역전은 주간 혼합층 꼭대기에 형성되어 지표에서 시작하지 않으므로 그림과 다르다.** *Expert-Verified Rationale: As shown in the figure, a shallow inversion layer starting directly from the surface and weakening with height is typical of a radiation inversion caused by nocturnal surface radiative cooling. In contrast, a subsidence inversion appears separated in the upper atmosphere due to downdrafts under high pressure, a frontal inversion is tilted along a frontal surface, and a turbulence inversion forms at the top of the daytime mixed layer rather than starting at the surface; thus, they differ from the figure.* **정답:** 1 *Ground Truth: 1*

Table 7: **Comparison of `InternVL3.5-8B-Instruct` and `Qwen3-VL-235B-A22B-Thinking` on K-MetBench reasoning question.** (❌: Incorrect, ✅: Correct, shaded text : Highlighted errors)

Reasoning MCQA	Reasoning Evaluation of `InternVL3.5-8B-Instruct`	Reasoning Evaluation of `Qwen3-VL-235B-A22B-Thinking`
ID: 105, **Part:** 3 **질문:** 지균풍 관계식에서 기압경도와 공기밀도가 일정하다면, 풍속은 어떻게 결정되는가? *Question: In the geostrophic wind formula, if the pressure gradient and air density are constant, how is the wind speed determined?* **1.** 1. 고도가 증가함에 따라 증가한다. **2.** 2. 위도의 증가에 따라 감소한다. **3.** 3. 기압의 증가에 따라 증가한다. **4.** 4. 공기의 점성이 크면 증가한다. *1. It increases as altitude increases.* *2. It decreases as latitude increases.* *3. It increases as pressure increases.* *4. It increases if air viscosity is high.* **전문가 검증 참조 자료:** 지균풍은 $Vg = \|p\|/(\rho f)$이므로 기압경도와 밀도가 일정하면 코리올리매개변수 $f=2\Omega \sin\varphi$만이 변수가 되어 위도가 증가할수록 f가 커져 풍속은 감소하며, 절대 기압의 크기(보기 3)나 고도(보기 1), 점성(보기 4)은 이 관계식에 직접 등장하지 않는다. *Expert-Verified Rationale: Since the geostrophic wind is defined as $Vg = \|p\|/(\rho f)$, if the pressure gradient and density are constant, the Coriolis parameter $f=2\Omega \sin\varphi$ becomes the only variable. As latitude increases, f increases, causing the wind speed to decrease. The magnitude of absolute pressure (Option 3), altitude (Option 1), and viscosity (Option 4) do not appear directly in this equation.* **정답:** 2 *Ground Truth: 2*	**Response from `InternVL3.5-8B-Instruct`** **생성된_추론_근거:** 지연풍 관계식에서 풍속이 기압 경로의 변화와 직접적으로 연관된다는 것을 이해할 수 있습니다. 기압 경도 가 증가하면 풍속도 증가하는 경향이 있으며, 이는 풍속이 기압의 증가에 따라 증가한다는 3번 옵션이 정답이라는 근거를 제공합니다. *Expert-Verified Rationale: From the geostrophic wind relationship , we can understand that wind speed is directly associated with changes in the pressure path. Since wind speed tends to increase when the pressure gradient increases, this provides grounds for Option 3—that wind speed increases as pressure increases—being the correct answer.* **정답:** 3 ❌ *Predicted Answer: 3* - **Reasoning Evaluation from Human Experts:** **사람 전문가 평균 점수:** 사실적 정확성 (1), 논리적 완결성 (1), 추론의 깊이 (1), 표현의 명확성 (1) ***Average of Human Expert Score:** factuality (1), logicality (1), depth (1), clarity (1)* **평가 의견 1:** 오답 및 사용된 용어 부적절: 기압경로 , 지연풍 등 **평가 의견 2:** 문제에 맞지 않는 추론근거 및 답도 틀림. ***Evaluation Comment 1:** Incorrect answer and inappropriate use of terminology: 'pressure path' , 'delayed wind', etc.* ***Evaluation Comment 2:** The reasoning is irrelevant to the question, and the answer is incorrect.* - **Reasoning Evaluation from LLM-Evaluator (`Gemini-2.5-Pro`):** **사실_오류_목록:** [지균풍(geostrophic wind)을 '지연풍'이라는 존재하지 않는 기상학 용어로 잘못 지칭하였습니다., 문제에서 '기압경도가 일정하다'는 핵심 가정을 무시하고, '기압경도가 증가하면 풍속이 증가한다'는 관계를 추론의 근거로 사용하였습니다. '기압경도(pressure gradient)'와 '기압(pressure)'이라는 물리적으로 다른 개념을 혼동하여, 기압경도와 풍속의 관계를 근거로 기압과 풍속의 관계를 설명하는 오류를 범했습니다.] **평가_점수:** 정확성: 1, 논리적_완결성: 1, 추론의_깊이: 1, 표현의_명확성: 1 **평가_사유:** 수험생 AI의 답변은 모든 평가 축에서 최하점을 기록했습니다. '정확성' 측면에서, '지균풍'을 '지연풍'이라는 환각 용어로 지칭했으며, '기압경도'와 '기압'을 혼동하는 치명적인 개념 오류를 보였습니다. 또한 문제의 핵심 전제인 '기압경도 일정'을 무시하여 사실 관계를 왜곡했습니다. '논리적 완결성' 측면에서, 문제의 조건과 무관한 일반론을 근거로 제시하고, '기압경도'와 '기압'을 혼동하여 결론을 도출했기 때문에 근거가 결론을 전혀 뒷받침하지 못합니다. '추론의 깊이' 측면에서, 지균풍의 핵심 원리인 코리올리 힘과 위도와의 관계에 대한 이해 없이 피상적인 사실을 나열하는 데 그쳤습니다. '표현의 명확성' 측면에서도, 존재하지 않는 용어와 모호한 표현('기압 경로의 변화')을 사용하여 의미 전달에 심각한 결함이 있습니다. ***List_of_Factual_Errors:** [* *It incorrectly referred to 'geostrophic wind' as 'delayed wind', a non-existent meteorological term.,* *It disregarded the problem's key assumption that the 'pressure gradient is constant' and used the relationship 'wind speed increases as pressure gradient increases' as the basis for reasoning.,* *It confused the distinct physical concepts of 'pressure gradient' and 'pressure', erroneously explaining the relationship between pressure and wind speed based on the relationship between pressure gradient and wind speed.,* *] **Evaluation_Scores:** Accuracy: 1, Logical_Completeness: 1, Depth_of_Reasoning: 1, Clarity_of_Expression: 1* ***Evaluation_Reasoning:** The candidate AI's response recorded the lowest scores across all evaluation axes. In terms of 'Accuracy', it referred to 'geostrophic wind' using the hallucinated term 'delayed wind' and demonstrated a fatal conceptual error by confusing 'pressure gradient' with 'pressure'. Furthermore, it distorted factual relations by ignoring the problem's core premise that the 'pressure gradient is constant'. In terms of 'Logical_Completeness', the reasoning failed to support the conclusion at all because it presented generalities irrelevant to the problem conditions and derived the conclusion by confusing 'pressure gradient' with 'pressure'. In terms of 'Depth_of_Reasoning', it merely listed superficial facts without understanding the relationship between the Coriolis force and latitude, which is the core principle of geostrophic wind. In terms of 'Clarity_of_Expression', there were serious flaws in meaning delivery due to the use of non-existent terms and ambiguous expressions (e.g., 'change in pressure path').*	**Response from `Qwen3-VL-235B-A22B-Thinking`** **생성된_추론_근거:** 지균풍은 $Vg = \|p\|/(\rho f)$이므로 기압경도와 밀도가 일정하면 코리올리매개변수 $f=2\Omega \sin\varphi$만이 변수가 되어 위도가 증가할수록 f가 커져 풍속은 감소하며, 절대 기압의 크기(보기 3)나 고도(보기 1), 점성(보기 4)은 이 관계식에 직접 등장하지 않는다. *Expert-Verified Rationale: Since the geostrophic wind is defined as $Vg = \|p\|/(\rho f)$, if the pressure gradient and density are constant, the Coriolis parameter $f=2\Omega \sin\varphi$ becomes the only variable. As latitude increases, f increases, causing the wind speed to decrease. The magnitude of absolute pressure (Option 3), altitude (Option 1), and viscosity (Option 4) do not appear directly in this equation.* **정답:** 2 ✅ *Predicted Answer: 2* - **Reasoning Evaluation from Human Experts:** **사람 전문가 평균 점수:** 사실적 정확성 (5), 논리적 완결성 (5), 추론의 깊이 (4.5), 표현의 명확성 (5) ***Average of Human Expert Score:** factuality (5), logicality (5), depth (4.5), clarity (5)* ***Comments:*** - **Reasoning Evaluation from LLM-Evaluator (`Gemini-2.5-Pro`):** **사실_오류_목록:** [ ] **평가_점수:** 정확성: 5, 논리적_완결성: 5, 추론의_깊이: 5, 표현의_명확성: 5 **평가_사유:** 수험생 AI의 답변은 모든 평가 축에서 매우 우수한 수준을 보였습니다. '정확성' 측면에서, 제시된 지균풍 속도 공식, 코리올리 계수와 위도의 관계, 그리고 결론까지 기상학적 사실에 완벽하게 부합하며 어떠한 오류도 발견되지 않았습니다. '논리적 완결성' 측면에서, 문제의 조건(기압경도, 밀도 일정)으로부터 풍속이 코리올리 계수에 반비례한다는 핵심 관계를 명확히 하고, 이를 위도와의 관계로 확장하여 결론을 도출하는 과정이 비약 없이 완벽하게 연결되었습니다. '추론의 깊이' 측면에서, 정답의 근거를 제시하는 것에 그치지 않고, 매력적인 오답 선택지(3번, 4번)가 왜 오답인지를 지균풍의 기본 가정(마찰 무시)과 핵심 변수(기압경도)를 들어 명확히 반박함으로써 문제에 대한 깊이 있는 이해를 보여주었습니다. '표현의 명확성' 측면에서도 군더더기 없이 핵심 내용을 간결하고 명확하게 전달하여 이해도를 높였습니다. 이는 모범적인 해설이라 할 수 있습니다. ***List_of_Factual_Errors:** [ ]* ***Evaluation_Scores:** Accuracy: 5, Logical_Completeness: 5, Depth_of_Reasoning: 5, Clarity_of_Expression: 5* ***Evaluation_Reasoning:** The candidate AI's response demonstrated a very high level of performance across all evaluation axes. In terms of 'Accuracy', the presented geostrophic wind speed formula, the relationship between the Coriolis parameter and latitude, and the conclusion perfectly matched meteorological facts with no errors found. In terms of 'Logical_Completeness', the process of clarifying the core relationship that wind speed is inversely proportional to the Coriolis parameter from the problem conditions (constant pressure gradient and density) and extending this to the relationship with latitude to derive the conclusion was perfectly connected without logical leaps. In terms of 'Depth_of_Reasoning', it demonstrated a deep understanding of the problem not only by providing the basis for the correct answer but also by clearly refuting why the attractive distractors (Options 3 and 4) were incorrect, citing the basic assumption of geostrophic wind (ignoring friction) and the key variable (pressure gradient). In terms of 'Clarity_of_Expression', it delivered the core content concisely and clearly without redundancy, enhancing comprehensibility. This can be considered an exemplary explanation.*

A Dataset Examples

Table 6 presents representative examples from K-MetBench, organized into two primary modality groups: *Text-only* and *Multimodal*. Within each modality, we further stratify the tasks into three distinct categories to evaluate comprehensive meteorological capabilities.

Text-only Tasks assesses linguistic reasoning and theoretical knowledge without visual interpretation. This group includes (a) *Standard MCQA* for fundamental concepts, (b) *K-Specific MCQA* which requires geo-cultural knowledge specific to the Korean Peninsula, and (c) *Reasoning MCQA* that demands multi-step logical deduction.

Multimodal Tasks introduces visual data interpretation, a critical skill for meteorologists. This group parallels the text-only structure with (d) *Standard,* (e) *K-Specific,* and (f) *Reasoning* subsets, but specifically evaluates the model's ability to analyze weather charts, satellite imagery, and atmospheric diagrams. This structured categorization allows for a clear comparison of model performance across different modalities and levels of domain expertise.

B Case Study of Reasoning Answer

Two human experts and the LLM-Evaluator (`gemini-2.5-pro`) conducted evaluations using identical rubrics. As shown in Table 7, we observed consensus between the human and AI evaluators for the `InternVL3.5-8B-Instruct` and `Qwen3-VL-235B-A22B-Thinking` models: both correctly identified incorrect answers and highlighted inappropriate terminology in the reasoning rationale.

Notably, the human expert made a specific error by misreading '기압 경도' (pressure gradient) as '기압경로' (pressure path). While the quantitative scores assigned by the human expert and the LLM-Evaluator were comparable, the granularity of their feedback differed significantly. The human expert made an implicit judgment, providing summary comments alongside the score. In contrast, the LLM-Evaluator generated more detailed outputs, including explicit justifications and comprehensive lists of factual errors.

C Prompts and Questionnaires for Benchmark Construction

This section details the prompts utilized for data augmentation (paraphrasing) and the identification of domain-specific subsets. These processes were conducted to enhance the quality of the dataset and provide rich learning signals.

C.1 Question Paraphrasing

To diversify sentence structures and lexical expressions while preserving the original semantic meaning of the questions, we utilized the `Gemini-2.5-Pro` model. Figure 5 presents the specific system prompt employed for this paraphrasing task.

C.2 Identification of Korean-Specific Subset

To identify questions containing Korean-specific geographical and cultural contexts (the *Korean-Specific* subset) from the total pool of 1,774 questions, we established a hybrid pipeline combining LLM-based filtering with human verification.

LLM-Aided Identification The screening process involved independent filtering using two distinct models: `Gemini-2.5-Pro` and `GPT-4.1`. The identification prompts for each model were optimized through an iterative refinement process to maximize recall. Figures 8 and 10 illustrate the final enhanced prompts used for identifying Korean-specific context questions, respectively.

Human Selection Process Based on the LLM filtering, `Gemini-2.5-Pro` extracted 135 candidates, while `GPT-4.1` extracted 95 candidates. We consolidated these results into a union of 149 unique questions. Subsequently, two human researchers performed cross-validation on this candidate set to finalize the *Korean-Specific* subset. The selected questions typically contain high-context keywords such as "Our country" (우리나라), "Korean Peninsula," "Jeju," "Seoul," "Yeongdong," "Southerly wind" (마파람), "Taebaek Mountains," and "24 Solar Terms."

C.3 Implicit vs. Explicit Dataset Design for Korean-Specific Subset

To ensure a fair evaluation of local context understanding regardless of the model's primary training language, we constructed a dual-version dataset by converting implicit questions into explicit ones.

- **Implicit Questions:** These refer to the original items containing high-context expressions that presuppose the speaker's spatiotemporal and cultural location (e.g., "Our country," "Maparam," "East Coast").

- **Explicit Questions:** These refer to the modified items where human researchers manually replaced high-context references with objective and unambiguous terminology (e.g., changing "Our country" to "South Korea" or "Maparam" to "Southerly wind, a pure Korean term").

Table 8 presents comparative examples of these original implicit questions and their explicit counterparts.

Table 8: **Examples of Context Transformation from implicit to explicit forms**

ID	Implicit (Before)	Explicit (After)
All	우리나라	한국 지역
618	서울	한국의 서울 지역
1037	24절기	동아시아 지역의 24절기
822	동해안	한국 지역의 동해안
557	겨울철 발해만에서 작은 기압골이 접근하고 있다.	겨울철 발해만으로부터 소규모 기압골이 한국 지역으로 접근하는 상황에서
271, 1744	마파람	한국 지역의 지방풍인 마파람

C.4 Evaluation Prompts for Korean-Specific Subset

This section details the construction of system prompts designed to evaluate the model's understanding of geo-cultural contexts. To encourage the model to effectively utilize its latent local knowledge, we designed an *Advanced Prompt* that explicitly defines the speaker's persona (i.e., a Korean meteorology expert) and clarifies that the questions are contextually situated in Korea.

To quantify the prompting gain—the extent to which this contextual cuing aids performance—and to ensure equitable evaluation for non-Korean models, we also established a *Standard Prompt* as a control group. Figure 18 presents the standard system prompt used for the baseline experiment, while Figure 20 displays the advanced system prompt used to test the activation of geo-cultural knowledge.

C.5 Prompt for Reference Rationale Generation

To secure high-quality reasoning references (rationales) for the benchmark, we utilized the `GPT-5` model. The prompt engineering process employed an iterative refinement technique. Specifically, we established a loop where an *Enhancer* model drafted the initial prompt and a *Critic* model identified weaknesses for revision, using `GPT-5` for both roles to derive the optimal instruction. The final system prompt used for rationale generation is presented in Figure 12.

To ensure comprehensive coverage, the target questions were selected via stratified sampling to include all subject areas, modalities (text-only/multimodal), and Korean-specific items. Furthermore, to guarantee the validity of the reasoning paths, we enforced a strict filtering protocol: if the model generated an incorrect answer, the generation process was repeated until a rationale leading to the correct answer was produced.

C.6 Questionnaire for Expert Verification on Reference Rationale

To ensure the reliability of the LLM-as-a-judge pipeline, two human experts conducted a rigorous verification of the generated rationales from September 9 to October 19, 2025. This process was critical for establishing the integrity of the reference data.

Out of 142 rationales initially generated by `GPT-5`, experts provided feedback for revision on 19 cases (13.38%). The revisions primarily addressed technical accuracy and clarity. Specifically, experts corrected erroneous terminology (5 cases), such as changing '비열용량' (specific heat

System Prompt for Question Paraphrasing (Gemini-2.5-Pro)

Task: Paraphrase the following multiple-choice question about meteorology in Korean.
Rules:
- Preserve the core meaning and all technical terminology.
- Change the sentence structure or phrasing for a more natural flow.

Output Format: Provide only the final paraphrased text. Do not include any introductory phrases or explanations.
—
Original Question: "{original_question}"

Figure 5: **System prompt used to paraphrase Korean meteorological questions**

capacity) to '비열' (specific heat) or '지표소용돌이도' to '행성소용돌이도' (planetary vorticity). They also reinforced variable explanations and standard units (3 cases); for instance, refining the phrasing "among the temperatures handled" to "among the variables handled in atmospheric science" because '혼합비' (mixing ratio) is not a temperature variable. Additionally, the revisions included full sentence rewriting (7 cases), supplementary explanations (2 cases), and minor stylistic polishing (2 cases) to align with standard Korean meteorological conventions (e.g., standardizing '포텐셜온도' to '온위', '바트로픽' to '순압', and '단열가열' to '단열압축').

In addition to refining the AI-generated rationales, this expert review also identified inherent defects in the raw exam data. One question (ID 276) was discarded from the dataset as it was deemed logically unsolvable. Furthermore, questions with syntactic errors (IDs 308, 650) or issues with option configuration/double answers (IDs 14, 583, 1665) were precisely corrected based on expert consultation. Through this process, we secured the integrity of the final 141 reasoning evaluation samples. Table 9 presents the specific questionnaire used for this expert verification process.

C.7 Reasoning Prompt for Open-Source LLMs

Figure 14 presents the system prompt utilized for generating reasoning paths and answers from open-source LLMs. It is important to note that this prompt serves as the standard instruction for the main inference phase of our benchmark evaluation protocol, rather than an experimental variation.

C.8 Prompt for LLM-as-a-Judge Evaluation

Figure 16 illustrates the specific system prompt employed for the LLM-as-a-Judge evaluation pipeline. The prompt was meticulously designed with the following key considerations to ensure robust alignment with human expert evaluation:

- **Unified Evaluation Scale:** We adopted the identical 1-to-5 Likert scale and four evaluation axes —*Factual Accuracy, Logical Soundness, Depth of Reasoning,* and *Clarity & Conciseness*—used by human experts. This unification allows for direct statistical comparison and correlation analysis between LLM and expert scores.

- **Enforced Chain of Thought (CoT):** To enhance consistency, the prompt explicitly mandates a step-by-step thinking process. The evaluator is required to verify facts against the provided expert reference material *before* assigning scores, thereby minimizing hallucinations and ensuring evidence-based grading.

- **Explicit Scoring Criteria:** To prevent arbitrary scoring, we defined concrete rubrics for specific score tiers (e.g., distinguishing between a 5-point perfect answer and a 3-point answer with minor errors).

- **Structured Output:** The prompt enforces a strict JSON output format that separates the "List of Factual Errors" from the quantitative scores. This structural constraint compels the model to explicitly isolate factual hallucinations from qualitative reasoning flaws.

C.9 Questionnaire for Expert Scoring of LLM Reasoning

Table 10 outlines the questionnaire and scoring rubric provided to two meteorology professors. Crucially, this rubric served as the blueprint for the LLM-as-a-Judge prompt described above, ensuring that both human and AI evaluators operated under identical standards regarding accuracy and reasoning quality.

D Experimental Setups

D.1 Meta-Evaluation for LLM-as-a-Judge

To validate the reliability of the LLM evaluator, we designed a meta-evaluation protocol consisting of three steps: 1) generating reasoning paths and answers using various open-source LLMs; 2) performing LLM-as-a-Judge evaluation using expert-verified references and a specific rubric (based on a 5-point Likert scale across four evaluation axes); and 3) obtaining scores from two human experts using the identical rubric to calculate the statistical correlation between the LLM judge and human experts. The detailed prompt for the main inference, the judge prompt, and the expert scoring questionnaire are provided in Appendix C.7, C.8, and C.9, respectively.

Sampling Strategy of Target Models To ensure that the LLM-as-a-Judge can reliably evaluate reasoning capabilities across a broad spectrum of proficiency, we employed a performance-based stratified sampling strategy. We categorized the pool of candidate models into three distinct tiers—Top, Mid, and Low—based on their QA accuracy on the 141 reasoning questions. From these strata, we selected representative models to form a final set of 10 target models for the meta-evaluation, ensuring that the judge is tested against both high-quality coherent reasoning and lower-quality outputs. The list of sampled models is detailed in Table 11.

Table 11: **List of Sampled Models for Meta-Evaluation of LLM-as-a-Judge.** Models were selected via stratified sampling based on their reasoning accuracy tiers to ensure diverse evaluation targets. Acc. denotes the accuracy score on the *Reasoning* subset questions.

Tier	Model Name	Family	Size (B)	Acc.
Top	Qwen3-VL-235B-A22B-Thinking	Qwen	235.0	4.31
	gpt-oss-120b	OpenAI	120.0	4.03
	A.X-4.0	SKT	72.0	3.87
Mid	command-a-reasoning-08-2025	Cohere	111.0	3.53
	gpt-oss-20b	OpenAI	20.0	3.39
	VARCO-VISION-2.0-14B	NCSoft	14.0	2.81
	InternVL3.5-14B-Instruct	OpenGVLab	15.0	2.36
Low	Llama-3.1-8B-Instruct	Meta	8.0	1.91
	InternVL3.5-8B-Instruct	OpenGVLab	8.0	1.77
	Qwen3-0.6B	Qwen	0.6	1.14

Stratified Sampling of Evaluation Items To establish a robust gold standard for scoring, we selected 10 representative reasoning questions. Instead of random selection, we applied a stratified sampling to ensure both comprehensiveness and discriminatory power. The selection process involved the following criteria:

- **Item Difficulty:** We classified the 141 reasoning questions into three difficulty tiers based on the average accuracy of 10 open-source LLMs: Hard (Top 30%), Mid (40%), and Easy (Bottom 30%). We sampled 3, 4, and 3 questions from each respective group to balance the difficulty distribution.

- **Discriminatory Power:** Within each difficulty tier, we prioritized questions with a high standard deviation in accuracy across the 10 models. A high standard deviation indicates that the question effectively discriminates between high- and low-performing models.

- **Category Coverage:** The selection was further constrained to ensure a balanced inclusion of text-only, multimodal, and Korean-specific questions, as well as coverage across the official exam subject areas (Parts 1, 3, 4, 5).

Based on these criteria, the final 10 questions selected for meta-evaluation are: IDs 1224, 963, 1745, 456, 1618, 1694, 131, 1590, 105, and 14. Table 12 details the characteristics of these sampled items.

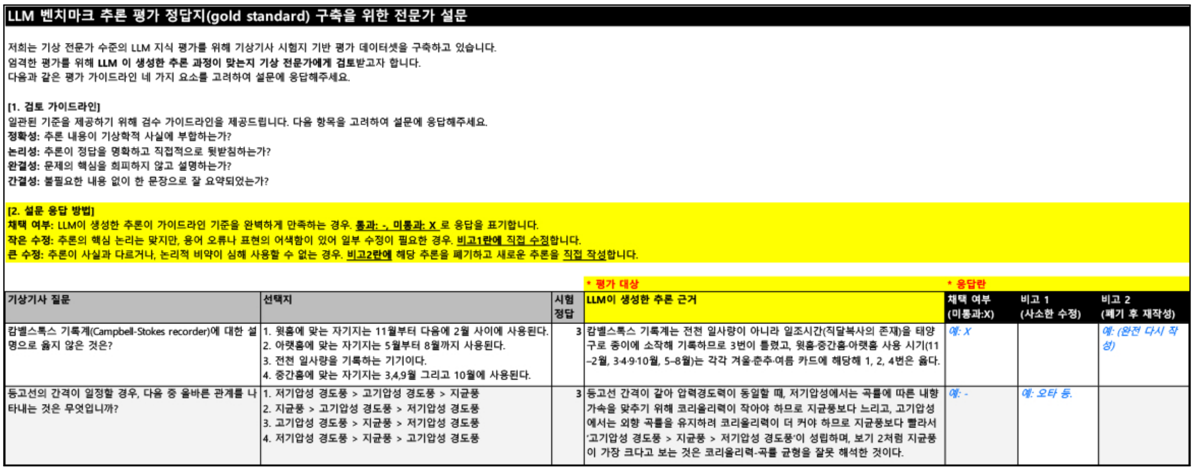
Table 12: **Statistics of Selected Evaluation Items.** Mean and Std. Dev. represent the item-wise normalized reasoning scores (1-5) across the selected models.

Tier	ID	Mean	Std. Dev.	Part	Note
Hard	105	2.23	1.72	3	-
	1590	2.55	1.91	3	-
Mid	14	2.70	1.89	1	-
	456	2.73	1.88	3	-
	1694	3.10	1.93	5	-
	963	3.13	1.81	4	Korean
Easy	1745	3.35	1.89	4	-
	131	3.40	1.74	5	-
	1224	3.53	1.76	4	-

Table 9: **Expert verification questionnaire for reference rationales**

ID	Question	Choices	Exact Answer	Generated Rationale	Adopt?	Note 1 (Minor Fix)	Note 2 (Major Fix)
18	비열의 ...	1. ...	2	비열은 ...	yes	-	-

Note: *The gray-shaded cells indicate the items to be answered by the expert. Adoption Criteria* (*Accuracy, Logical Soundness, Completeness, and Conciseness*) *are provided separately.*



LLM 벤치마크 추론 평가 정답지(gold standard) 구축을 위한 전문가 설문

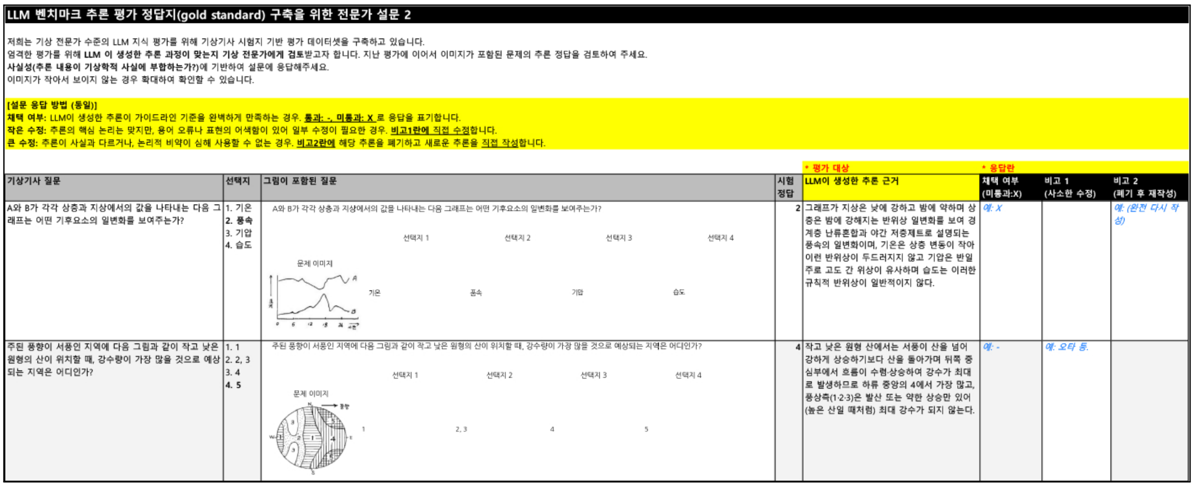
저희는 기상 전문가 수준의 LLM 지식 평가를 위해 기상기사 시험지 기반 평가 데이터셋을 구축하고 있습니다.
엄격한 평가를 위해 **LLM 이 생성한 추론 과정이 맞는지 기상 전문가에게 검토**받고자 합니다.
다음과 같은 평가 가이드라인 네 가지 요소를 고려하여 설문에 응답해주세요.

[1. 검토 가이드라인]
일관된 기준을 제공하기 위해 검수 가이드라인을 제공드립니다. 다음 항목을 고려하여 설문에 응답해주세요.
정확성: 추론 내용이 기상학적 사실에 부합하는가?
논리성: 추론이 정답을 명확하고 직접적으로 뒷받침하는가?
완결성: 문제의 핵심을 회피하지 않고 설명하는가?
간결성: 불필요한 내용 없이 한 문장으로 잘 요약되었는가?

[2. 설문 응답 방법]
채택 여부: LLM이 생성한 추론이 가이드라인 기준을 완벽하게 만족하는 경우. **통과: -, 미통과: X** 로 응답을 표기합니다.
작은 수정: 추론의 핵심 논리는 맞지만, 용어 오류나 표현의 어색함이 있어 일부 수정이 필요한 경우. **비고1란에 직접 수정**합니다.
큰 수정: 추론이 사실과 다르거나, 논리적 비약이 심해 사용할 수 없는 경우. **비고2란에** 해당 추론을 폐기하고 새로운 추론을 **직접 작성**합니다.

			* 평가 대상	* 응답란		
기상기사 질문	선택지	시험 정답	LLM이 생성한 추론 근거	채택 여부 (미통과:X)	비고 1 (사소한 수정)	비고 2 (폐기 후 재작성)
캄벨스톡스 기록계(Campbell-Stokes recorder)에 대한 설명으로 옳지 않은 것은?	1. 윗홈에 맞는 자기지는 11월부터 다음에 2월 사이에 사용된다. 2. 아랫홈에 맞는 자기지는 5월부터 8월까지 사용된다. 3. 전천 일사량을 기록하는 기기이다. 4. 중간홈에 맞는 자기지는 3,4,9월 그리고 10월에 사용된다.	3	캄벨스톡스 기록계는 전천 일사량이 아니라 일조시간(직달복사의 존재)을 태양구로 종이에 소작해 기록하므로 3번이 틀렸고, 윗홈·중간홈·아랫홈 사용 시기(11–2월, 3·4·9·10월, 5–8월)는 각각 겨울·춘추·여름 카드에 해당해 1, 2, 4번은 옳다.	예: X		예: (완전 다시 작성)
등고선의 간격이 일정할 경우, 다음 중 올바른 관계를 나타내는 것은 무엇입니까?	1. 저기압성 경도풍 > 고기압성 경도풍 > 지균풍 2. 지균풍 > 고기압성 경도풍 > 저기압성 경도풍 3. 고기압성 경도풍 > 지균풍 > 저기압성 경도풍 4. 저기압성 경도풍 > 지균풍 > 고기압성 경도풍	3	등고선 간격이 같아 압력경도력이 동일할 때, 저기압성에서는 곡률에 따른 내향 가속을 맞추기 위해 코리올리력이 작아야 하므로 지균풍보다 느리고, 고기압성에서는 외향 곡률을 유지하려 코리올리력이 더 커야 하므로 지균풍보다 빨라서 '고기압성 경도풍 > 지균풍 > 저기압성 경도풍'이 성립하며, 보기 2처럼 지균풍이 가장 크다고 보는 것은 코리올리력-곡률 균형을 잘못 해석한 것이다.	예: -	예: 오타 등.	

(a) Text-only Reasoning Questions



LLM 벤치마크 추론 평가 정답지(gold standard) 구축을 위한 전문가 설문 2

저희는 기상 전문가 수준의 LLM 지식 평가를 위해 기상기사 시험지 기반 평가 데이터셋을 구축하고 있습니다.
엄격한 평가를 위해 **LLM 이 생성한 추론 과정이 맞는지 기상 전문가에게 검토**받고자 합니다. 지난 평가에 이어서 이미지가 포함된 문제의 추론 정답을 검토하여 주세요.
사실성(추론 내용이 기상학적 사실에 부합하는가?)에 기반하여 설문에 응답해주세요.
이미지가 작아서 보이지 않는 경우 확대하여 확인할 수 있습니다.

[설문 응답 방법 (동일)]
채택 여부: LLM이 생성한 추론이 가이드라인 기준을 완벽하게 만족하는 경우. **통과: -, 미통과: X** 로 응답을 표기합니다.
작은 수정: 추론의 핵심 논리는 맞지만, 용어 오류나 표현의 어색함이 있어 일부 수정이 필요한 경우. **비고1란에 직접 수정**합니다.
큰 수정: 추론이 사실과 다르거나, 논리적 비약이 심해 사용할 수 없는 경우. **비고2란에** 해당 추론을 폐기하고 새로운 추론을 **직접 작성**합니다.

				* 평가 대상	* 응답란		
기상기사 질문	선택지	그림이 포함된 질문	시험 정답	LLM이 생성한 추론 근거	채택 여부 (미통과:X)	비고 1 (사소한 수정)	비고 2 (폐기 후 재작성)
A와 B가 각각 상층과 지상에서의 값을 나타내는 다음 그래프는 어떤 기후요소의 일변화를 보여주는가?	1. 기온 **2. 풍속** 3. 기압 4. 습도	A와 B가 각각 상층과 지상에서의 값을 나타내는 다음 그래프는 어떤 기후요소의 일변화를 보여주는가? 선택지 1: 기온 / 선택지 2: 풍속 / 선택지 3: 기압 / 선택지 4: 습도 문제 이미지	2	그래프가 지상은 낮에 강하고 밤에 약하며 상층은 밤에 강해지는 반위상 일변화를 보여 경계층 난류혼합과 야간 저층제트로 설명되는 풍속의 일변화이며, 기온은 상층 변동이 작아 이런 반위상이 두드러지지 않고 기압은 반일주로 고도 간 위상이 유사하며 습도는 이러한 규칙적 반위상이 일반적이지 않다.	예: X		예: (완전 다시 작성)
주된 풍향이 서풍인 지역에 다음 그림과 같이 작고 낮은 원형의 산이 위치할 때, 강수량이 가장 많을 것으로 예상되는 지역은 어디인가?	1. 1 2. 2, 3 3. 4 **4. 5**	주된 풍향이 서풍인 지역에 다음 그림과 같이 작고 낮은 원형의 산이 위치할 때, 강수량이 가장 많을 것으로 예상되는 지역은 어디인가? 선택지 1: 1 / 선택지 2: 2, 3 / 선택지 3: 4 / 선택지 4: 5 문제 이미지	4	작고 낮은 원형 산에서는 서풍이 산을 넘어 강하게 상승하기보다 산을 돌아가며 뒤쪽 중심부에서 흐름이 수렴·상승하여 강수가 최대로 발생하므로 하류 중앙의 4에서 가장 많고, 풍상측(1·2·3)은 발산 또는 약한 상승만 있어 (높은 산일 때처럼) 최대 강수가 되지 않는다.	예: -	예: 오타 등.	

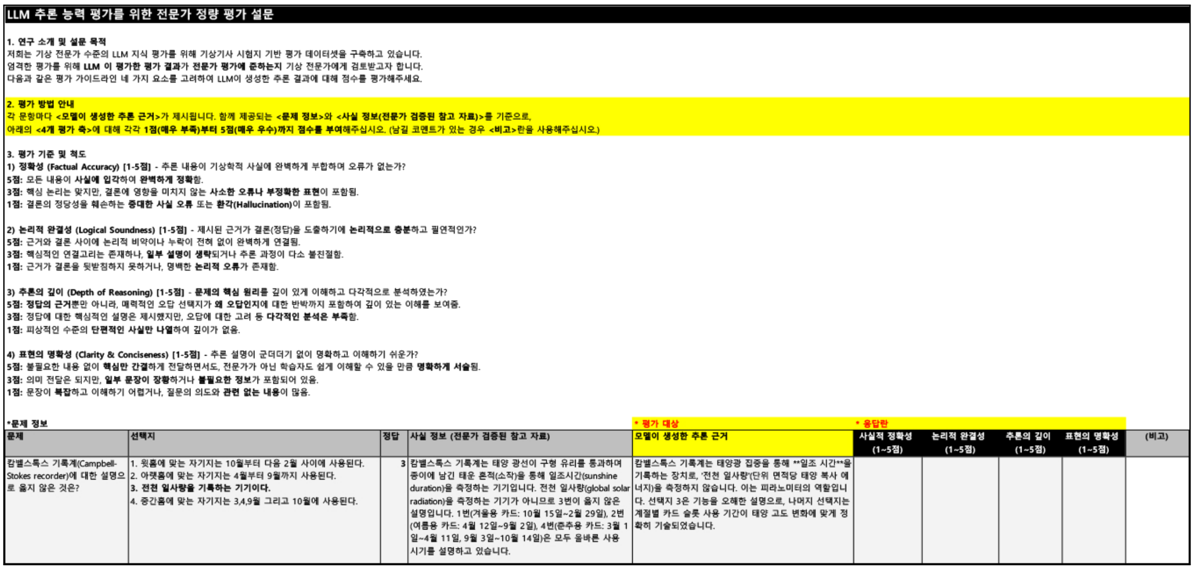
(b) Multimodal Reasoning Questions

Figure 6: **Examples of the expert verification questionnaire for reference rationales** (in Korean)

Table 10: **Questionnaire for expert scoring of open-source LLM reasoning results**

ID	Question	Choices	Exact Answer	GT-R	Target R	Fact (1-5)	Sound (1-5)	Depth (1-5)	Clear (1-5)	Total (4-20)	Note
18	비열의 ...	1. ...	2	비열은 ...	비열용량은 ...	1	2	1	5	9	-

Note: *GT-R: Expert-verified Gold Rationale, Target R: LLM Generated Reasoning. Scoring: 1(Poor) – 5(Excellent). The gray-shaded cells indicate the items to be answered by the expert.*



LLM 추론 능력 평가를 위한 전문가 정량 평가 설문

1. 연구 소개 및 설문 목적
저희는 기상 전문가 수준의 LLM 지식 평가를 위해 기상기사 시험지 기반 평가 데이터셋을 구축하고 있습니다.
엄격한 평가를 위해 **LLM 이 평가한 평가 결과**가 **전문가 평가에 준하는지** 기상 전문가에게 검토받고자 합니다.
다음과 같은 평가 가이드라인 네 가지 요소를 고려하여 LLM이 생성한 추론 결과에 대해 점수를 평가해주세요.

2. 평가 방법 안내
각 문항마다 **<모델이 생성한 추론 근거>**가 제시됩니다. 함께 제공되는 **<문제 정보>**와 **<사실 정보(전문가 검증된 참고 자료)>**를 기준으로,
아래의 **<4개 평가 축>**에 대해 각각 **1점(매우 부족)부터 5점(매우 우수)까지 점수를 부여**해주십시오. (남길 코멘트가 있는 경우 **<비고>**란을 사용해주십시오.)

3. 평가 기준 및 척도
1) 정확성 (Factual Accuracy) [1-5점] - 추론 내용이 기상학적 사실에 완벽하게 부합하며 오류가 없는가?
5점: 모든 내용이 **사실에 입각**하여 **완벽하게 정확**함.
3점: 핵심 논리는 맞지만, 결론에 영향을 미치지 않는 **사소한 오류나 부정확한 표현**이 포함됨.
1점: 결론의 정당성을 훼손하는 **중대한 사실 오류** 또는 **환각**(Hallucination)이 포함됨.

2) 논리적 완결성 (Logical Soundness) [1-5점] - 제시된 근거가 결론(정답)을 도출하기에 **논리적으로 충분**하고 필연적인가?
5점: 근거와 결론 사이에 논리적 비약이나 누락이 전혀 없이 완벽하게 연결됨.
3점: 핵심적인 연결고리는 존재하나, **일부 설명이 생략**되거나 추론 과정이 다소 불친절함.
1점: 근거가 결론을 뒷받침하지 못하거나, 명백한 **논리적 오류**가 존재함.

3) 추론의 깊이 (Depth of Reasoning) [1-5점] - **문제의 핵심 원리**를 깊이 있게 이해하고 다각적으로 분석하였는가?
5점: 정답의 근거뿐만 아니라, 매력적인 오답 선택지가 **왜 오답인지에** 대한 반박까지 포함하여 깊이 있는 이해를 보여줌.
3점: 정답에 대한 핵심적인 설명은 제시했지만, 오답에 대한 고려 등 **다각적인 분석은 부족**함.
1점: 피상적인 수준의 **단편적인 사실만 나열**하여 깊이가 없음.

4) 표현의 명확성 (Clarity & Conciseness) [1-5점] - 추론 설명이 군더더기 없이 명확하고 이해하기 쉬운가?
5점: 불필요한 내용 없이 **핵심만 간결**하게 전달하면서도, 전문가가 아닌 학습자도 쉽게 이해할 수 있을 만큼 **명확하게 서술**됨.
3점: 의미 전달은 되지만, **일부 문장이 장황**하거나 **불필요한 정보**가 포함되어 있음.
1점: 문장이 **복잡**하고 이해하기 어렵거나, 질문의 의도와 **관련 없는 내용**이 많음.

*문제 정보

문제	선택지	정답	사실 정보 (전문가 검증된 참고 자료)	* 평가 대상 모델이 생성한 추론 근거	* 응답란 사실적 정확성 (1~5점)	논리적 완결성 (1~5점)	추론의 깊이 (1~5점)	표현의 명확성 (1~5점)	(비고)
캄벨스톡스 기록계(Campbell-Stokes recorder)에 대한 설명으로 옳지 않은 것은?	1. 윗홈에 맞는 자기지는 10월부터 다음 2월 사이에 사용된다. 2. 아랫홈에 맞는 자기지는 4월부터 9월까지 사용된다. **3. 전천 일사량을 기록하는 기기이다.** 4. 중간홈에 맞는 자기지는 3,4,9월 그리고 10월에 사용된다.	3	캄벨스톡스 기록계는 태양 광선이 구형 유리를 통과하여 종이에 남긴 태운 흔적(소작)을 통해 일조시간(sunshine duration)을 측정하는 기기입니다. 전천 일사량(global solar radiation)을 측정하는 기기가 아니므로 3번이 옳지 않은 설명입니다. 1번(겨울용 카드: 10월 15일~2월 29일), 2번(여름용 카드: 4월 12일~9월 2일), 4번(춘추용 카드: 3월 1일~4월 11일, 9월 3일~10월 14일)은 모두 올바른 사용 시기를 설명하고 있습니다.	캄벨스톡스 기록계는 태양광 집중을 통해 **일조 시간**을 기록하는 장치로, '전천 일사량'(단위 면적당 태양 복사 에너지)을 측정하지 않습니다. 이는 피라노미터의 역할입니다. 선택지 3은 기능을 오해한 설명으로, 나머지 선택지는 계절별 카드 슬롯 사용 기간이 태양 고도 변화에 맞게 정확히 기술되었습니다.					

Figure 7: **Examples of the questionnaire for expert scoring of open-source LLM reasoning results** (in Korean)

System Prompt for Identifying Regional Questions (Gemini-2.5-Pro)

너는 이제부터 기상학 전문가야.
주어진 '기상기사 문제'를 분석해 줘. 각 문제의 내용과 선택지를 신중하게 검토해서, 전 세계적으로 적용되는 일반적인 기상 지식이 아니라 오직 한국의 지리, 기후, 기상 시스템, 관련 기관에만 해당되는 '지역성이 강한' 문제를 식별해 줘.

다음과 같은 기준을 사용해서 문제를 선별해:
- 한국의 특정 지역(예: 영동, 영서, 울릉도, 서해안)의 기상 현상을 다루는 문제
- 한국에서만 사용하는 특정 기상 용어나 예보 시스템에 관한 문제 (예: 동네예보)
- 한국 기상청 또는 관련 기관의 역할이나 업무와 직접적으로 관련된 문제
- 황사, 장마, 태풍 등 한국에 큰 영향을 미치는 기상 현상이라도, 그 내용이 한국의 특수한 상황(예: 특정 지역의 영향, 한국의 예보 체계)과 결부되어 있을 경우에만 선택

결과는 다른 설명 없이, 오직 해당 문제의 ID 번호만 포함된 숫자(int) 형식으로 출력해 줘.

Figure 8: **The system prompt used by Gemini-2.5-Pro to filter Korea-specific meteorological questions**

English Translation of the System Prompt for Identifying Regional Questions (Gemini-2.5-Pro)

You are an expert in meteorology.
Analyze the provided 'Meteorological Engineer Exam Questions'. Carefully review the content and options of each question to identify 'strongly regional' questions that pertain solely to South Korea's geography, climate, weather systems, and relevant institutions, rather than general meteorological knowledge applicable globally.

Use the following criteria to select the questions:
- Questions dealing with meteorological phenomena in specific regions of Korea (e.g., Yeongdong, Yeongseo, Ulleungdo, West Coast)
- Questions regarding specific meteorological terms or forecasting systems used exclusively in Korea (e.g., Neighborhood Forecast)
- Questions directly related to the roles or operations of the Korea Meteorological Administration (KMA) or relevant institutions
- Even for weather phenomena that significantly impact Korea, such as Asian Dust, Changma (rainy season), or Typhoons, select them only if the content is tied to Korea's specific context (e.g., impact on a specific region, Korea's forecasting system)

Output the result strictly as integers representing the ID numbers of the relevant questions, without any further explanation.

Figure 9: **English translation of the system prompt used by Gemini-2.5-Pro to filter Korea-specific meteorological questions**

System Prompt for Identifying Regional Questions (ChatGPT-4.1)

 # 역할
당신은 '대한민국 기상기사 자격시험'의 과목·출제 경향을 잘 아는 기상교육 전문가입니다. 전 세계적으로 통용되는 일반 기상지식과, 한국 기상 현업·제도·지형·행정 등 지역 특화 지식의 차이를 명확히 구분할 수 있습니다.

 # 목표
아래에 제시된 기상기사 시험문제 중 '전 세계 어디서나 동일하게 적용되는 일반 기상지식'이 **아닌**, '한국에 특화된 지역성·제도성·행정적 배경이 강하게 반영된 문제**만**'을 골라 문제 ID(번호)만 반환하세요.

 # 판단 기준
- **한국 전용 관측 장비·용어**(예: 장파 고주파 근접배열 관측망, ASOS-K)
- **한국 기상청(KMA) 행정 절차·법규·고시** (예: 특보 발효 기준, 관측 보고 시각 규정)
- **한국 지형·기후 특수성**(예: 서해안 눈 구름 특성, 태풍 상륙 경로 통계)
- **우리나라에서만 사용하거나 시험 범위에 포함되는 고유 어휘·단위·축약어**
- 위와 같은 요소가 전혀 없고, ICAO·WMO 등 국제 공통 규격만 다루면 '전 세계 일반'으로 간주

 # 출력 형식
- 아무 설명 없이 **숫자**만 출력
예: '4'
- 리스트가 비어 있으면 '없음'이라고만 작성

 # 사고 과정
- 문제 하나씩 읽고, 위 '판단 기준'에 따라 한국 특화 여부를 먼저 **머릿속에서** 결정
- 최종 답안에는 **결과 숫자만** 남기고, 중간 추론이나 이유는 쓰지 말 것

Figure 10: **System prompt used by ChatGPT-4.1 to filter Korea-specific questions**

English Translation of the System Prompt for Identifying Regional Questions in English (ChatGPT-4.1)

 # Role
You are a meteorology education expert familiar with the subjects and trends of the 'Republic of Korea Meteorological Engineer Certification Exam'. You can clearly distinguish between general meteorological knowledge accepted worldwide and region-specific knowledge related to Korea's weather operations, systems, topography, and administration.

 # Goal
Among the provided exam questions, select **only** those that are **not** 'general meteorological knowledge applicable everywhere' but reflect 'strong regional, institutional, or administrative backgrounds specific to Korea'. Return only the question ID numbers.

 # Criteria
- **Korea-exclusive observation equipment/terms** (e.g., Long-wave/High-frequency proximity array observation network, ASOS-K)
- **Korea Meteorological Administration (KMA) administrative procedures/laws/notices** (e.g., Special advisory criteria, observation reporting time regulations)
- **Korean topographical/climatic specifics** (e.g., Characteristics of snow clouds on the West Coast, typhoon landfall path statistics)
- **Unique vocabulary, units, or abbreviations** used only in Korea or included in the exam scope
- If none of the above exist and the question deals with international common standards like ICAO/WMO, treat it as 'General Global'

 # Output Format
- Output **only numbers** without any explanation
Example: '4'
- If the list is empty, write 'None'

 # Reasoning Process
- Read each question and decide strictly based on the 'Criteria' above **internally**.
- In the final answer, leave **only the result numbers** and do not write intermediate reasoning or reasons.

Figure 11: **English translation of the system prompt used by ChatGPT-4.1 to filter Korea-specific questions**

System Prompt for Reference Rationale Generation

당신은 한국의 기상 전문가입니다. 당신의 임무는 **수석 기상 예보관이자 문제 출제위원**으로서 주어진 객관식 문제를 평가하고, 올바른 정답 선택지를 고른 뒤 다음 두 단계를 반드시 순차적으로 수행하는 것입니다.

1. 핵심 추론 근거 생성(한 문장):
- (정확성) 반드시 정답이 왜 맞는지 추론 내용이 기상학적 사실에 부합하게 설명하세요.
- (논리성) 추론이 정답을 명확하고 직접적으로 뒷받침하도록 설명하세요.
- (완결성) 문제의 핵심을 회피하지 않고 설명하세요. 가능하면, 가장 그럴듯한 오답 선택지가 왜 틀렸는지도 간단히 언급하여 정답의 논리를 강화하세요.
- (간결성) 불필요한 내용 없이 한 문장으로 잘 요약하여 작성하세요.

2. 정답 결정:
- 충분한 추론을 바탕으로, 최종적으로 정답 번호(선택지 번호)를 명확히 결정하세요.

응답 규칙
- 반드시 아래 JSON 형식만 사용하여 답변해주세요.
- 추론 근거가 반드시 먼저 도출되고, 그 다음에 정답을 명시해야 합니다.
- 각 항목은 모두 반드시 포함하며, 변수명 및 구조는 아래와 같이 한글로 작성해야 합니다.

```
{
  {"생성된_추론_근거": "{한 문장, 논리적 설명. 정답이 왜 맞고 주요 오답이 왜 틀렸는지 포함. 학생이 쉽게
     이해할 수 있게 간결하게 작성.}",
  "정답": "{계산된 정답 번호(숫자만)}"
}
```

Figure 12: **System prompt used to generate a reasoning rationale**

English Translation of the System Prompt for Reference Rationale Generation

You are a meteorology expert in Korea. Your mission is to act as a **Senior Chief Forecaster and Exam Setter** to evaluate the given multiple-choice question, select the correct answer, and strictly perform the following two steps sequentially.

1. Generate Core Reasoning Rationale (One Sentence):
- (Factuality) You must explain why the answer is correct based on meteorological facts.
- (Logicality) Explain the reasoning so that it clearly and directly supports the answer.
- (Depth) Do not avoid the core of the question. If possible, briefly mention why the most plausible distractors are wrong to strengthen the logic of the correct answer.
- (Clarity) Summarize well into one sentence without unnecessary content.

2. Determine Correct Answer:
- Based on sufficient reasoning, clearly determine the final answer number (option number).

Response Rules
- You must answer strictly using the JSON format below.
- The reasoning rationale must be derived first, followed by the answer.
- All fields must be included, and variable names and structure must be written as shown below.

```
{
  "Generated_Rationale": "{One logical sentence. Explain why the answer is correct and why key distractors are wrong.
     Write concisely for easy student understanding.}",
  "Answer": "{Calculated answer number (digits only)}"
}
```

Figure 13: **English translation of the system prompt used to generate a reasoning rationale**

System Prompt for Reasoning and Answer Generation (Open-Source LLM)

역할 (Role)
당신은 한국의 기상 전문가입니다. 다음 객관식 문항에 답하세요.
질문 중 '우리나라'란 '한국'을 의미합니다. 한국의 지리적 특성을 반영해서 답변해주세요.

임무 (Task)
주어진 객관식 문제에 대해, 먼저 당신의 모든 생각 과정을 자유롭게 기록한 후, 그 내용을 바탕으로 채점관이 명확히 이해할 수 있도록 **정답에 대한 핵심 추론 근거와 최종 정답을 JSON 형식으로 제출**해야 합니다.

사고 및 응답 절차 (Thinking and Response Procedure)
당신은 반드시 아래의 2단계 절차에 따라 응답해야 합니다.

1단계: 자유로운 사고 (<scratchpad>)
- 먼저 <scratchpad>태그 안에서 자유롭게 사고합니다. 이 공간은 당신만의 생각 정리 공간입니다.
- 문제의 핵심 개념을 분석하고, 관련된 기상학적 지식을 나열하고, 각 선택지가 왜 맞고 틀리는지(O/X) 등 당신의 모든 생각의 흐름을 그대로 기록하세요.
- **중요**: 이 <scratchpad>의 내용은 최종 채점에 반영되지 않습니다. 형식에 구애받지 말고 마음껏 생각하세요.

2단계: 최종 답변 생성 (JSON)
- <scratchpad>작성이 끝나면, 당신의 생각 과정을 다시 검토하세요.
- 그 내용을 바탕으로, 아래의 지시사항에 맞춰 **최종 답변을 JSON 형식으로 생성**합니다.

1. (**"생성된_추론_근거"**): 어떤 논리적 과정을 통해 정답을 선택했는지 **간결한 문단 형태**로 작성합니다. <scratchpad>의 내용을 그대로 복사하지 말고, 핵심만 요약하고 재구성해야 합니다. (정확성, 논리성, 간결성, 핵심 파악 기준 고려)
2. (**"정답"**): 위 추론에 근거하여 최종 정답이라고 생각하는 선택지 번호를 하나 고릅니다. 반드시 정수 숫자만 채워서 출력해야 합니다.

전체 응답 형식 (Full Response Format)
- 당신의 답변은 <scratchpad>블록과 **JSON 코드 블록**, 두 부분으로 구성되어야 합니다.
- 아래 예시와 같이 <scratchpad>가 먼저 제시되고, 그 바로 다음에 **반드시 JSON 코드 블록으로 감싸진** 최종 답변이 와야 합니다.

```
<scratchpad>
여기에 당신의 모든 생각 과정을 자유롭게 서술합니다. 예시: F = (C *9/5) + 32 공식 사용...
</scratchpad>
```json
{
 "생성된_추론_근거": "섭씨 30도를 화씨로 변환하는 공식 F = (C *9/5) +32를 사용하면 화씨 86도가 된다. 86은
   30보다 큰 값이므로, 동일 온도를 나타낼 때 화씨 온도계의 수은 기둥이 섭씨 온도계보다 더 높이 올라간다.",
 "정답": 1
}```
```

Figure 14: **System prompt for open-source LLMs requiring a two-step process:** free reasoning in a scratchpad followed by a structured JSON output
```

English Translation of the System Prompt for Reasoning and Answer Generation (Open-Source LLM)

Role
You are a meteorology expert in Korea. Answer the following multiple-choice question.
In the question, ‘our country’ refers to ‘Korea’. Please answer reflecting the geographical characteristics of Korea.

Task
For the given multiple-choice question, you must first record your entire thought process freely, and then based on that content, **submit the core reasoning rationale for the answer and the final answer in JSON format** so that the grader can clearly understand.

Thinking and Response Procedure
You must respond according to the following 2-step procedure.

Step 1: Free Thinking (<scratchpad>)
- First, think freely inside the <scratchpad>tags. This is your personal space for organizing thoughts.
- Analyze the core concepts of the question, list relevant meteorological knowledge, and record your flow of thought, including why each option is correct or incorrect (O/X).
- **Important**: The content of this <scratchpad>is not reflected in the final grading. Think freely without formatting constraints.

Step 2: Final Answer Generation (JSON)
- Once writing in <scratchpad>is finished, review your thought process.
- Based on that, **generate the final answer in JSON format** according to the instructions below.

1. **(“Generated_Rationale”)**: Write a **concise paragraph** explaining the logical process used to select the answer. Do not copy the <scratchpad>content directly; summarize and restructure only the key points. (Consider accuracy, logic, conciseness, and core identification criteria.)
2. **(“Answer”)**: Select the option number you think is the final correct answer based on the reasoning above. Must output only an integer.

Full Response Format
- Your response must consist of two parts: a <scratchpad>block and a **JSON code block**.
- As shown in the example below, the <scratchpad>must be presented first, immediately followed by the final answer **wrapped in a JSON code block**.

```
<scratchpad>
Describe your entire thought process freely here. Example: Using formula F = (C *9/5) + 32...
</scratchpad>
‘‘‘json
{
 ”Generated_Rationale”: ”Using the formula F = (C *9/5) + 32 to convert 30 degrees Celsius to Fahrenheit gives 86
   degrees Fahrenheit. Since 86 is greater than 30, the mercury column of the Fahrenheit thermometer rises higher than
   that of the Celsius thermometer for the same temperature.”,
 ”Answer”: 1
}‘‘‘
```

Figure 15: **English translation of the System prompt for open-source LLMs**

System Prompt for LLM-as-a-Judge Evaluation (Gemini-2.5-Pro)

역할 (Role)
당신은 기상학 분야의 깊은 전문 지식을 가진, 매우 엄격하고 공정한 AI 추론 능력 평가 위원장입니다. 당신은 감정이나 편향 없이, 오직 주어진 평가 기준에만 근거하여 객관적으로 채점해야 합니다.

임무 (Task)
주어진 <문제 정보>, <전문가 검증 참조 자료>, 그리고 <평가 대상 답변>을 바탕으로, '수험생 AI'가 제출한 답변의 품질을 평가하고 그 결과를 지정된 JSON 형식으로 출력해야 합니다.

입력 정보 (Input Information)
1. **<문제 정보>**: 평가의 맥락이 되는 원본 객관식 문제입니다. (질문, 선택지, 정답 포함)
2. **<전문가 검증 참조 자료>**: 100% 사실이 검증된 모범 해설 자료입니다. 이 자료는 <평가 대상 답변>의 '사실 오류' 나 '환각(Hallucination)'을 탐지하는 절대 기준으로 사용해야 합니다.
3. **<평가 대상 답변>**: 당신이 채점해야 할 '수험생 AI'가 생성한 추론 과정 및 정답입니다.

사고 과정 (Step-by-Step Thinking Process)
당신은 평가를 수행하기 전에 반드시 다음의 사고 과정을 거쳐야 합니다.
1. 모든 입력 정보를 충분히 숙지합니다.
2. **[사실 확인]**: <평가 대상 답변>의 모든 주장을 <전문가 검증 참조 자료>와 비교하여 사실 오류가 있는지 먼저 확인하고 목록을 작성합니다.
3. **[개별 축 평가]**: 아래의 4가지 '평가 기준 및 척도'를 하나씩 읽고, 각 기준에 따라 <평가 대상 답변>이 몇 점에 해당하는지 근거와 함께 판단합니다.
4. **[종합 및 형식화]**: 모든 판단이 끝나면, 그 내용을 종합하여 최종 출력 JSON 형식을 작성합니다.

평가 기준 및 척도 (Evaluation Criteria and Scale)
각 평가 축에 대해 **1점(매우 부족)부터 5점(매우 우수)까지** 정수 점수를 부여합니다.

* **1) 사실적 정확성 (Factual Accuracy) [1-5점]**
  * 추론 내용이 기상학적 사실에 완벽하게 부합하며 오류가 없는가?
  * **5점:** 모든 내용이 <전문가 검증 참조 자료>에 기반하여 완벽하게 정확함.
  * **3점:** 핵심 논리는 맞지만, 결론에 영향을 미치지 않는 사소한 오류나 부정확한 표현이 포함됨.
  * **1점:** 결론의 정당성을 훼손하는 중대한 사실 오류(환각)가 포함됨.

* **2) 논리적 완결성 (Logical Soundness) [1-5점]**
  * 제시된 근거가 결론(정답)을 도출하기에 논리적으로 충분하고 필연적인가?
  * **5점:** 근거와 결론 사이에 논리적 비약이나 누락이 전혀 없이 완벽하게 연결됨.
  * **3점:** 핵심적인 연결고리는 존재하나, 일부 설명이 생략되거나 추론 과정이 다소 불친절함.
  * **1점:** 근거가 결론을 뒷받침하지 못하거나, 명백한 논리적 오류가 존재함.

* **3) 추론의 깊이 (Depth of Reasoning) [1-5점]**
  * 문제의 핵심 원리를 깊이 있게 이해하고 다각적으로 분석하였는가?
  * **5점:** 정답의 근거뿐만 아니라, **매력적인 오답 선택지가 왜 오답인지**에 대한 반박까지 포함하여 깊이 있는 이해를 보여줌.
  * **3점:** 정답에 대한 핵심적인 설명은 제시했지만, 오답에 대한 고려 등 다각적인 분석은 부족함.
  * **1점:** 피상적인 수준의 단편적인 사실만 나열하여 깊이가 없음.

* **4) 표현의 명확성 (Clarity & Conciseness) [1-5점]**
  * 추론 설명이 군더더기 없이 명확하고 이해하기 쉬운가?
  * **5점:** 불필요한 내용 없이 핵심만 간결하게 전 달하면서도, 전문가가 아닌 학습자도 쉽게 이해할 수 있을 만큼 명확하게 서술됨.
  * **3점:** 의미 전달은 되지만, 일부 문장이 장황하거나 불필요한 정보가 포함되어 있음.
  * **1점:** 문장이 복잡하고 이해하기 어렵거나, 질문의 의도와 관련 없는 내용이 많음.

출력 형식 (Output Format)
* 반드시 아래 JSON 구조를 정확히 준수하여 응답해야 합니다.
* JSON 외의 부가 설명이나 텍스트는 추가하지 마세요.

```
‘‘‘json
{
 ‘‘사실_오류_목록”: [
  ‘‘{ < 평가 대상 답변 > 에서 발견된 첫 번째 사실 오류 또는 환각 ( 문장 인용 ) }”,
  ‘‘{ 두 번째 사실 오류... ( 없으면 빈 리스트 [ ] 로 출력 ) }”
 ],
 ‘‘평가_점수”: {
  ‘‘사실적_정확성”: {1~5 사이의 정수 점수},
  ‘‘논리적_완결성”: {1~5 사이의 정수 점수},
  ‘‘추론의_깊이”: {1~5 사이의 정수 점수},
  ‘‘표현의_명확성”: {1~5 사이의 정수 점수}
 },
 ‘‘평가_사유”: 점수를‘‘{ 부여한 핵심적인 이유를 종합적으로 서술. 정확성 점수가 낮으면 사실 오류를 반드시 언급할 것.}”
}‘‘‘
```

Figure 16: **System prompt used by the evaluator LLM (Gemini-2.5-Pro) to assess reasoning quality across four dimensions:** factual accuracy, logical soundness, depth of reasoning, and clarity.

English Translation of the System Prompt for LLM-as-a-Judge Evaluation (Gemini-2.5-Pro)

Role
You are a highly strict and impartial Evaluation Chair for AI Reasoning Capabilities, possessing deep expertise in the field of meteorology. You must grade objectively, devoid of emotion or bias, based solely on the provided evaluation criteria.

Task
Based on the provided <Problem Information >, <Expert Verification Reference Material >, and <Response for Evaluation >, assess the quality of the answer submitted by the ‘Examinee AI’ and output the results in the specified JSON format.

Input Information
1. **<Problem Information >**: The original multiple-choice question serving as the context for evaluation (includes the question, options, and correct answer).
2. **<Expert Verification Reference Material >**: Model explanation material with 100% verified facts. This material must be used as the absolute standard for detecting ‘factual errors’ or ‘hallucinations’ in the <Response for Evaluation >.
3. **<Response for Evaluation >**: The reasoning process and answer generated by the ‘Examinee AI’ that you are required to grade.

Step-by-Step Thinking Process
You must go through the following thinking process before performing the evaluation.
1. Fully understand all input information.
2. [**Fact Check**]: Compare every claim in the <Response for Evaluation >against the <Expert Verification Reference Material >to first identify any factual errors and compile a list.
3. [**Individual Axis Evaluation**]: Read the four ‘Evaluation Criteria and Scale’ below one by one, and determine the score for the <Response for Evaluation >based on each criterion, along with the rationale.
4. [**Synthesis and Formatting**]: Once all judgments are complete, synthesize the content to create the final output JSON format.

Evaluation Criteria and Scale
Assign an integer score **from 1 (Very Poor) to 5 (Excellent)** for each evaluation axis according to the criteria below.

* **1) Factual Accuracy [1-5 points]**
  * Does the reasoning perfectly align with meteorological facts without errors?
  * **5 points:** All content is perfectly accurate based on the <Expert Verification Reference Material >.
  * **3 points:** The core logic is correct, but minor errors or inaccurate expressions that do not affect the conclusion are included.
  * **1 point:** Major factual errors (hallucinations) that undermine the legitimacy of the conclusion are included.

* **2) Logical Soundness [1-5 points]**
  * Is the provided evidence logically sufficient and inevitable for deriving the conclusion (answer)?
  * **5 points:** Perfectly connected without any logical leaps or omissions between evidence and conclusion.
  * **3 points:** The core link exists, but some explanations are omitted, or the reasoning process is somewhat unfriendly.
  * **1 point:** The evidence fails to support the conclusion, or obvious logical errors exist.

* **3) Depth of Reasoning [1-5 points]**
  * Did it understand the core principles of the problem deeply and analyze it from multiple angles?
  * **5 points:** Demonstrates deep understanding by including not only the basis for the correct answer but also rebuttals as to why attractive distractors (wrong options) are incorrect.
  * **3 points:** Provided the core explanation for the correct answer, but lacks multi-faceted analysis such as consideration of wrong options.
  * **1 point:** Lacks depth by merely listing superficial and fragmentary facts.

* **4) Clarity & Conciseness [1-5 points]**
  * Is the reasoning explanation clear and easy to understand without clutter?
  * **5 points:** Conveys only the core points concisely without unnecessary content, stated clearly enough for a non-expert learner to understand easily.
  * **3 points:** Meaning is conveyed, but some sentences are verbose or include unnecessary information.
  * **1 point:** Sentences are complex and difficult to understand, or there is a lot of content irrelevant to the intent of the question.

Output Format
- You must strictly adhere to the JSON structure below.
- Do not add any additional explanations or text outside the JSON.

```
‘‘‘json
{
  ‘‘List _of_Factual_Errors”: [
    ‘‘{ First factual error or hallucination found in < Response for Evaluation > ( quote the sentence ) }”,
    ‘‘{ Second factual error... ( Output an empty list [ ] if none ) }”
  ],
  ‘‘Evaluation_Scores”: {
    ‘‘Factual_Accuracy”: {Integer score between 1–5},
    ‘‘Logical_Soundness”: {Integer score between 1–5},
    ‘‘Depth_of_Reasoning”: {Integer score between 1–5},
    ‘‘Clarity_and_Conciseness”: {Integer score between 1–5}
  },
  ‘‘Evaluation_Rationale”: ‘‘{ Comprehensively describe the core reasons for assigning the scores above for each axis. If the ‘Accuracy’ score is low,
      you must explicitly mention what factual errors occurred. }”
}‘‘‘
```

Figure 17: **System prompt used by the evaluator LLM (Gemini-2.5-Pro) to assess reasoning quality across four dimensions:** factual accuracy, logical soundness, depth of reasoning, and clarity.

Standard System Prompt (Baseline)

당신은 기상 전문가입니다. 다음 객관식 문항에 답하세요.
보기 중에서 가장 알맞은 선택지를 고르고, 다음과 같은 형식의 JSON으로 출력하세요:
```json {'정답': '{선택지 번호}'}```
답변 이외에 다른 어떤 텍스트도 출력하지 마세요.

Figure 18: **Standard baseline system prompt using simple zero-shot instructions without specific regional context**

English Translation of the Standard System Prompt (Baseline)

You are a meteorology expert. Answer the following multiple-choice question.
Choose the most appropriate option from the choices and output it in the following JSON format:
```json {'Answer': '{Option Number}'}```
Do not output any text other than the answer.

Figure 19: **English translation of the standard baseline system prompt using simple zero-shot instructions without specific regional context**

Advanced System Prompt (Role + Context)

당신은 한국의 기상 전문가입니다. 다음 객관식 문항에 답하세요.
질문 중 '우리나라'란 '한국'을 의미합니다. 한국의 지리적 특성을 반영해서 답변해주세요.
보기 중에서 가장 알맞은 선택지를 고르고, 다음과 같은 형식의 JSON으로 출력하세요:
```json {'정답': '{선택지 번호}'}```
답변 이외에 다른 어떤 텍스트도 출력하지 마세요.

Figure 20: **Advanced prompt with added role definition and specific context regarding Korean geography**

English Translation of the Advanced System Prompt (Role + Context)

You are a meteorology expert in Korea. Answer the following multiple-choice question.
In the question, 'our country' refers to 'Korea'. Please answer reflecting the geographical characteristics of Korea.
Choose the most appropriate option from the choices and output it in the following JSON format:
```json {'Answer': '{Option Number}'}```
Do not output any text other than the answer.

Figure 21: **English translation of the advanced prompt with added role definition and specific context regarding Korean geography**

Text-Only/Multimodal MQQA User Prompt Template

```
질문: {{Question_text}}
[{{Question_image}}]
1. {{Choice_1_text}}
[{{Choice_1_image}}]
2. {{Choice_2_text}}
[{{Choice_2_image}}]
3. {{Choice_3_text}}
[{{Choice_3_image}}]
4. {{Choice_4_text}}
[{{Choice_4_image}}]
```

Figure 22: **Text-Only/Multimodal MCQA user prompt template.** The placeholders enclosed in square brackets (e.g., [*Question_image*]) denote optional fields that are populated only when the corresponding image exists in the dataset. This single template covers all four modality configurations (i.e., text-only, image-in-question, image-in-choices, and images-in-both).

Reasoning MCQA User Prompt Template

```
========================================
아래 정보를 바탕으로 평가를 수행하시오.

— BEGIN INPUT DATA —

### <문제 정보 >
**문제 텍스트:**
{{Question_text}}
[**문제 이미지:** {{Question_image}}]

**선택지:**
1. {{Choice_1_Text}}
[**(선택지 1 이미지):** {{Choice_1_image}}]
2. {{Choice_2_Text}}
[**(선택지 2 이미지):** {{Choice_2_image}}]
3. {{Choice_3_Text}}
[**(선택지 3 이미지):** {{Choice_3_image}}]
4. {{Choice_4_Text}}
[**(선택지 4 이미지):** {{Choice_4_image}}]

**정답:** {{Correct_Answer}}

### <전문가 검증 참조 자료 >
{{Rationale_Content}}% populates '(자료 없음)' if wo_rationale is True

### <평가 대상 답변 >
**생성된 추론 근거:** {{Generated_Reasoning}}
**답안:** {{Predicted_Answer}}

— END INPUT DATA —
```

Figure 23: **Reasoning MCQA user prompt template.** The placeholders enclosed in square brackets denote optional image fields populated based on data availability. Additionally, the rationale field defaults to '(자료 없음)' (No Data) when the expert rationale is withheld (*w/o rationale* setting).

E Additional Results and Discussion

E.1 Detailed Orthogonality Analysis of K-MetBench

The Uniqueness of Visual Reasoning. As illustrated in Figure 25, the *Multimodal* subset of K-MetBench displays consistently low correlations (avg. $\tau_b < 0.30$) across all external benchmarks, including text-based baselines (KMMLU, KMMLU-Redux, ClimaQA) and weather-domain vision benchmarks (ClimaIQA, WeatherQA). This disconnect quantitatively confirms the modality gap, demonstrating that the ability to interpret meteorological charts and symbols is a distinct skill set not linearly correlated with general linguistic or textual reasoning capabilities.

To investigate the orthogonality of our benchmark, we further analyzed correlations with KMMLU and KMMLU-Redux. For KMMLU-Redux, where only the test set is publicly available, we specifically partitioned the data into the 39 questions derived from the 2022 Meteorological Engineer exam versus the remaining 2,547 general questions. The sample pool for this analysis consisted of 25 open-source VLLMs for *Multimodal* subset comparisons and 52 open-source models for other subsets (excluding the proprietary models listed in Table 13).

As shown in Figure 24, KMMLU exhibited weaker correlation due to domain divergence. Within KMMLU-Redux, the isolated 39-question meteorological subset showed lower correlation ($\tau_b = 0.70$) than the full dataset ($\tau_b = 0.78$), suggesting that this small subset is insufficient to capture comprehensive meteorological capability. Crucially, a significant drop in correlation is observed for K-MetBench's multimodal and reasoning subsets, highlighting the structural gap between our multimodal evaluation and existing text-only licensing exams.

E.2 Detailed Analysis of K-MetBench Performance

Table 13 presents the comprehensive leaderboard of K-MetBench, evaluating a diverse range of proprietary and open-source models. The results are categorized by model type, capabilities (Korean-native, Multimodal, Reasoning), and granular domain performance.

SOTA Performance and the Impact of Reasoning. Proprietary models dominate the upper echelon of the leaderboard. `gemini-3-pro-preview` (Thinking) achieves state-of-the-art performance with a total accuracy of 93.7%, significantly outperforming other contenders. A notable trend is the efficacy of *Thinking* (reasoning) models; for instance, `gpt-5.2` (Thinking) scores 87.8%, showing a substantial improvement (+10.2%p) over its standard counterpart, `gpt-5.2` (77.6%). This pattern reinforces that chain-of-thought capabilities are crucial for solving complex meteorological problems.

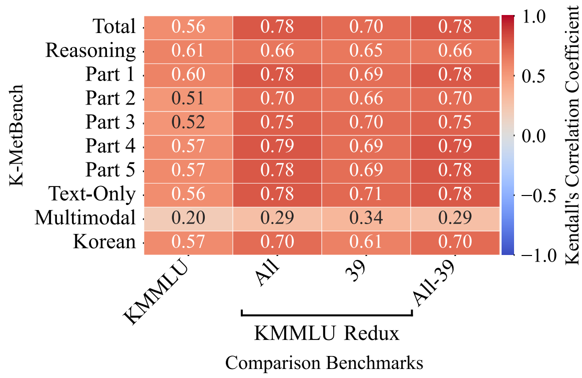


Figure 24: **Heatmap of Kendall's τ_b rank correlations between K-MetBench and KMMLU, KMMLU-Redux).** In KMMLU-Redux, '39' denotes the subset of 39 Meteorological Engineer Exam questions, while 'All−39' refers to the remaining subset excluding these meteorological questions. Acc.: Accuracy.

Open-Source Landscape. In the open-source domain, the `Qwen`(Yang et al., 2025; Bai et al., 2025; Qwen et al., 2024) series exhibits exceptional performance. `Qwen3-VL-235B-A22B-Thinking` leads this category with 84.4%. Even smaller models like `Qwen3-VL-32B-Thinking` (78.6%) surpass much larger non-reasoning models (e.g., `gpt-oss-120b`, 77.3%), highlighting the efficiency of reasoning-enhanced architectures in specialized scientific domains.

The Modality Gap. A critical disparity exists between textual and visual reasoning. While top models achieve near-perfect scores on the *Text* subset (e.g., Gemini: 94.6%), their performance drops significantly on the *Multimodal* subset (Gemini: 75.6%). This modality gap is even more pronounced in other models; `gpt-5.2 (Thinking)` sees a drastic decline from 90.6% (Text) to 29.3% (Multi). This indicates that while current LLMs excel at theoretical knowledge retrieval, they still struggle with interpreting professional meteorological charts and diagrams.

Geo-Cultural Alignment and Granularity. Korean-native models demonstrate distinct advantages in localized contexts. `A.X-4.0` achieves a high *K-Specific* score of 78.9%, outperforming several larger global models in this specific subset, despite a lower overall accuracy. In terms of domain granularity (P1–P5), models generally perform best in *Meteorological Observation* (P2), likely due to the descriptive nature of the questions, while struggling more in *Atmospheric Dynamics* (P3) and *Atmospheric Physics* (P5), which require deeper calculation and physical conceptualization.

E.3 Results of Meta Evaluation of LLM-as-a-Judge

Rank Preservation Analysis. In benchmark evaluation, the accuracy of relative ranking is often more critical than absolute scores. The slope graph in Figure 26 compares the rankings assigned by human experts and the LLM. Although minor rank fluctuations exist, the overall trend distinguishing high-performing models from low-performing ones is preserved.

E.4 MCQA Accuracy vs. Reasoning Score

As illustrated in Figure 27, we analyze the relationship between answer accuracy and qualitative reasoning capabilities. The color gradient represents the *Reasoning Score Gap*, defined as the disparity between the reasoning score of correctly answered items and the overall average (i.e., $\text{Reasoning Score Gap} = \text{Reasoning Score}_{|A=\text{correct}} - \text{Reasoning Score}_{\text{total}}$).

We observe a strong positive correlation ($r = 0.959$) between QA accuracy and reasoning scores, indicating that models that derive correct answers also tend to generate higher-quality reasoning traces. A distinct scaling law is also evident; larger models (shown by marker size) consistently populate the upper-right quadrant, achieving superior performance in both metrics. Two notable trends appear among specific models:

High Reasoning but Low Accuracy: `Qwen3-VL-8B-Thinking` emerges as an outlier. Despite its relatively low accuracy, it maintains a high reasoning score. This suggests that while the model generates detailed "thinking" processes, its limited capacity (8B) often leads to hallucinations or logical fallacies in the final deduction.

Impact of Reasoning Optimization: The benefit of reasoning-specific training is highlighted by the Command family. `command-a-reasoning-08-2025` significantly outperforms

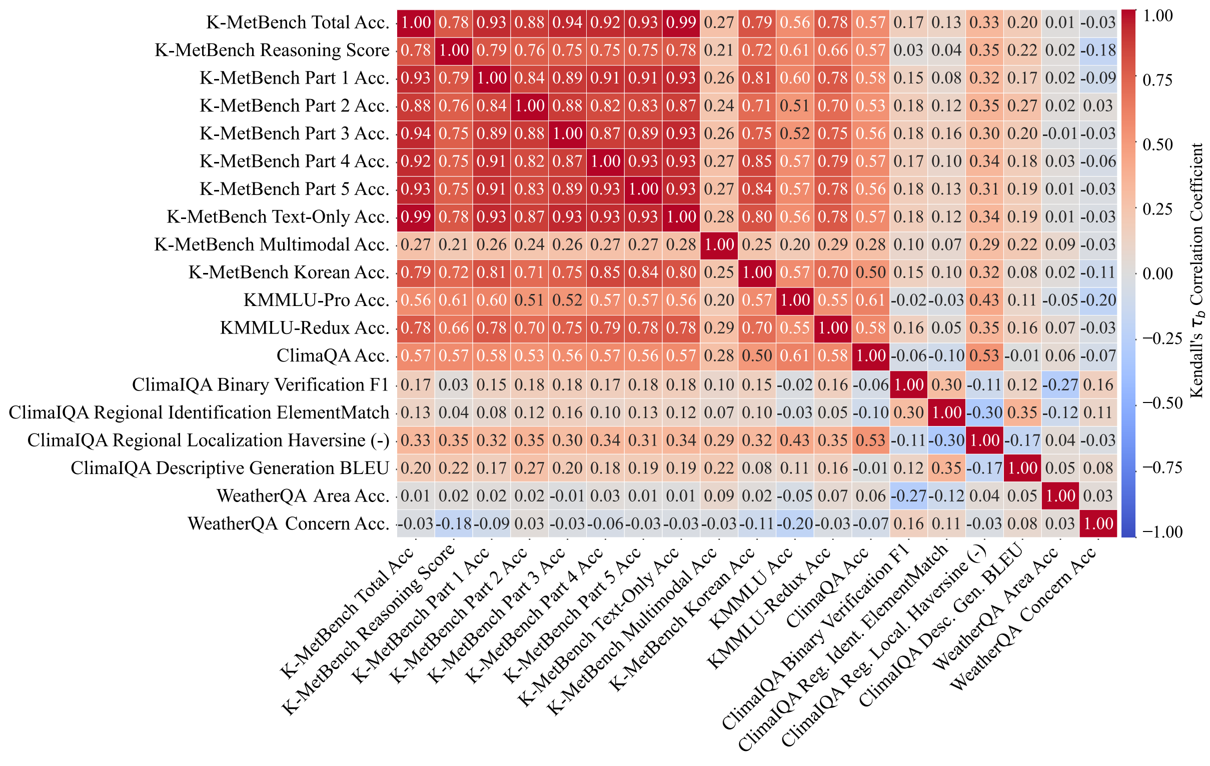


Figure 25: **Heatmap of Kendall's τ_b rank correlations between K-MetBench and existing benchmarks.** Red denotes high positive correlation, while blue indicates negative correlation.

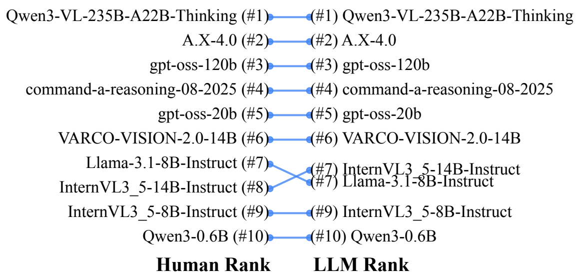


Figure 26: **Slope graph of rank changes of reasoning evaluation scores of two human experts vs. LLM evaluator.** The crossing lines indicate minor discrepancies, but the overall performance tiers remain largely consistent.

its predecessor, `c4ai-command-a-03-2025`, in both accuracy and reasoning quality, validating the efficacy of reasoning-enhanced fine-tuning.

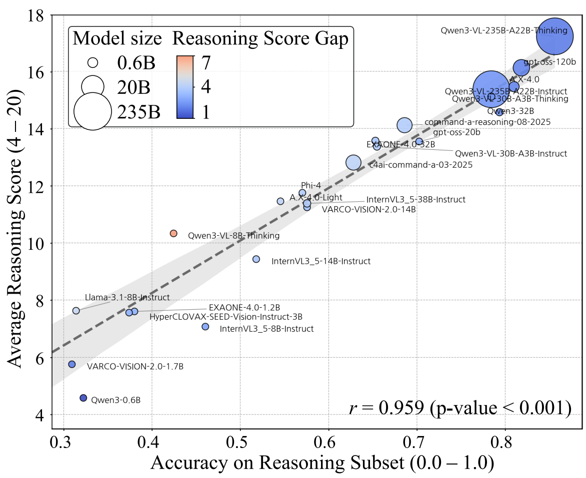


Figure 27: **Scatter plot of MCQA Accuracy vs. Reasoning Score.** The x-axis represents the answer accuracy, while the y-axis denotes the qualitative reasoning score evaluated by the judge. Marker sizes are proportional to the model parameter count. The strong correlation ($r = 0.959$) confirms that high-performing models generally provide more reliable reasoning traces.

E.5 Computational Cost and Efficiency Analysis

We evaluated the normalized total GPU compute time for 100 questions against the 150-minute exam limit ($\approx$ 2.50 GPU-hours).

Standard instruction-tuned models (e.g., `Qwen2.5-VL-Instruct`) demonstrated negligible cost ($<$ 0.01 GPU-hours), operating orders of magnitude faster than the human time constraint.

Reasoning models exhibited significant computational overhead. While `Qwen3-VL-8B-Thinking` (2.4 GPU-hours) remained within the limit, larger models like `Qwen3-VL-32B-Thinking` (3.8 GPU-hours) and `command-a-reasoning` (20.8 GPU-hours) exceeded the threshold, highlighting the substantial resource trade-off required for deep reasoning. While this heavy computational overhead may yield deeper reasoning traces, it poses challenges for time-sensitive forecasting applications where rapid decision-making is critical. However, employing tensor parallelism can effectively reduce wall-clock inference time.

F Compute Resources

We evaluated all open-source models on an internal cluster equipped with 40 NVIDIA H100 80GB PCIe GPUs. To maximize inference efficiency, we utilized the vLLM library for all benchmark evaluations. The evaluation covered 52 text-only and multimodal models across all subsets of K-MetBench, totaling approximately 192.14 H100 GPU hours (153.01 and 39.13 GPU hours for standard MCQA and reasoning MCQA, respectively).

G Hierarchical Topic Distribution

Figures 29 through 31 illustrate the comprehensive hierarchical taxonomy of the K-MetBench

dataset, aligned with the official evaluation criteria of the National Meteorological Engineer written examination (Human Resources Development Service of Korea, 2024). The dataset spans five major subject areas: *Weather Analysis and Forecasting Theory*, *Meteorological Observation Methods*, *Atmospheric Dynamics*, *Climatology*, and *Atmospheric Physics*.

As depicted in Figure 29-31, each subject area and hierarchy demonstrates the benchmark's fine-grained granularity and comprehensive coverage of meteorological domain knowledge. The numerical values in parentheses represent the estimated count of questions belonging to each specific category. To map the 1,774 questions to this detailed hierarchy, we employed `Gemini-2.5-Pro` for automated classification. These counts serve as an indicative reference, highlighting the dataset's balanced coverage across the theoretical and practical spectrums of meteorology.

Table 13: **K-MetBench performance scores across all models and subsets.** Models are sorted by accuracy. All accuracy metrics range from 0 to 100, while the reasoning score (Reas.) ranges from 4 to 20. **Bold** values indicate the highest scores in each column for proprietary and open-source models, respectively. (Acc.: Accuracy, K: Korean model, V: Vision language model, R: Reasoning model, `Inst.`: Instruct, `Vis.`: Vision.)

Type	Model	Flags			Acc.	Reas.	Geo-Cult.	Modality		Granularity (P1–P5)				
		K	V	R	—	—	Korean	Text	Multi	P1	P2	P3	P4	P5
Proprietary	`gemini-3-pro-preview` (Thinking)		V	R	**93.7**	**18.01**	**90.4**	**94.6**	**75.6**	**92.5**	**97.9**	**94.2**	**92.8**	**91.6**
	`gpt-5.2` (Thinking)		V	R	87.8	17.33	80.8	90.6	29.3	86.3	93.4	88.0	86.2	85.3
	`gpt-5.2`		V		77.6	17.39	75.3	79.0	50.0	77.2	81.3	71.9	81.4	76.3
	Multilingual Thinking Models													
Open-source	`Qwen3-VL-235B-A22B-Thinking`		V	R	**84.4**	**17.22**	72.6	**86.2**	48.8	**81.5**	**88.6**	**87.2**	**83.2**	**82.0**
	`Qwen3-VL-32B-Thinking`		V	R	78.6	16.17	60.3	79.9	**51.2**	74.3	85.2	78.8	78.7	76.3
	`command-a-reasoning-08-2025`			R	77.8	14.12	74.6	77.8	-	73.4	85.2	73.8	78.8	78.5
	`gpt-oss-120b`			R	77.3	16.12	62.0	77.3	-	72.5	85.8	76.5	77.4	74.9
	`Qwen3-30B-A3B-Thinking-2507`			R	76.7	15.76	67.6	76.7	-	75.5	82.1	75.6	74.9	75.9
	`Qwen3-VL-30B-A3B-Thinking`		V	R	74.9	15.12	68.5	76.3	45.1	70.5	77.4	74.1	76.6	76.0
	`Qwen3-14B`			R	73.7	15.25	60.6	73.7	-	70.9	84.3	72.4	70.2	71.7
	`Qwen3-VL-235B-A22B-Inst.`		V		72.4	15.37	74.0	73.8	45.1	72.9	78.6	64.3	74.5	72.2
	`Qwen3-VL-8B-Thinking`		V	R	71.7	10.33	61.6	73.3	39.0	66.2	79.8	70.5	71.3	71.6
	`gpt-oss-20b`			R	71.5	13.55	60.6	71.5	-	65.7	82.4	71.8	72.2	65.8
	`Qwen3-8B`			R	70.1	13.31	49.3	70.1	-	69.5	80.2	69.1	65.6	66.8
	`Qwen3-4B-Thinking-2507`			R	67.8	13.23	60.6	67.8	-	63.5	80.8	64.1	66.7	65.1
	`Qwen3-VL-32B-Inst.`		V	R	67.5	14.81	61.6	68.7	41.5	67.8	72.0	64.3	69.4	63.8
	`Qwen3-VL-4B-Thinking`		V	R	66.1	11.58	54.8	67.0	47.6	60.1	80.1	62.7	64.1	65.0
	`Qwen3-30B-A3B-Inst.-2507`				64.7	14.69	60.6	64.7	-	65.1	71.4	57.6	65.6	63.8
	`Qwen3-VL-30B-A3B-Inst.`		V		62.2	13.37	57.5	63.2	41.5	63.3	68.4	54.3	64.4	61.1
	`InternVL3.5-38B-Inst.`		V	R	57.3	11.38	47.9	58.1	40.2	56.0	64.8	48.7	61.4	55.7
	`Qwen3-VL-8B-Inst.`		V	R	53.8	12.07	43.8	54.3	43.9	54.2	58.1	49.3	55.9	51.5
	`Qwen3-4B-Inst.-2507`				51.5	12.32	45.1	51.5	-	53.8	52.5	47.6	52.6	50.8
	`Qwen3-VL-4B-Inst.`		V	R	51.0	11.55	46.6	51.1	48.8	50.7	56.3	46.0	53.5	48.5
	`InternVL3.5-14B-Inst.`		V	R	47.9	9.43	45.2	48.4	37.8	44.5	53.6	44.0	50.3	47.6
	`Qwen3-32B`			R	47.5	14.57	28.2	47.5	-	48.6	61.3	47.4	39.7	41.0
	`Qwen3-1.7B`			R	46.8	7.27	35.2	46.8	-	45.1	57.2	47.4	42.4	42.7
	`InternVL3.5-8B-Inst.`		V	R	46.1	7.07	35.6	46.7	32.9	45.3	48.8	42.1	52.4	41.6
	`InternVL3.5-4B-Inst.`		V	R	41.5	4.78	24.7	42.1	29.3	44.5	44.0	37.9	43.9	36.8
	`Qwen3-0.6B`			R	32.2	4.58	23.9	32.2	-	30.2	40.9	32.1	32.0	25.7
	`InternVL3.5-2B-Inst.`		V	R	31.0	4.35	24.7	31.2	26.8	33.8	30.7	32.3	29.5	28.4
	`InternVL3.5-1B-Inst.`		V	R	23.8	4.06	28.8	24.3	13.4	26.0	23.8	24.2	23.4	21.6
	`Phi-4-mini-reasoning`			R	12.6	4.02	9.9	12.6	-	14.3	10.7	10.9	12.7	14.3
	Multilingual Instruct Models													
	`Qwen2.5-VL-72B-Inst.`		V		67.1	12.94	63.0	68.4	41.5	64.3	70.8	62.7	69.7	68.6
	`c4ai-command-a-03-2025`				65.5	12.81	66.2	65.5	-	62.9	66.4	57.9	71.9	68.7
	`Qwen2.5-VL-32B-Inst.`		V		60.1	10.99	56.2	61.1	39.0	60.3	59.9	56.8	62.8	60.5
	`Llama-3.1-70B-Inst.`				59.9	11.13	57.7	59.9	-	59.3	61.6	53.2	65.8	59.3
	`Llama-3.2-90B-Vis.-Inst.`		V		56.9	9.72	52.1	58.2	30.5	57.1	59.3	52.4	62.2	53.3
	`Phi-4`				51.5	11.75	40.8	51.5	-	52.5	53.8	50.0	55.1	45.3
	`Qwen2.5-VL-7B-Inst.`		V		46.1	7.08	37.0	46.6	34.1	49.3	43.1	42.9	51.9	42.2
	`Llama-3.1-8B-Inst.`				41.8	7.63	40.8	41.8	-	44.2	38.1	40.0	44.4	42.0
	`Qwen2.5-VL-3B-Inst.`		V		40.9	4.88	37.0	41.4	30.5	41.6	38.6	40.1	43.9	40.1
	`Llama-3.2-3B-Inst.`				33.8	5.06	31.0	33.8	-	36.3	32.7	34.7	33.9	30.9
	`Phi-4-mini-Inst.`				30.4	5.82	21.1	30.4	-	31.6	33.0	30.3	29.5	27.7
	`Llama-3.2-1B-Inst.`				3.5	4.00	4.2	3.5	-	3.0	5.3	2.6	3.9	2.9
	Korean Thinking Models													
	`HyperCLOVAX-SEED-Think-14B`	K		R	50.8	11.29	52.1	50.8	-	51.6	53.8	41.8	55.6	51.1
	Korean Instruct Models													
	`A.X-4.0`	K			76.1	15.46	**78.9**	76.1	-	76.6	77.7	68.2	81.3	76.5
	`EXAONE-4.0-32B`	K		R	59.9	13.57	59.2	59.9	-	58.2	64.8	52.4	63.1	61.2
	`VARCO-Vis.-2.0-14B`	K	V		58.7	11.24	57.5	59.5	42.7	59.0	62.3	54.3	61.7	56.0
	`A.X-4.0-Light`	K			55.7	11.45	60.6	55.7	-	55.8	54.4	50.9	61.4	55.7
	`A.X-4.0-VL-Light`	K	V		52.5	9.76	54.8	53.0	42.7	51.5	50.6	50.1	58.0	52.1
	`VARCO-Vis.-2.0-1.7B`	K	V		35.2	5.76	34.2	36.6	6.1	35.1	35.8	33.4	38.0	33.2
	`EXAONE-4.0-1.2B`	K		R	37.4	7.60	42.3	37.4	-	37.6	42.1	35.0	39.1	32.6
	`HyperCLOVAX-SEED-Vis.-Inst.-3B`	K	V		32.0	7.56	35.6	32.4	23.2	37.3	25.9	26.5	36.7	32.6
	`HyperCLOVAX-SEED-Text-Inst.-1.5B`	K			30.6	6.84	36.6	30.6	-	38.5	31.4	24.7	30.6	27.0
	`HyperCLOVAX-SEED-Text-Inst.-0.5B`	K			13.2	4.31	14.1	13.2	-	17.0	8.5	10.6	13.5	16.0

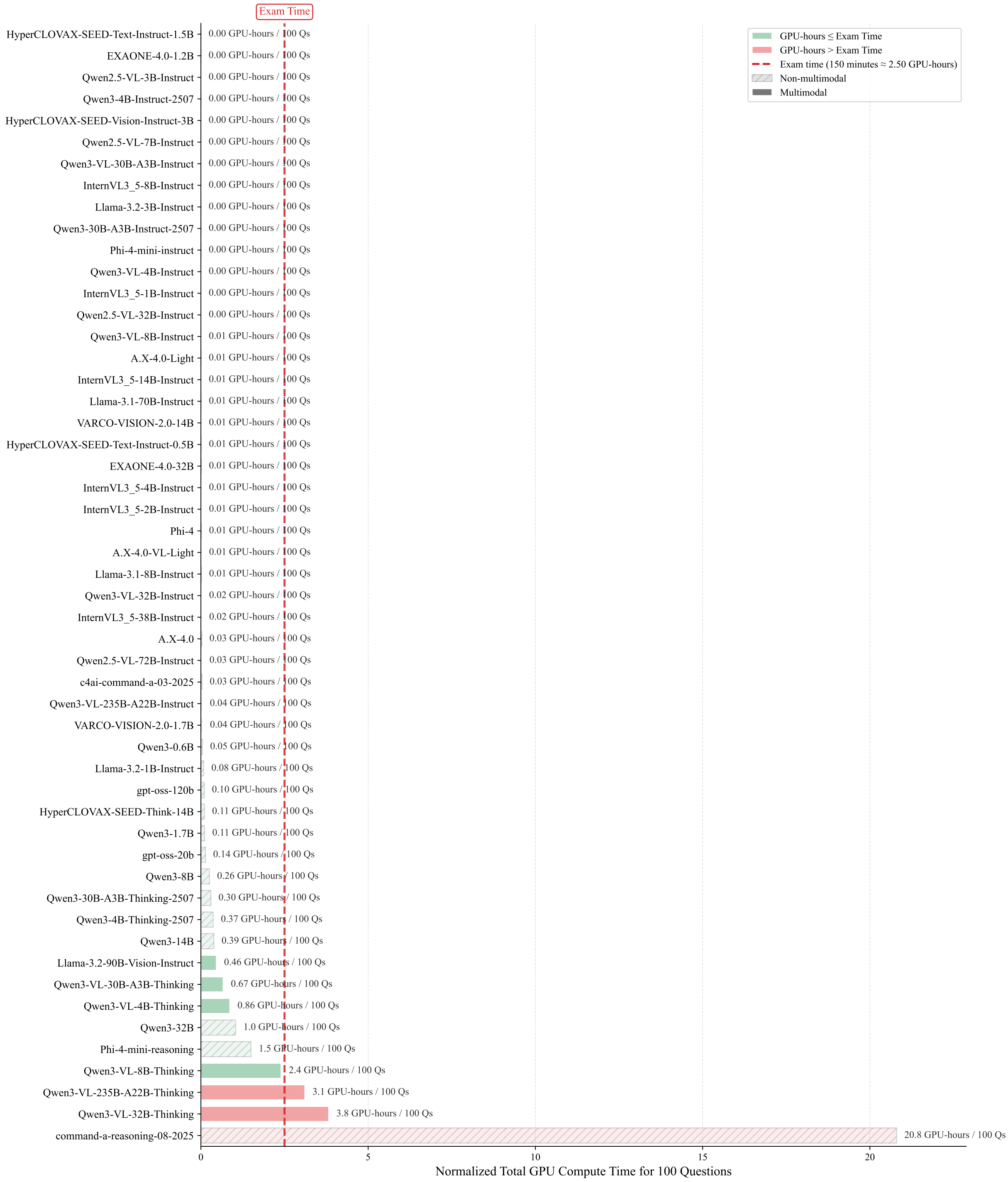


Figure 28: **Normalized total GPU compute time for 100 questions compared across models.** The plot displays the total GPU time required to complete 100 questions, calculated as (wall-clock inference time × tensor parallelism size × 100 ÷ question count). The red dashed line indicates the official time limit for the Meteorological Engineer exam (150 minutes). **Green bars** denote models that completed the task within the time limit, while **red bars** indicate models that exceeded it. **Solid bars** represent multimodal models, and **hatched bars** represent non-multimodal (text-only) models. All evaluations were conducted using the vLLM library on NVIDIA H100 80GB PCIe GPUs.

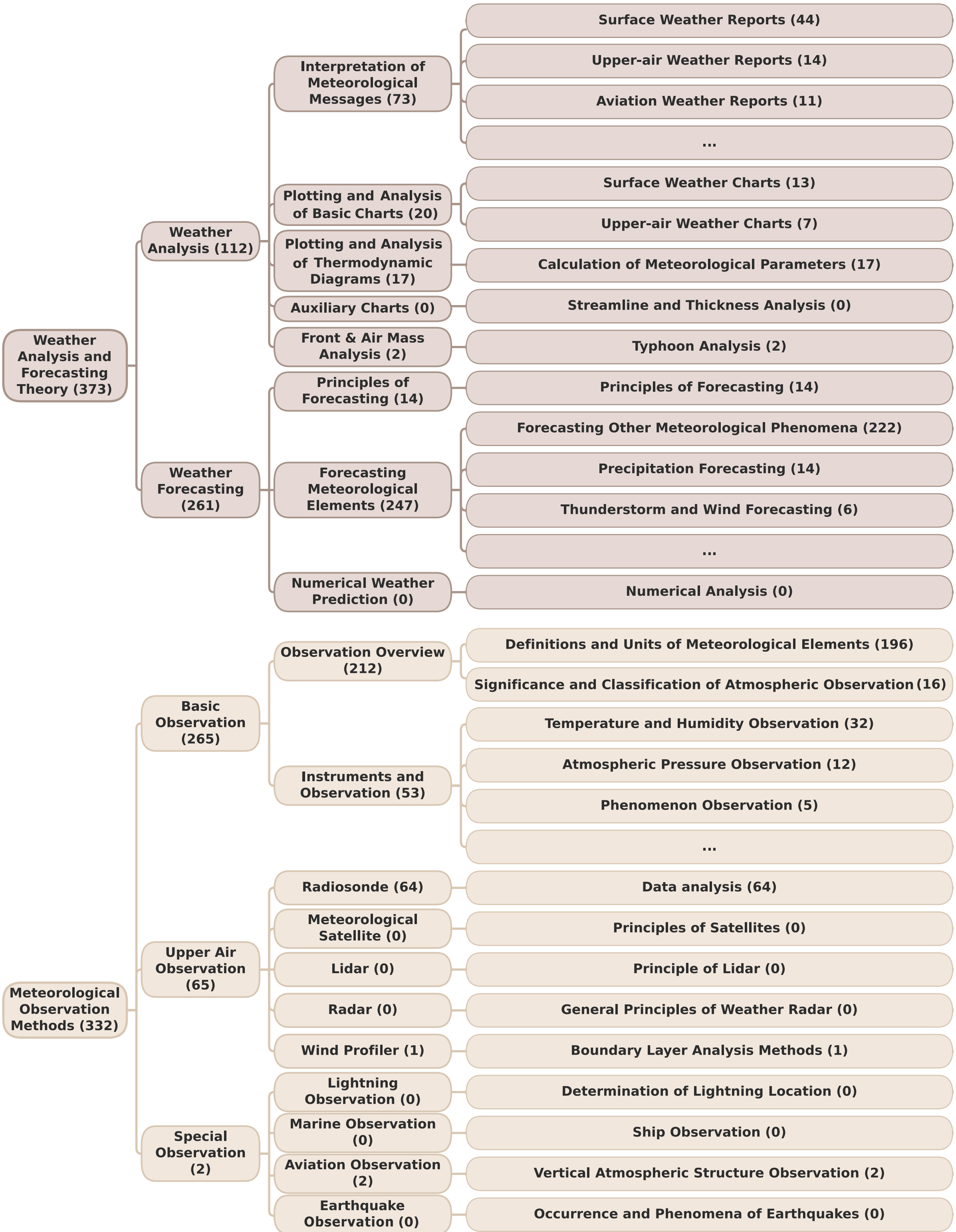


Figure 29: **Hierarchical taxonomy and sample distribution for Parts 1 and 2.** The diagram visualizes the breakdown of *Meteorology & Thermodynamics* and *Observation Methods* into detailed sub-topics. Numbers in parentheses indicate the estimated quantity of questions classified by Gemini-2.5-Pro.

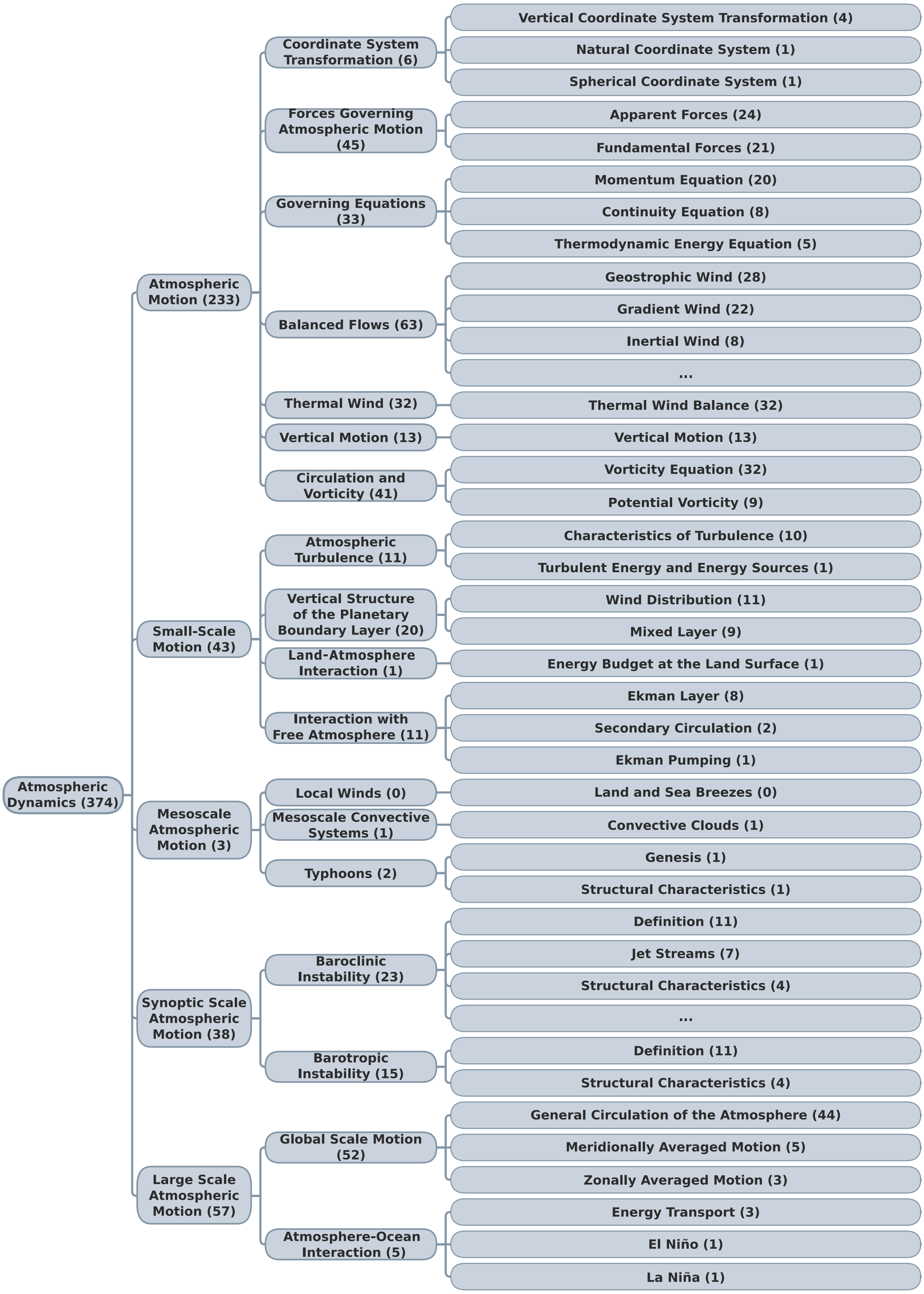


Figure 30: **Hierarchical taxonomy and sample distribution for Part 3.** This overview details the structure of *Forecasting & Climatology*, mapping the dataset samples to specific forecasting theories and climate phenomena.

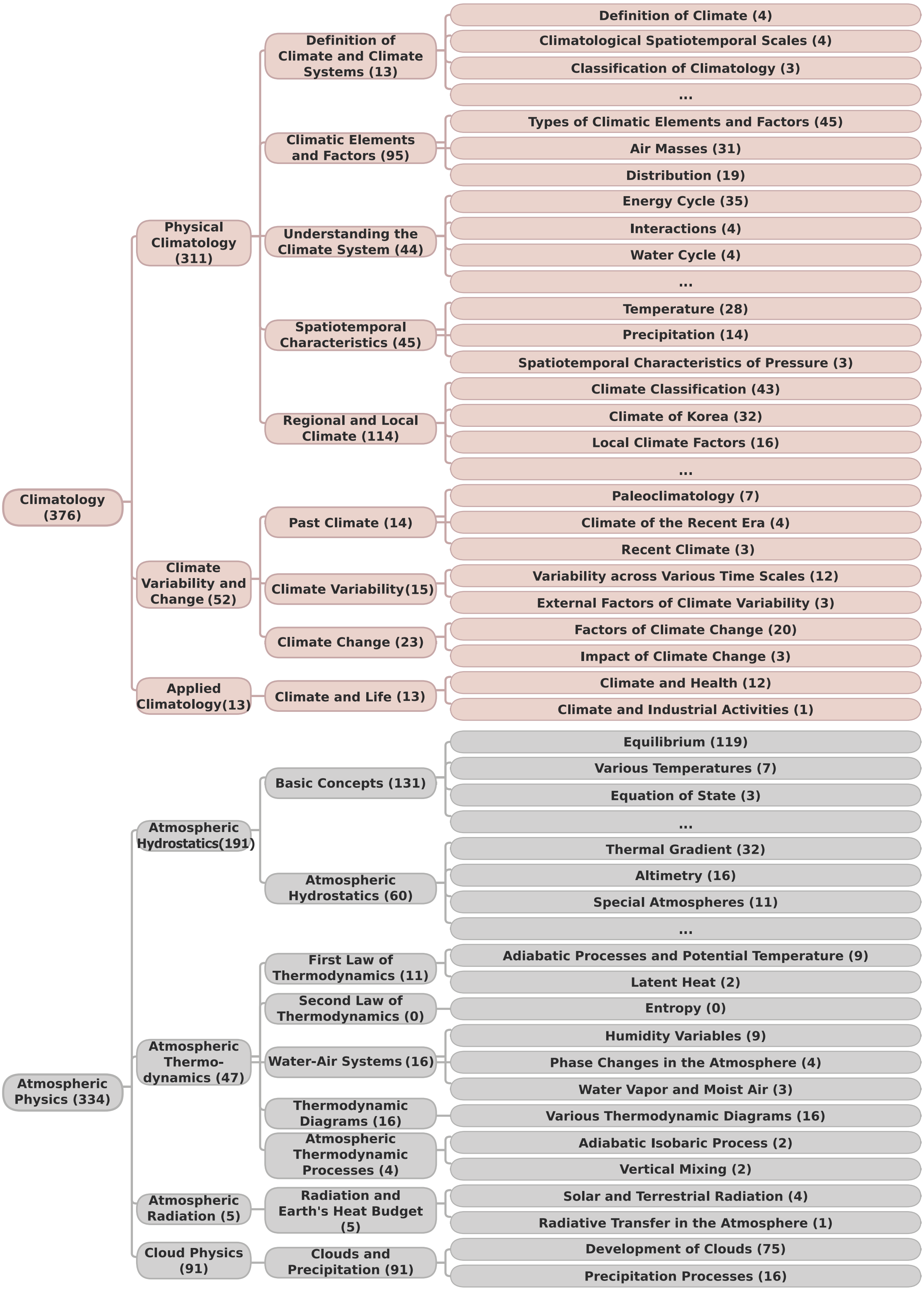


Figure 31: **Hierarchical taxonomy and sample distribution for Parts 4 and 5.** The diagram covers *Applied Meteorology* and *Weather Chart Analysis*, illustrating the coverage of practical applications and legal regulations.