SEAICE-BENCH: A BENCHMARK FOR SEA-ICE PREDICTION AND ANALYSIS CAPABILITIES OF LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sea ice analysis requires models to connect remote-sensing imagery, long-horizon geophysical records, and specialized scientific knowledge, yet existing large language model (LLM) benchmarks do not evaluate this combination of capabilities in a sea-ice-specific setting. We introduce SEAICE-BENCH, a benchmark for assessing LLMs and multimodal LLMs on three evidence sources: native Sentinel-2 imagery, historical numerical records, and domain literature. SEAICE-BENCHcomprises 136,500 multispectral satellite images across full-scene and crop-based splits, monthly sea ice extent records from November 1978 to January 2026, weekly sea ice motion vectors from November 1978 to December 2024, and a curated corpus of more than 10,000 sea-ice research articles and textbooks that yields 53,785 question-answer pairs spanning six task families and four answer formats. We evaluate ten frontier models, seven proprietary and three open-weight, in a unified zero-shot protocol on sea ice concentration estimation, next-step sea ice extent and motion vector reasoning, and literature-grounded question answering. The results reveal substantial headroom for current systems: models remain weak at processing multiple large image files to estimate sea ice concentration in specific regions and at predicting multidimensional motion vectors. These findings position SEAICE-BENCHas a realistic testbed for measuring progress toward expert-facing LLM systems for sea-ice analysis, forecasting support, and scientific literature synthesis.

1 INTRODUCTION

Sea ice is central to the climate system, polar ecosystems, navigation, operations, and environmental monitoring (Sandven et al., 2023; Meier & Stroeve, 2022). Evaluating large language models (LLMs) in this domain requires more than factual recall: models must connect multispectral satellite imagery, long-horizon numerical records, and specialized scientific literature (Stokholm, 2023; Fetterer et al., 2025; Tschudi et al., 2019). A useful sea-ice benchmark should therefore test remote-sensing interpretation, quantitative trend and motion-field reasoning, and expert scientific knowledge.

Existing benchmarks cover only parts of this evidence chain. ScienceQA is broad and multimodal but targets school curricula; EarthSE is domain-specific but text-centric; and MSEarth uses figures and captions from publications rather than native sensing products (Lu et al., 2022; Xu et al., 2026; Zhao et al., 2025). Sea-ice resources such as IceBench instead evaluate task-specific vision systems (Taleghan et al., 2025). A dedicated benchmark for LLM and multimodal LLM (MLLM) evaluation across sea-ice imagery, numerical records, and literature therefore remains absent.

We present SEAICE-BENCH, a specialized benchmark organized around these three evidence sources. Its visual component contains 6,500 original-resolution images from 500 locations and 130,000 cropped images from 10,000 locations, with 12 spectral bands and a true-color image (TCI) per location. Its numerical component covers monthly sea ice extent from November 1978 to January 2026 and weekly sea ice motion vectors from November 1978 to December 2024 in both hemispheres (Fetterer et al., 2025; Tschudi et al., 2019). Its textual component derives 53,785 question–answer pairs from more than 10,000 sea-ice articles and textbooks. Together, these components probe perception over native imagery, quantitative reasoning over geophysical records, and scientific understanding, supporting realistic assessment of LLMs for sea-ice analysis, forecasting support, and literature synthesis.

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

Contributions. Our contributions are:

- the first sea-ice-specific LLM and MLLM benchmark, to our knowledge, spanning visual, numerical, and textual evidence;
- native full-scene and cropped multispectral imagery together with long-horizon extent and regional motion-vector tasks; and
- 53,785 literature-grounded items that test domain knowledge and scientific reasoning rather than only general science competency.

2 Related Work

Relevant prior work spans general multimodal science QA, Earth- and climate-focused language benchmarks, broader Earth-science LLM evaluation, and sea-ice vision datasets. Each covers only part of the evidence chain required for expert sea-ice analysis.

General multimodal science QA. ScienceQA contains 21,208 multiple-choice questions, including 10,332 with image context, across natural, social, and language science topics from elementary and high-school curricula (Lu et al., 2022). It provides a strong reference for general image-aware scientific reasoning, but does not target remote sensing, geophysical records, or expert sea-ice literature.

Earth and climate language benchmarks. Domain-specific QA resources remain primarily textual. Pirá 2.0 provides 2,258 bilingual QA sets from ocean and climate abstracts and reports (Pirozelli et al., 2024). GeoBench contains 1,578 exam-style geoscience questions (Deng et al., 2024); SciQA contains 2,565 questions over the Open Research Knowledge Graph (Auer et al., 2023); and ClimaQA combines 566 expert-annotated and 3,000 synthetic graduate-textbook questions (Manivannan et al., 2025). These benchmarks test valuable climate and geoscience language capabilities, but do not combine native satellite imagery with numerical geophysical records.

Table 1 summarizes the closest QA benchmarks: existing datasets are generally broad and multimodal but not sea-ice-specific, or domain-specific but text-only.

Data set	#Q	#I	Source	Domain
ScienceQA (Lu et al., 2022)	21,208	10,332	Grade 1–12 textbooks	General science
Pirá 2.0 (Pirozelli et al., 2024)	2,258	0	Abstracts and reports	Ocean and climate
GeoBench (Deng et al., 2024)	1,578	0	Exam questions	Geoscience
SciQA (Auer et al., 2023)	2,565	0	ORKG	Scholarly communication
ClimaQA (Manivannan et al., 2025)	3,566	0	Graduate textbooks	Climate
SeaIce-Bench	53,785	136,500	Literature and satellite imagery	Sea ice

Table 1: Comparison with related QA benchmarks. #Q and #I denote question–answer pairs and images. ClimaQA combines its Gold and Silver splits.

Broader Earth-science LLM and MLLM benchmarks. EarthSE is built from 100,000 research papers and evaluates closed-form QA and open-ended Earth-science exploration (Xu et al., 2026). MSEarth adds more than 289K publication figures with refined captions for multimodal reasoning (Zhao et al., 2025). Both broaden Earth-science evaluation, but their evidence is textual or derived from publication figures rather than native sea-ice sensing products.

Sea-ice-specific data resources. Sea-ice resources emphasize specialized perception. AI4Arctic supports automated chart production from Sentinel-1 SAR, AMSR2 observations, auxiliary weather variables, and expert ice charts (Stokholm, 2023); IceBench standardizes deep-learning-based sea-ice type classification (Taleghan et al., 2025). These resources are valuable for cryospheric vision research, but do not evaluate LLMs across imagery, long-horizon numerical records, and literature-grounded reasoning.

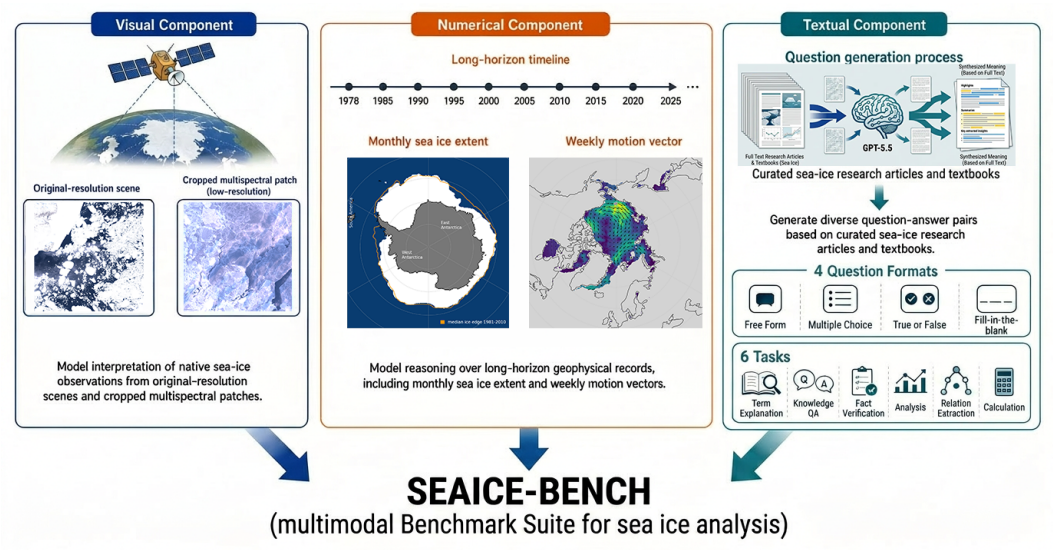


Figure 1: **Framework of SEAICE-BENCH.**

Positioning of SEAICE-BENCH. General science benchmarks provide multimodality without sea-ice expertise; Earth and climate benchmarks provide domain text without native imagery or long-horizon records; and sea-ice vision resources lack LLM-oriented QA. SEAICE-BENCH fills this gap by unifying satellite imagery, historical numerical data, and literature-grounded QA in one LLM and MLLM evaluation suite.

3 SEAICE-BENCH: SEA ICE BENCHMARK

3.1 OVERVIEW

Figure 1 summarizes the benchmark framework, while Figure 2(a) details the tasks defined over its three evidence sources. The visual, numerical, and textual components measure observation, prediction, and scientific understanding, respectively. Each component can be evaluated independently, enabling comparisons among text-only models, multimodal models, and task-specific baselines.

3.2 MONTHLY SEA ICE EXTENT DATA

We use the NOAA/NSIDC Sea Ice Index, Version 4, for monthly Northern- and Southern-Hemisphere extent from November 1978 through January 2026 (Fetterer et al., 2025). Each instance contains a rolling 40-year context (480 observations) and the immediately following month as the target. This formulation preserves seasonal and multi-decadal structure while supporting comparable one-step forecasting in both hemispheres (Chen et al., 2023; Brennan et al., 2023; Meier & Stroeve, 2022).

3.3 WEEKLY REGIONAL SEA ICE MOTION VECTOR DATA

We derive a higher-dimensional task from the Polar Pathfinder 25 km EASE-Grid weekly motion product, whose along-grid u, v components are distributed in georeferenced netCDF files (Tschudi et al., 2019; 2020; Rew & Davis, 1990). Full polar grids over a 40-year weekly history exceed a model-comparable context budget, so we retain one fixed regional window per hemisphere and serialize its cells as CSV (Shafranovich, 2005). Each instance uses roughly 2,080 weekly fields to predict the complete u, v field for the next week. The task therefore tests regional spatial forecasting, not pan-polar skill; fixed windows ensure identical spatial support across models and splits.

3.4 SEA ICE SATELLITE IMAGE DATA

The visual component uses Copernicus Sentinel-2 Collection 1 MSI Level-1C imagery, a high-resolution source previously used for lead detection, surface classification, and concentration or melt analysis (Drusch et al., 2012; Muchow et al., 2021; Petty et al., 2021; Ludwig et al., 2020; Buckley et al., 2023). We retain acquisitions from 2017–2026 in Arctic and Antarctic sea-ice regions, including some cloudy scenes to reflect realistic observation conditions. Because the evaluated MLLMs do not accept native JPEG2000 files, we convert imagery to PNG. The full-scene split contains 6,500 images from 500 locations; the crop split contains 130,000 images from 10,000 additional locations. Every location provides 12 spectral bands and one TCI, and the target is sea-ice concentration. We construct the reference concentration from a pixelwise snow/ice mask based on the Normalized Difference Snow Index (NDSI) and near-infrared reflectance, as detailed in Appendix B.

3.5 DOMAIN LITERATURE

We assemble more than 10,000 articles and textbooks through Semantic Scholar queries, citation expansion, public PDF links, and legitimate institutional access; we neither scrape publisher sites nor bypass paywalls (Wade, 2022; Kinney et al., 2023; Weeks, 2010; Untersteiner, 1986). GPT-5.5 generates candidates from source passages across the six task families in Table 2 and four formats: free form, multiple choice, true/false, and fill-in-the-blank (Singh et al., 2025). Rule-based checks remove malformed, duplicated, or leaking items, and semantic review removes unsupported or ambiguous questions. Reasoning traces are used only for curation and are withheld at evaluation. Sea-ice experts manually reviewed 5,770 items (10.7%); Appendix M reports the composition.

Group	Task	Capability
Reasoning	Analysis (Cai et al., 2025)	Draw conclusions from evidence reported in sea-ice literature.
	Relation extraction (Li et al., 2026)	Identify relations such as sensor–product, process–variable, and cause–effect links.
	Calculation (Stephan et al., 2025)	Reason over values, trends, errors, resolutions, and derived quantities.
Understanding	Term explanation (Heiman, 2025)	Define technical concepts and explain their scientific roles.
	Knowledge QA (Xu et al., 2024)	Answer direct domain questions grounded in a source.
	Fact verification (Zhang & Gao, 2023)	Determine whether a claim is supported and identify the relevant evidence.

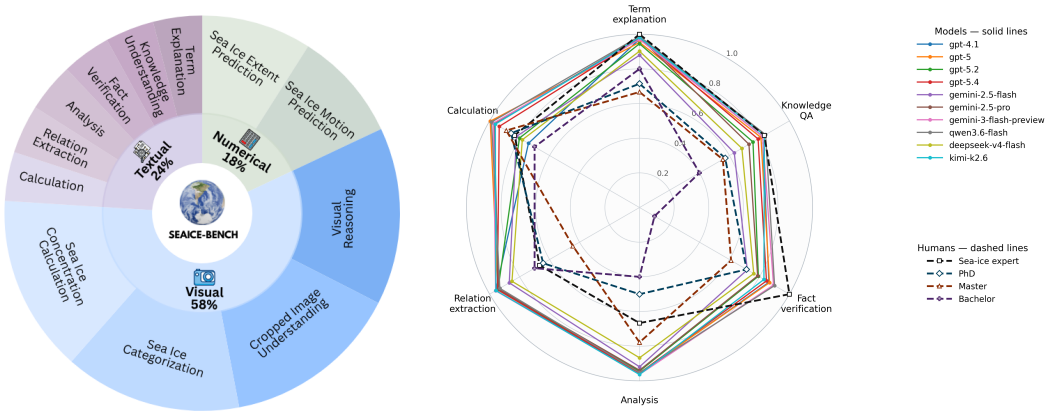
Table 2: Literature-grounded task families in SEAICE-BENCH.

4 EXPERIMENTS

We evaluate whether current frontier proprietary LLMs and MLLMs can use the three evidence sources in SEAICE-BENCH: numerical records, satellite imagery, and domain literature. The visual, numerical, and textual components are evaluated as separate but complementary tasks under a common protocol rather than chained into an end-to-end workflow. All LLM and MLLM experiments follow a unified zero-shot setup with provider-hosted APIs only; we do not fine-tune any foundation model on benchmark data. For every benchmark task and answer format, we use a fixed instruction template shared across models. Separately, for the visual task, we train a task-specific non-LLM baseline adapted from MMSeaIce on the visual training split; its architectural changes and input preprocessing are detailed in Appendix G.

4.1 MODEL SUITE AND UNIFIED PROTOCOL

The evaluation suite contains ten frontier models—seven proprietary and three open-weight—with native support for the input modalities required by SEAICE-BENCH. We exclude GPT-5.5 because it was used to generate the question–answer corpus, thereby avoiding generator–evaluator overlap.



(a) Benchmark task composition. (b) Task-level model and human accuracy.

Figure 2: **Benchmark composition and textual-task accuracy.** (a) Visual, numerical, and textual components of SEAICE-BENCH and their associated tasks. (b) Closed-form accuracy of the evaluated models and multiple-choice survey accuracy of the human participant groups across the six textual task families.

The full provider/model list, serving stacks, and version notes are deferred to Appendix O. Across all experiments, we disable tool use, search, and code execution so that results reflect the models' native responses to the benchmark inputs rather than external augmentation. Numerical inputs are serialized as CSV files, and visual inputs are passed as PNG images plus a shared textual instruction header. For the literature-based QA evaluation, questions span six task categories: Term Explanation, Knowledge QA, Fact Verification, Analysis, Relation Extraction, and Calculation. Each question is presented using one of four standardized prompt templates according to its answer format, with the placeholder `{question}` replaced by the corresponding benchmark question. The complete templates are provided in Appendix F.3.

5 Results and Discussion

Across SEAICE-BENCH, performance depends strongly on both input modality and output dimensionality. The evaluated models are most effective on literature-grounded question answering and on scalar, one-step sea-ice extent forecasting. Performance deteriorates on tasks that require joint interpretation of multiple multispectral images or generation of a spatially coupled vector field. Because WCP-RMSE penalizes invalid outputs, the numerical and visual results reflect both predictive error and the ability to return a complete, machine-readable answer. The following subsections discuss these modality-specific patterns and their implications.

5.1 Visual Understanding and Prediction

We report visual results on the full-scene and crop-based Sentinel-2 splits. The target is sea ice concentration, expressed as a percentage in $[0, 100\%]$, estimated from the TCI image together with the 12 spectral-band images. The exact prompt used for this task is provided in Appendix F.1. All ten MLLMs are evaluated on the crop split, while the full-scene split is used as an additional stress test for the three MLLMs that stably accept the complete 13-image prompt. We also evaluate Adapted MMSeaIce, a task-specific non-LLM baseline derived from the U-Net-based MMSeaIce method (Chen et al., 2024), on both splits. The published formulation processes a single image and predicts SIC in 10-percentage-point intervals, whereas our task provides 13 co-registered images and requires a continuous numerical prediction. We therefore modify the input layer, replace the original output with a scalar-regression head, and train the adapted model from scratch on our training data; Appendix G provides the full architectural details. Table 3 reports worst-case-penalized RMSE (WCP-RMSE), which assigns the maximum possible error to invalid outputs, together with completion rate (CR). For this single-value task, a completion is valid only when the entire response consists of one finite

Model	Full-scene images		Cropped images	
	WCP-RMSE (↓)	CR (%) (↑)	WCP-RMSE (↓)	CR (%) (↑)
GPT-4.1	—	—	46.23 ± 2.14	100.00
GPT-5	—	—	36.80 ± 1.87	100.00
GPT-5.2	—	—	34.14 ± 2.23	100.00
GPT-5.4	—	—	34.35 ± 1.67	100.00
Gemini 2.5 Flash	58.29 ± 3.43	76.47	55.06 ± 1.50	87.07
Gemini 2.5 Pro	43.18 ± 2.46	100.00	38.26 ± 1.59	100.00
Gemini 3 Flash Preview	49.82 ± 2.28	97.96	40.89 ± 1.52	97.84
Qwen 3.6 Flash	—	—	79.91 ± 0.91	79.08
DeepSeek V4 Flash	—	—	75.46 ± 1.03	82.67
Kimi K2.6	—	—	99.81 ± 0.13	0.54
MMSeaIce (Chen et al., 2024)(non-LLM)	$\mathbf{2.82} \pm 0.70$	100.00	$\mathbf{2.06} \pm 1.13$	100.00

Table 3: Sea ice concentration estimation results on the full-scene and cropped Sentinel-2 splits using the TCI together with all 12 spectral-band images. Adapted MMSeaIce is our task-specific modification of the published MMSeaIce architecture (Chen et al., 2024). Worst-case-penalized RMSE (WCP-RMSE) is reported as mean $\pm$ standard deviation in percentage points and assigns invalid outputs the maximum error over the $[0, 100]$ target range. Bold marks the lowest WCP-RMSE in each split. Lower WCP-RMSE is better; higher completion rate (CR) is better. A dash indicates that the model was not evaluated on the full-scene split.

numerical value in $[0, 100]$; refusals, missing-image claims, and responses containing explanations or alternative estimates are invalid. This joint reporting prevents low completion from making a model appear artificially accurate.

The visual task exhibits the largest gap between general-purpose MLLMs and a specialized baseline. On full-scene images, Gemini 2.5 Pro is the strongest MLLM, with a WCP-RMSE of 43.18 and a completion rate of 100.00%, whereas Adapted MMSeaIce achieves 2.82 with full completion. On cropped images, GPT-5.2 obtains the best MLLM result at 34.14 with a 100.00% completion rate, but remains substantially behind Adapted MMSeaIce at 2.06. Thus, even when the strongest MLLMs return valid answers for every instance, successful execution does not imply accurate spatial prediction. Models with low completion are penalized even more strongly: most notably, Kimi K2.6 completes only 0.54% of cropped-image instances and therefore has a WCP-RMSE of 99.81.

Qualitative inspection shows that failures arise at both the input-processing and prediction stages. Models may deny that the attached images are present, refuse the task, access nominally identical multi-image inputs inconsistently, rely primarily on one rendered image, or violate the required single-value output format. Even format-compliant responses can substantially misestimate concentration, indicating difficulty in combining spectral information with the geometry and spatial distribution of ice. Appendix I distinguishes these completion-level failures from prediction errors in valid numerical outputs. Cropping reduces the visual and contextual burden for the strongest MLLMs, but the persistent order-of-magnitude gap to Adapted MMSeaIce shows that current general-purpose models are not yet reliable substitutes for task-specific remote-sensing models.

5.2 Numerical prediction

We report numerical results on two tracks derived from the monthly extent and weekly motion records in Section 3. The first track is one-step-ahead monthly extent forecasting: each instance provides 480 months of Arctic or Antarctic extent and asks the model to predict the immediately following month. The exact evaluation prompt and its single-value output constraints are provided in Appendix F.2. The second track evaluates weekly motion prediction over the fixed hemispheric windows defined above. Rather than reducing a window to an average or a small set of probe points, we ask the model to generate the complete next-week u, v field for every retained cell while preserving direction, magnitude, and spatial coherence. This task remains substantially more structured than scalar extent forecasting, but it is deliberately regional rather than full-grid so that every evaluated model receives the same 40-year history under a feasible input budget. Table 4 combines next-month extent and

Model	Sea-ice extent				Motion vector			
	Northern region		Southern region		North		South	
	WCP-RMSE (↓)	CR (%) (↑)	WCP-RMSE (↓)	CR (%) (↑)	WCP-RMSE (↓)	CR (%) (↑)	WCP-RMSE (↓)	CR (%) (↑)
GPT-4.1	2.031 ± 0.335	100.0	3.457 ± 0.515	100.0	7.91 ± 0.60	100.0 ± 0.0	53.80 ± 13.09	80.0 ± 6.8
GPT-5	0.258 ± 0.040	100.0	0.383 ± 0.075	100.0	6.87 ± 0.58	100.0 ± 0.0	38.39 ± 13.62	93.3 ± 3.8
GPT-5.2	1.321 ± 0.363	100.0	1.037 ± 0.275	100.0	7.50 ± 0.50	100.0 ± 0.0	98.04 ± 9.54	40.2 ± 8.1
GPT-5.4	0.583 ± 0.121	100.0	0.635 ± 0.117	100.0	6.77 ± 0.50	100.0 ± 0.0	43.90 ± 14.18	86.7 ± 6.3
Gemini 2.5 Flash	4.552 ± 1.905	92.5	3.537 ± 2.109	97.5	61.22 ± 4.06	40.0 ± 5.3	70.71 ± 4.90	34.8 ± 5.0
Gemini 2.5 Pro	0.464 ± 0.275	100.0	0.691 ± 0.097	100.0	19.82 ± 11.48	96.8 ± 2.5	35.59 ± 10.64	93.2 ± 3.1
Gemini 3 Flash Preview	$\mathbf{0.226} \pm 0.058$	100.0	$\mathbf{0.316} \pm 0.054$	100.0	164.1 ± 0.23	0.2 ± 0.1	123.68 ± 72.65	72.7 ± 15.6
Qwen 3.6 Flash	0.602 ± 0.165	100.0	1.825 ± 1.801	98.5	90.66 ± 5.10	14.5 ± 5.1	79.77 ± 5.76	29.8 ± 6.3
DeepSeek V4 Flash	1.063 ± 0.447	100.0	1.421 ± 0.486	100.0	89.81 ± 6.46	18.3 ± 5.5	101.66 ± 16.81	32.9 ± 6.9
Kimi K2.6	10.600 ± 1.333	60.5	11.999 ± 1.732	63.3	63.22 ± 27.47	72.5 ± 5.7	101.08 ± 34.63	74.9 ± 6.8
Non-LLM methods								
EA-LSTM (Ali et al., 2021)	0.365 ± 0.127	100.0	0.429 ± 0.134	100.0	—	—	—	—
Attention-LSTM (Wei et al., 2022; Hoffman et al., 2023)	—	—	—	—	$\mathbf{6.52 \pm 1.01}$	100.0 ± 0.0	$\mathbf{6.88 \pm 1.30}$	100.0 ± 0.0
Persistence (PS) (Bilgili, 2026)	1.931 ± 0.462	100.0	2.977 ± 0.787	100.0	7.17 ± 1.03	100.0 ± 0.0	7.95 ± 1.02	100.0 ± 0.0

Table 4: Combined numerical forecasting results for next-month sea-ice extent and weekly sea-ice motion vector prediction in the Northern and Southern regions. Both tasks report worst-case-penalized RMSE (WCP-RMSE) as mean $\pm$ standard deviation, together with completion rate. Extent WCP-RMSE is measured in millions of square kilometers; motion-vector completion is reported as a percentage. Bold marks the lowest error in each column. Lower error is better; higher completion is better. An em dash indicates that a method was not evaluated on that task.

regional motion-vector forecasting results, reporting window-specific WCP-RMSE and completion rates.

LLMs are considerably more competitive on next-month sea-ice extent forecasting, where both the input and output admit a comparatively simple temporal representation. Gemini 3 Flash Preview achieves the lowest WCP-RMSE in both regions, with 0.226 ± 0.058 in the Northern region and 0.316 ± 0.054 in the Southern region, improving on EA-LSTM at 0.365 ± 0.127 and 0.429 ± 0.134, respectively. GPT-5 also outperforms EA-LSTM in both regions, with errors of 0.258 ± 0.040 and 0.383 ± 0.075, while maintaining a 100% completion rate. These results show that capable LLMs can match or exceed a specialized sequence baseline on one-step scalar forecasting, although this finding is specific to the benchmark setup and does not establish superiority for longer forecast horizons.

This advantage does not transfer consistently to regional motion-vector prediction. In the Northern window, GPT-5.4 obtains a WCP-RMSE of 6.77 ± 0.50, approaching Attention-LSTM at 6.52 ± 1.01 and slightly improving on persistence at 7.17 ± 1.03. In the Southern window, however, the strongest LLM result is 35.59 ± 10.64 from Gemini 2.5 Pro, compared with 6.88 ± 1.30 for Attention-LSTM and 7.95 ± 1.02 for persistence. Completion rates also decline sharply for several models. The contrast between extent and motion forecasting therefore suggests that performance degrades as the target changes from a single temporally varying scalar to a spatially coupled two-component field. The regional asymmetry further cautions against inferring general motion-forecasting competence from the favorable Northern-window result.

Qualitative inspection identifies distinct failure modes across the two tracks. For extent forecasting, invalid completions arise from non-finite `NaN` values or verbose responses that violate the required single-value format; Appendix J provides a detailed analysis and a representative failed completion. Some responses nevertheless articulate a sensible computational procedure, such as estimating a historical month-to-month change, which helps explain why models can perform well on the scalar task. Tool use and code execution were disabled, however, so these results reflect in-context reasoning rather than externally executed calculations. For motion vectors, models often emit code scaffolds, code-wrapped data, or incorrect delimiters instead of a complete CSV field. These failures, analyzed in Appendix K, show that procedural reasoning alone does not overcome inadequate representations of complex spatial dynamics or the difficulty of generating high-dimensional structured outputs.

5.3 Question answering

Textual results are reported on the literature-grounded QA corpus across all six task families and all four answer formats. Table 5 reports the detailed quantitative results for the model variants evaluated on this component. Following the format-aware evaluation practice used in recent scientific QA benchmarks (Manivannan et al., 2025) and the metrics formalized in Section H, we use accuracy (Acc.) for exact-match closed-form scoring and semantic similarity (SS) for cosine-similarity scoring of free-form responses. Because each task family appears across multiple answer formats, we report

Model	Term Expl.		Knowledge QA		Fact Verif.	
	SS	Acc.	SS	Acc.	SS	Acc.
gpt-4.1	0.750 ± 0.26	0.972 ± 0.17	0.713 ± 0.20	0.822 ± 0.38	0.635 ± 0.28	0.863 ± 0.34
gpt-5	0.770 ± 0.23	0.953 ± 0.21	0.732 ± 0.19	0.810 ± 0.39	0.605 ± 0.21	0.867 ± 0.34
gpt-5.2	0.765 ± 0.22	0.945 ± 0.23	0.721 ± 0.19	0.757 ± 0.43	0.618 ± 0.22	0.793 ± 0.41
gpt-5.4	0.771 ± 0.23	0.974 ± 0.16	0.732 ± 0.20	0.791 ± 0.41	0.618 ± 0.24	0.847 ± 0.36
gemini-2.5-flash	0.729 ± 0.26	0.879 ± 0.33	0.701 ± 0.21	0.632 ± 0.48	0.582 ± 0.23	0.719 ± 0.45
gemini-2.5-pro	$\mathbf{0.793 \pm 0.22}$	0.981 ± 0.14	0.705 ± 0.21	0.730 ± 0.45	$\mathbf{0.650 \pm 0.24}$	0.788 ± 0.41
gemini-3-flash-preview	0.758 ± 0.24	0.970 ± 0.17	$\mathbf{0.734 \pm 0.19}$	0.821 ± 0.38	0.609 ± 0.22	0.894 ± 0.31
qwen3.6-flash	0.706 ± 0.27	$\mathbf{0.988 \pm 0.11}$	0.660 ± 0.22	$\mathbf{0.841 \pm 0.37}$	0.594 ± 0.16	$\mathbf{0.901 \pm 0.30}$
deepseek-v4-flash	0.782 ± 0.25	0.902 ± 0.30	0.697 ± 0.21	0.683 ± 0.47	0.635 ± 0.22	0.762 ± 0.43
kimi-k2.6	0.758 ± 0.28	0.981 ± 0.14	0.623 ± 0.28	0.833 ± 0.38	0.619 ± 0.24	0.830 ± 0.38

Model	Analysis		Relation Extr.		Calculation	
	SS	Acc.	SS	Acc.	SS	Acc.
gpt-4.1	0.724 ± 0.24	0.957 ± 0.20	0.688 ± 0.26	0.959 ± 0.20	0.785 ± 0.17	0.741 ± 0.44
gpt-5	0.729 ± 0.23	0.955 ± 0.21	0.699 ± 0.24	0.944 ± 0.23	$\mathbf{0.837 \pm 0.18}$	$\mathbf{0.996 \pm 0.06}$
gpt-5.2	0.727 ± 0.22	0.944 ± 0.23	0.689 ± 0.23	0.944 ± 0.23	0.805 ± 0.18	0.801 ± 0.40
gpt-5.4	$\mathbf{0.741 \pm 0.23}$	0.955 ± 0.21	0.697 ± 0.24	0.948 ± 0.22	0.814 ± 0.18	0.936 ± 0.25
gemini-2.5-flash	0.700 ± 0.26	0.919 ± 0.27	0.656 ± 0.27	0.869 ± 0.34	0.783 ± 0.17	0.821 ± 0.38
gemini-2.5-pro	0.722 ± 0.22	0.939 ± 0.24	0.714 ± 0.23	0.941 ± 0.24	0.807 ± 0.15	0.970 ± 0.17
gemini-3-flash-preview	0.714 ± 0.24	0.963 ± 0.19	0.694 ± 0.24	$\mathbf{0.960 \pm 0.20}$	0.780 ± 0.17	0.974 ± 0.16
qwen3.6-flash	0.701 ± 0.20	0.952 ± 0.22	0.663 ± 0.25	0.932 ± 0.25	0.733 ± 0.19	0.987 ± 0.11
deepseek-v4-flash	0.713 ± 0.24	0.867 ± 0.34	0.685 ± 0.27	0.851 ± 0.36	0.788 ± 0.19	0.782 ± 0.42
kimi-k2.6	0.711 ± 0.26	$\mathbf{0.964 \pm 0.19}$	$\mathbf{0.725 \pm 0.24}$	0.958 ± 0.20	0.778 ± 0.21	0.962 ± 0.19

Table 5: Task-level textual evaluation results. For each task family, Acc. reports exact-match accuracy on closed-form items and SS reports cosine-based semantic similarity on free-form items. Each entry is shown as mean $\pm$ standard deviation, with the standard deviation rounded to two decimals. Higher is better for both Acc. and SS. The best result in each column is boldfaced.

both metrics for every task family, with each entry shown as mean $\pm$ standard deviation. We do not report a separate completion rate for textual QA: missing, malformed, or incorrect answers remain in the evaluation set and lower Acc. or SS as appropriate. Verified examples sampled from the question dataset and administered in the human survey are reproduced, together with their reference answers and verification notes, in Appendix N.5.

Question-answering performance is substantially stronger and more consistent than visual prediction performance; Figure 2(b) summarizes the task-level accuracy patterns for models and human participant groups. Qwen 3.6 Flash obtains the highest exact-match accuracies for Term Explanation (0.988), Knowledge QA (0.841), and Fact Verification (0.901), while GPT-5 achieves the highest Calculation accuracy (0.996). The leading Analysis and Relation Extraction accuracies are also high, reaching 0.964 for Kimi K2.6 and 0.960 for Gemini 3 Flash Preview, respectively. No model dominates every task family or metric: for example, the best semantic-similarity results are distributed among Gemini 2.5 Pro, Gemini 3 Flash Preview, GPT-5.4, Kimi K2.6, and GPT-5. The generally lower and more variable semantic-similarity scores indicate that strong closed-form accuracy does not always translate into equally close free-form agreement with the references. Nevertheless, the overall pattern shows that current LLMs are effective at retrieving, organizing, and applying sea-ice knowledge in text-based tasks.

Consistent with the unified protocol above, the model receives only the question at inference time; the source passage and intermediate reasoning traces used during data construction are withheld during evaluation. Results are reported by task family, with Acc. and SS capturing both answer quality and failures to provide a correct, scoreable response.

To contextualize these model results with human performance, we also conducted an approximately 20-minute participant survey on the same six textual task families. Each participant answered two multiple-choice questions per task family (12 questions total). The survey design, participant safeguards, sample-size rationale, and quality-control criteria are provided in Appendix N, the

administered question examples in Appendix N.5, and the task-level human results in Table 8. After filtering, the human baseline comprised 24 participants. Sea-ice experts achieved the highest pooled accuracy (83.3%), followed by other PhD holders (66.7%), master's degree holders (65.7%), and bachelor's degree holders (51.7%). Many task-level LLM accuracies in Table 5 are numerically above the pooled non-expert survey accuracies and close to or above the pooled sea-ice-expert accuracy. This comparison is only suggestive, however, because the survey used a small multiple-choice subset whereas the model table aggregates a much larger collection of closed-form items across answer formats. It should not be interpreted as a matched human–model ranking or as evidence that the models have attained expert-level sea-ice competence.

The survey is an auxiliary, exploratory reference for benchmark difficulty rather than a formally powered population study or the principal contribution of the paper. Small, task-oriented human reference groups are common in AI benchmark evaluation: WebArena estimated human performance using five computer-science graduate students on 170 sampled tasks; MMMU used three senior students per subject, with each participant answering a subject-specific subset; and MMLU reported crowd-worker performance while estimating its frequently cited expert level from examination statistics and informed estimates rather than a matched expert cohort (Zhou et al., 2024; Yue et al., 2024; Hendrycks et al., 2021). Accordingly, our original design did not include an a priori power analysis for estimating a causal effect or precise population parameter. We interpret the 24-participant results only as a transparent descriptive baseline; the detailed rationale and limitations are given in Appendix N.

6 Conclusion

We introduced SeaIce-Bench, a multi-task, multimodal benchmark that combines native Sentinel-2 imagery, long-horizon numerical records, and literature-grounded question answering in one evaluation suite. Its tasks correspond to complementary components of sea-ice science: observation through sea ice concentration estimation, prediction through next-step extent and motion-vector forecasting, and scientific understanding through article-grounded QA. The benchmark brings together 136,500 satellite images, monthly sea ice extent records from November 1978 to January 2026, weekly sea ice motion vectors from November 1978 to December 2024, and 53,785 question–answer pairs derived from more than 10,000 sea-ice research articles and textbooks. The textual component further broadens scientific QA beyond factual recall by evaluating explanation, verification, analysis, relation extraction, and calculation. To the best of our knowledge, SeaIce-Bench is the first benchmark specifically designed to evaluate foundation and multimodal models across this range of sea-ice and closely related ocean-science capabilities.

Our experiments show that the capabilities of current frontier models are uneven rather than uniformly weak. They are already promising for literature-grounded question answering and can be competitive for one-step scalar extent forecasting, but their performance falls sharply on multispectral concentration estimation and spatially coupled motion-vector prediction. Completion-rate gaps further show that producing a scoreable structured output remains difficult for several models. The motion results apply only to the two fixed evaluation windows and do not establish pan-polar forecasting performance. On the visual task, Adapted MMSeaIce substantially outperforms every evaluated MLLM on both full-scene and cropped imagery, highlighting the continuing value of specialized models. Current LLMs may therefore be useful as scientific question-answering and numerical-reasoning assistants, but they should not replace domain-specific methods where accurate remote-sensing interpretation or high-dimensional spatial forecasting is required. We do not claim that this first release resolves every aspect of sea-ice modelling or constitutes a complete end-to-end workflow. Rather, it establishes an initial, extensible evaluation foundation for a technically important and currently underrepresented domain whose relevance spans the climate system, polar ecosystems, ocean–atmosphere interactions, and high-latitude operations.

Ethical and Licensing Considerations

SeaIce-Bench is assembled from Earth-observation imagery, public geophysical records, and copyrighted scientific literature, so we make its release boundaries explicit. For the visual component, we use Sentinel-2 imagery whose finest native spatial resolution is 10 m (Drusch et al., 2012). In addition, our extraction points are sea-ice regions in the Arctic and Antarctic rather than populated street-level settings (Appendix A). Consequently, the satellite images in SeaIce-Bench are not suitable for identifying individual persons or inferring person-level attributes. We do not intend the benchmark for surveillance or any person-focused downstream use. For licensing, the satellite-image component follows the Copernicus Sentinel data terms, which permit free, full, and open use subject to the required source notice for redistributed or modified data. The literature component is handled more conservatively. We extract textbooks and research articles only to construct derived question–answer pairs for evaluation, and we do not distribute the original books, articles, scans, or any reconstructed text corpus. Copyright and licence obligations for those source documents remain with their respective rights holders.

Generative AI Disclosure

In this work, we used GPT-5.5 to generate candidate synthetic question–answer data sets from source passages for the textual component of SeaIce-Bench. All generated items were subjected to rule-based checks for malformed, duplicated, and answer-leaking content and to semantic review for unsupported or ambiguous content; in addition, sea-ice experts manually reviewed 5,770 randomly sampled pairs (10.7% of the corpus), as detailed in Appendix M. The other LLMs and MLLMs reported in this paper were used as experimental subjects: they produced outputs for the benchmark tasks under the evaluation protocol described in Section 4. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Formulate mathematical claims, provide critical ingredients for proving mathematical claims, are not applicable to this work. Additionally, we used generative AI tools for manuscript proofreading and typographical-error checking. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility Statement

The SeaIce-Bench dataset is publicly available at `https://huggingface.co/datasets/paperupload/SEAICEBENCH`, and the accompanying code is publicly available at `https://github.com/paperupload409-beep/sea_ice_bench`.

References

Sahara Ali, Yiyi Huang, Xin Huang, and Jianwu Wang. Sea ice forecasting using attention-based ensemble LSTM. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. URL `https://ccai-papers.s3.us-east-1.amazonaws.com/icml2021/50/paper.pdf`.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1): 7240, 2023.

Mehmet Bilgili. Multivariate deep learning approach to forecast of an ice-free arctic summer under the SSP2-4.5 and SSP5-8.5 scenarios. *Polar Science*, 47:101308, 2026. doi: 10.1016/j.polar.2025.101308.

M. Kathleen Brennan, Gregory J. Hakim, and Edward Blanchard-Wrigglesworth. Monthly arctic sea-ice prediction with a linear inverse model. *Geophysical Research Letters*, 50(7):e2022GL101656, 2023. doi: 10.1029/2022GL101656.

Ellen M. Buckley, Sinéad L. Farrell, Ute C. Herzfeld, Melinda A. Webster, Thomas Trantow, Oliwia N. Baney, Kyle A. Duncan, Huilin Han, and Matthew Lawson. Observing the evolution of summer melt on multiyear sea ice with ICESat-2 and sentinel-2. *The Cryosphere*, 17:3695–3719, 2023. doi: 10.5194/tc-17-3695-2023.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Wang Changxin, Zhifeng Gao, Hongshuai Wang, Li Yongge, Mujie Lin, et al. Sciassess: Benchmarking llm proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2335–2357, 2025.

Siwen Chen, Kehan Li, Hongpeng Fu, Ying Cheng Wu, and Yiyi Huang. Sea ice extent prediction with machine learning methods and subregional analysis in the arctic. *Atmosphere*, 14(6):1023, 2023. doi: 10.3390/atmos14061023.

Xinwei Chen, Muhammed Patel, Fernando J. Pena Cantu, Jinman Park, Javier Noa Turnes, Linlin Xu, K. Andrea Scott, and David A. Clausi. MMSeaIce: A collection of techniques for improving sea ice mapping with a multi-task model. *The Cryosphere*, 18(4):1621–1632, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Yi Xu, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 161–170, 2024.

R Devika, Subramaniyaswamy Vairavasundaram, C Sakthi Jay Mahenthar, Vijayakumar Varadarajan, and Ketan Kotecha. A deep learning model based on bert and sentence transformer for semantic keyphrase extraction on big social data. *IEEE Access*, 9:165252–165261, 2021.

Matthias Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernández, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, Francesco Marchese, and P. Bargellini. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012. doi: 10.1016/j.rse.2011.11.026.

Florence Fetterer, Ken Knowles, W. N. Meier, Matt Savoie, Ann Windnagel, and Trey Stafford. Sea ice index, version 4, 2025. URL `https://nsidc.org/data/G02135/versions/4`.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL `https://jmlr.org/papers/v24/21-0048.html`.

Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2):349–360, 2009. doi: 10.1093/poq/nfp031.

John W Hager, Larry L Fry, Sandra S Jacks, and David R Hill. Datums, ellipsoids, grids, and grid reference systems. Technical report, DEFENSE MAPPING AGENCY HYDROGRAPHIC/ TOPOGRAPHIC CENTER WASHINGTON DC, 1992.

Thomas M. Haladyna, Steven M. Downing, and Michael C. Rodriguez. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3): 309–333, 2002. doi: 10.1207/S15324818AME1503_5.

Alice Heiman. The accuracy, robustness, and readability of llm-generated sustainability-related word definitions. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pp. 104–109, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Lauren Hoffman, Matthew R. Mazloff, Sarah T. Gille, Donata Giglio, Cecilia M. Bitz, Patrick Heimbach, and Kayli Matsuyoshi. Machine learning for daily forecasts of arctic sea ice motion: An attribution assessment of model predictive skill. *Artificial Intelligence for the Earth Systems*, 2 (4):e230004, 2023. doi: 10.1175/AIES-D-23-0004.1.

Dahyeon Jeong, Shilpa Aggarwal, Jonathan Robinson, Naresh Kumar, Alan Spearot, and David S. Park. Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics*, 161:102992, 2023. doi: 10.1016/j.jdeveco.2022.102992.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023. doi: 10.48550/arXiv.2301.10140.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. Llm with relation classifier for document-level relation extraction. *IEEE Transactions on Big Data*, 2026.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Valentin Ludwig, Gunnar Spreen, and Leif Toudal Pedersen. Evaluation of a new merged sea-ice concentration dataset at 1 km resolution from thermal infrared and passive microwave satellite data in the arctic. *Remote Sensing*, 12(19):3183, 2020. doi: 10.3390/rs12193183.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. ClimaQA: An automated evaluation framework for climate question answering models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=goFpCuJalN.

Adam W. Meade and S. Bartholomew Craig. Identifying careless responses in survey data. *Psychological Methods*, 17(3):437–455, 2012. doi: 10.1037/a0028085.

Walter N. Meier and Julienne Stroeve. An updated assessment of the changing arctic sea ice cover. *Oceanography*, 35(3–4):10–19, 2022. doi: 10.5670/oceanog.2022.114.

Marek Muchow, Amelie U. Schmitt, and Lars Kaleschke. A lead-width distribution for antarctic sea ice: a case study for the weddell sea with high-resolution sentinel-2 images. *The Cryosphere*, 15: 4527–4537, 2021. doi: 10.5194/tc-15-4527-2021.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report: Ethical principles and guidelines for the protection of human subjects of research. Technical report, U.S. Department of Health and Human Services, 1979. URL https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html.

Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45 (4):867–872, 2009. doi: 10.1016/j.jesp.2009.03.009.

Alek A. Petty, Marco Bagnardi, Nathan T. Kurtz, Rachel Tilling, Steven Fons, Thomas Armitage, Christopher Horvat, and Ronald Kwok. Assessment of ICESat-2 sea ice surface classification with sentinel-2 imagery: Implications for freeboard and new estimates of lead and floe geometry. *Earth and Space Science*, 8(3):e2020EA001491, 2021. doi: 10.1029/2020EA001491.

Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR Costa, and Fabio G Cozman. Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63, 2024.

Melanie Revilla and Carlos Ochoa. Ideal and maximum length for a web survey. *International Journal of Market Research*, 59(5):557–565, 2017. doi: 10.2501/IJMR-2017-039.

Russ Rew and Glenn Davis. Netcdf: An interface for scientific data access. *IEEE Computer Graphics and Applications*, 10(4):76–82, 1990. doi: 10.1109/38.56302.

Stein Sandven, Gunnar Spreen, Georg Heygster, Fanny Girard-Ardhuin, Sin'ead L. Farrell, Wolfgang Dierking, and Richard A. Allard. Sea ice remote sensing—recent developments in methods and climate data sets. *Surveys in Geophysics*, 44:1653–1689, 2023. doi: 10.1007/s10712-023-09781-0.

Yakov Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180, October 2005. URL https://www.rfc-editor.org/info/rfc4180.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 759–773, 2025.

Anders R. Stokholm. Ai4arctic sea ice challenge dataset, 2023. Data set, Technical University of Denmark Data Repository.

S. A. Taleghan, A. P. Barrett, W. N. Meier, and F. Banaei-Kashani. Icebench: A benchmark for deep-learning-based sea-ice type classification. *Remote Sensing*, 17(9):1646, 2025.

Mark Tschudi, Walter N. Meier, J. Scott Stewart, Charles Fowler, and Jim Maslanik. Polar pathfinder daily 25 km ease-grid sea ice motion vectors, version 4, 2019. URL https://nsidc.org/data/NSIDC-0116/versions/4.

Mark A. Tschudi, Walter N. Meier, and J. Scott Stewart. An enhancement to sea ice motion and age products at the national snow and ice data center (nsidc). *The Cryosphere*, 14(5):1519–1536, 2020. doi: 10.5194/tc-14-1519-2020.

Norbert Untersteiner (ed.). *The Geophysics of Sea Ice*. Springer, New York, NY, 1986. doi: 10.1007/978-1-4899-5352-0.

Alex D. Wade. The semantic scholar academic graph (s2ag). In *Companion Proceedings of the Web Conference 2022*, WWW '22, pp. 739, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3527147. URL https://doi.org/10.1145/3487553.3527147.

Yuhao Wang, Yusheng Liao, Heyang Liu, Hongcheng Liu, Yanfeng Wang, and Yu Wang. MM-SAP: A comprehensive benchmark for assessing self-awareness of multimodal large language models in perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9192–9205, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.498. URL https://aclanthology.org/2024.acl-long.498/.

Willy Weeks. *On Sea Ice*. University of Alaska Press, Fairbanks, AK, 2010. doi: 10.5876/9781602231016.

Jianfen Wei, Renlong Hang, and Jing-Jia Luo. Prediction of pan-arctic sea ice using attention-based LSTM neural networks. *Frontiers in Marine Science*, 9:860403, 2022. doi: 10.3389/fmars.2022.860403.

Wei Xu, Xinyue Zhao, Yuxin Zhou, Xiaofei Yue, Ben Fei, Fei Ling, Weichen Zhang, and Lei Bai. Earthse: A benchmark evaluating earth scientific exploration capability for large language models. In *International Conference on Learning Representations*, 2026.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. Generate-on-graph: Treat llm as both agent and kg for incomplete knowledge graph question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 18410–18430, 2024.

Ting Yan and Roger Tourangeau. Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1):51–68, 2008. doi: 10.1002/acp.1331.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.624. URL https://aclanthology.org/2024.findings-acl.624/.

Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the asia-pacific chapter of the association for computational linguistics (volume 1: Long papers)*, pp. 996–1011, 2023.

Xiangyu Zhao, Wanghan Xu, Bo Liu, Yuhao Zhou, Fenghua Ling, Ben Fei, Xiaoyu Yue, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. Msearth: A multimodal scientific dataset and benchmark for phenomena uncovering in earth science. *arXiv preprint arXiv:2505.20740*, 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

A Satellite Image Extraction Regions

To construct the satellite-image component of SeaIce-Bench, we extract samples from the marginal ice zone in both hemispheres under Military Grid Reference System (MGRS) (Hager et al., 1992). As shown in Figure 3, the extraction region is restricted to locations above 64° North in the Arctic and below 64° South in the Antarctic.

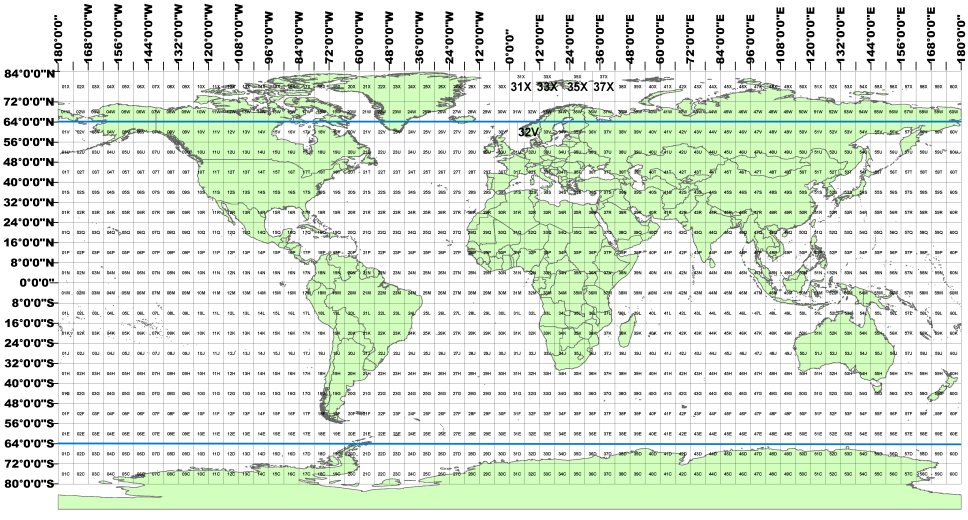


Figure 3: **Geographic coverage of extracted image locations.** We extract image samples in the marginal ice zone above 64° North and below 64° South, as illustrated by the spatial distribution of locations in the two hemispheres.

B Construction of Ground-Truth Sea-Ice Concentration

We construct the reference sea-ice concentration (SIC) label for each Sentinel-2 scene or crop from a pixelwise snow/ice mask. Let $\rho_{B03}(p)$ and $\rho_{B11}(p)$ denote the reflectance of pixel p in the green and short-wave-infrared (SWIR) bands, respectively. We first compute the Normalized Difference Snow Index (NDSI) as

$$\mathrm{NDSI}(p) = \frac{\rho_{B03}(p) - \rho_{B11}(p)}{\rho_{B03}(p) + \rho_{B11}(p)}. \tag{1}$$

High NDSI values are indicative of snow or ice, but water can also produce high NDSI values. We therefore require both $\mathrm{NDSI}(p) > 0.4$ and near-infrared reflectance $\rho_{B08}(p) > 0.11$, where Band 8 is the Sentinel-2 near-infrared (NIR) band. The resulting binary ice mask is

$$M_{\mathrm{ice}}(p) = \mathbb{I}[\mathrm{NDSI}(p) > 0.4 \ \wedge \ \rho_{B08}(p) > 0.11]\,, \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Finally, for the set Ω of valid pixels in the scene or crop, we calculate SIC as the percentage of pixels classified as ice:

$$\mathrm{SIC} = 100\% \times \frac{\sum_{p\in\Omega} M_{\mathrm{ice}}(p)}{|\Omega|}. \tag{3}$$

Thus, the target lies in $[0, 100\%]$, with 0 indicating no detected sea ice and 100 % indicating that all valid pixels are classified as ice.

C All-band Satellite Image Example

To illustrate the full image input used in the satellite-image task, Figures 4 show all 12 spectral bands together with the true-color image (TCI). Each subfigure caption reports the band identifier, the corresponding band description, and the native spatial resolution.

D QA Construction

When we construct QA questions, the prompts we use will vary depending on the format and task of the question. For questions of different formats, we use the following prompts.

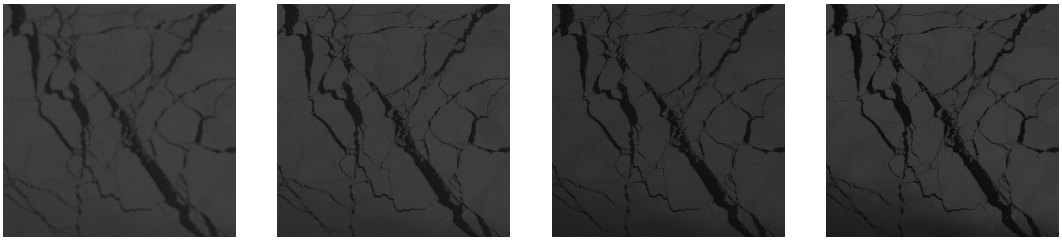
Prompt for Different Question Formats

Free Form QA
Free-form questions and answers are designed for open-ended responses. These questions may require detailed explanations, recommendations, comparisons, descriptions, or reasoning-based answers. The response should be flexible and adapted to the complexity of the question.

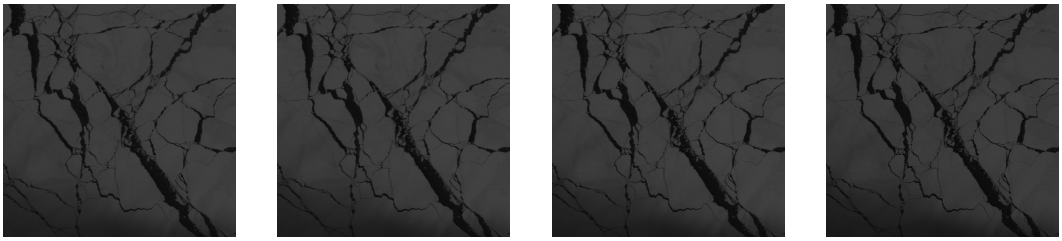
For example, a question about predicting sea ice concentration may require a list of evaluation measures, each with a brief explanation of its purpose and relevance.

Format Rules:
1. Answers may be written in any appropriate format, including sentences, bullet points, numbered lists, tables, or paragraphs.
2. There is no strict restriction on answer length.
3. The structure of the answer should match the question type and complexity.
4. When the question asks for explanations, provide sufficient detail and reasoning.
5. When the question asks for recommendations or methods, include practical suggestions and, where appropriate, explain why they are suitable.
6. Avoid unnecessary formatting constraints unless explicitly requested by the question.
7. The answer should be clear, complete, and directly relevant to the question.

Multiple Choice
Multiple-choice questions consist of a question prompt followed by four answer options,



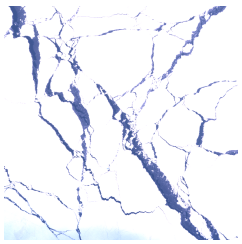
(a) B01: Coastal aerosol, 60 m (b) B02: Blue, 10 m (c) B03: Green, 10 m (d) B04: Red, 10 m



(e) B05: Vegetation red edge, 20 m (f) B06: Vegetation red edge, 20 m (g) B07: Vegetation red edge, 20 m (h) B08: NIR, 10 m



(i) B09: Water vapour, 60 m (j) B10: SWIR–Cirrus, 60 m (k) B11: SWIR, 20 m (l) B12: SWIR, 20 m



(m) TCI: True-color, 10m

Figure 4: **Example multi-band satellite image** from spectral Bands 1–12 and the corresponding true-color image (TCI).

labelled A, B, C, and D. Exactly one option is correct. The task is to select the correct answer by returning only the corresponding letter.

Format Rules:
1. Each question must include one clear question prompt.
2. Each question must provide exactly four answer options labelled A, B, C, and D.
3. Exactly one option must be correct.
4. The answer must be a single uppercase letter: A, B, C, or D.
5. The answer must not include the full text of the selected option.
6. The answer must not include explanations, reasoning, punctuation, or any additional text.

True or False

True/false questions require determining whether a given statement is correct or incorrect. The task is to evaluate the statement and return only whether it is true or false.

Format Rules:
1. The answer must be exactly one of the following: True or False.
2. Do not include explanations, reasoning, punctuation, or any additional text.
3. Do not rewrite or restate the question.
4. Use the exact capitalisation: True or False.

Fill-in-the-blank

Fill-in-the-blank questions provide a sentence, statement, formula, or short passage with one or more missing words, phrases, values, or symbols. The task is to complete the blank(s) using the most accurate answer based on the question context.

Format Rules:
1. The answer must contain only the exact word, phrase, number, value, or symbol that correctly fills the blank.
2. Do not include explanations, reasoning, punctuation, or additional text unless explicitly requested.
3. If there is more than one blank, provide the answers in the same order as the blanks appear.
4. For multiple blanks, separate answers using commas unless another format is specified.
5. Preserve required units, symbols, spelling, and capitalisation when they are necessary for correctness.
6. Do not rewrite the full sentence unless explicitly requested.

Figure 5: **Prompt for Different Question Formats.**

The prompts for different tasks are shown in Table 2. The complete QA prompt is shown in Figure 7.

Prompt for Different Task Formats

Term Explanation

This task requires defining a technical term, concept, method, variable, or domain-specific expression and explaining its meaning clearly. The answer should describe what the term means, how it is used, and why it is important in the relevant context.

Example:
Question: Define “sea ice concentration” and explain its significance.

Expected Answer:
The response should provide a clear definition of sea ice concentration and explain its relevance, such as its role in climate monitoring, sea ice modelling, navigation, and polar

environmental research.

Response Guidelines:
1. Start with a clear and concise definition.
2. Explain the concept in context.
3. Include its significance, application, or relevance where appropriate.
4. Use simple but accurate language.
5. Include examples if they help clarify the concept.
6. Avoid unnecessary technical details unless the question requires them.

Knowledge QA

This task requires answering factual questions by providing clear, accurate, and well-structured explanations. The response should describe relevant facts, characteristics, patterns, distributions, processes, or relationships in sufficient detail.

Example:
Question: Describe the distribution of sea ice extent that occurs in the Arctic.

Expected Answer:
The response should explain how Arctic sea ice extent varies spatially and seasonally, including where sea ice is commonly found, how it expands and retreats through the year, and what factors influence its distribution.

Response Guidelines:
1. Provide a direct factual answer to the question.
2. Describe key characteristics, patterns, or distributions clearly.
3. Include relevant spatial, temporal, physical, or environmental factors where appropriate.
4. Use precise domain terminology while keeping the explanation understandable.
5. Organise the answer logically, such as from general description to specific details.
6. Include examples where they help clarify the answer.
7. Avoid unsupported speculation or unrelated information.

Fact Verification

This task requires checking whether a given claim, statement, value, or reported result is accurate. The response should assess the claim against reliable evidence, data, definitions, or accepted domain knowledge.

Example:
Question: Verify whether the reported sea ice concentration data is correct.

Expected Answer:
The response should determine whether the reported sea ice concentration data is accurate by comparing it with the relevant source data, expected range, calculation method, metadata, or scientific definition.

Response Guidelines:
1. Clearly state whether the claim appears to be correct, incorrect, partially correct, or unverifiable.
2. Explain the basis for the verification.
3. Compare the claim with reliable evidence, reference data, or accepted definitions where available.
4. Identify any inconsistencies, missing information, or uncertainty.
5. Avoid making unsupported conclusions when the evidence is insufficient.
6. Provide a concise explanation of the verification result.

Analysis

This task requires using logical reasoning to examine information, interpret data, identify

relationships, and draw well-supported conclusions. The response should explain how the evidence, observations, or given information support a particular conclusion.

Example:
Question: Determine which evidence supports a given conclusion.

Expected Answer:
The response should evaluate the available evidence, identify the most relevant supporting information, and explain why it supports the conclusion. It should distinguish strong evidence from weak, unrelated, or insufficient evidence.

Response Guidelines:
1. Identify the main question, claim, or conclusion being analysed.
2. Examine the relevant data, evidence, observations, or arguments.
3. Explain the logical relationship between the evidence and the conclusion.
4. Distinguish between strong, weak, irrelevant, or contradictory evidence.
5. Draw a clear conclusion based on the available information.
6. Mention uncertainty or limitations when the evidence is incomplete.
7. Avoid unsupported assumptions or conclusions not justified by the evidence.

Relation Extraction

This task requires identifying, analysing, and summarising relationships or interconnections between entities, concepts, variables, processes, or events. The response should explain how the entities are connected, the nature of their relationship, and the significance of that relationship in the relevant context.

Example:
Question: Summarise the relationships between sea ice and ocean currents.

Expected Answer:
The response should describe how sea ice and ocean currents influence each other, such as how ocean currents affect sea ice transport, melting, formation, and distribution, and how sea ice can modify ocean circulation through freshwater input, brine rejection, and changes in surface heat exchange.

Response Guidelines:
1. Identify the main entities or concepts involved.
2. Describe the relationship between them clearly.
3. Explain the direction of influence where relevant, such as whether one entity affects the other or whether the relationship is mutual.
4. Include mechanisms, causes, effects, or dependencies that explain the relationship.
5. Mention contextual factors, such as spatial, temporal, physical, or environmental conditions, where appropriate.
6. Distinguish direct relationships from indirect or uncertain relationships.
7. Avoid listing entities without explaining how they are connected.

Calculation

This task requires performing mathematical or numerical operations to obtain a correct answer. The calculation may involve one or more steps, such as applying formulas, aggregating values, comparing quantities, deriving statistics, or computing results from given data.

Example:
Question: Calculate the annual minimum sea ice extent from the data.

Expected Answer:
The response should identify the relevant sea ice extent values for each year, determine the minimum value within each annual period, and report the calculated annual minimum sea ice extent clearly.

Response Guidelines:
1. Identify the required quantity to be calculated.
2. Use the relevant data, formula, or mathematical operation.
3. Show intermediate steps when the calculation is multi-step or when explanation is needed.
4. Provide the final numerical answer clearly.
5. Include units where applicable.
6. Ensure the result is consistent with the data and the calculation method.
7. Mention any assumptions, missing data, or uncertainty that may affect the result.

Figure 6: **Prompt for Different Task Formats.**

Complete QA Construction Prompt

System Prompt
You are a highly skilled scientist with deep expertise in reading, interpreting, and analysing scientific literature. You will be given the content of a research paper. Your task is to generate one high-quality question and its corresponding answer based strictly on the paper content and the instructions below.

Your output must be precise, scientifically accurate, self-contained, and directly grounded in a specific detail from the paper.

Inputs:
1. Format Prompt:
{Format_Prompt}

2. Task Prompt:
{Task_Prompt}

3. Research Paper:
input_file with {file_id}

Instructions
1. Follow the Specified QA Format
Generate the question and answer according to the provided Format Prompt. The answer must follow all formatting rules specified in the Format Prompt.

2. Follow the Specified Task Type
Generate the question according to the provided Task Prompt. The question should match the intended task type, such as term explanation, knowledge QA, fact verification, analysis, relation extraction, calculation, multiple choice, true/false, or fill-in-the-blank.

3. Select One Specific Detail
Do not generate a broad or general question about the whole paper. Select one specific and meaningful detail from the paper, such as:
- a key finding;
- a methodological choice;
- a dataset characteristic;
- an experimental setting;
- a numerical result;

- a model assumption;
- a limitation;
- a discussion point;
- a comparison between methods;
- a scientific interpretation.

The question should require careful understanding of that specific detail. Avoid overly simple, generic, or surface-level questions.

4. Ensure Content Relevance and Accuracy
The question, answer, and explanation must be directly derived from the paper content. Do not introduce facts, claims, assumptions, or interpretations that are not supported by the paper.

5. Make the Question Self-Contained
The question must be understandable without access to the original paper. Do not use phrases such as:
- “this paper”;
- “this article”;
- “the study”;
- “according to the paper”;
- “as mentioned above”;
- “the authors”;
- “we” or “our.”

Instead, rewrite the context as an independent statement inside the question when needed.

6. Make the Answer Self-Contained
The answer must directly answer the question and must not depend on references to the paper. Do not use vague references such as “it,” “they,” or “this result” unless the meaning is completely clear from the question.

7. Provide an Answer Explanation
Each generated question must include an explanation for the answer. The explanation should:
- justify why the answer is correct;
- refer to the relevant scientific reasoning, data, method, or statement;
- be written as independent statements;
- avoid unsupported analysis or speculation;
- be clear and logically organised.

8. Avoid Paper-Level Commentary
Do not analyse or summarise the paper as a whole. Focus only on the selected detail needed to construct the question, answer, and explanation.

9. Output Format
Return exactly one dictionary-like JSON object with the following keys:

"question": "The generated self-contained question.",
"answer": "The correct answer to the question.",
"explanation": [
"Explanation step 1.",
"Explanation step 2.",
"Explanation step 3."
]

Output Rules:
- Return only the dictionary-like JSON object.

- Do not include markdown formatting.
- Do not include additional comments before or after the output.
- Ensure the output is valid JSON-compatible text.
- The explanation must be a list of strings.

Figure 7: **Complete QA Construction Prompt.**

After constructing the QA pair, we will also refine the QA. The main purpose is to remove extra references that may exist in the QA, such as mentioning papers or non-existent figures in the question. The refine prompt is as follows.

QA Refine Prompt

System Prompt

You are an expert editor for scientific question-answer datasets. Your task is to revise the provided question, answer, and explanation so that they are fully self-contained and do not refer to any source document, paper, article, study, manuscript, or similar external text.

The revised output must preserve the original meaning, factual content, and dictionary structure.

Input:
QA_Dictionary

Instructions

1. Remove References to Source Documents
Remove or rewrite any expressions that refer to the original source, including phrases such as:
- “this paper”
- “this article”
- “the study”
- “the manuscript”
- “according to the paper”
- “as described in the article”
- “the authors”
- “we” or “our”
- any similar source-dependent wording.

2. Make the QA Self-Contained Revise the question, answer, and explanation so that they can be understood independently without requiring access to the original paper or article.

The question should include enough context to be answered directly.

The answer should directly respond to the question without referring to the source document.

The explanation should justify the answer using complete, independent statements.

3. Preserve the Original Meaning Do not change the scientific meaning, factual content, correct answer, or logical reasoning. Only revise wording to improve independence, clarity, and completeness.

4. Preserve the Dictionary Format Return the output using the same dictionary structure with the following keys:
- "question"
- "answer"

- "explanation"

The "explanation" field must remain a list of strings.

Output Format:
Return exactly one JSON-compatible dictionary:

"question": "The revised self-contained question.",
"answer": "The revised answer.",
"explanation": ["Revised explanation step 1.",
"Revised explanation step 2.",
"Revised explanation step 3."
]
Output Rules:
- Return only the dictionary.
- Do not include markdown formatting.
- Do not include comments before or after the dictionary.
- Do not add new information that was not present in the original QA.
- Do not change the original answer unless wording changes are required for self-containment.

Figure 8: **QA Refine Prompt.**

E QA EXAMPLES

QA Example 1 (Free QA)

Question
In polarimetric synthetic aperture radar (SAR) feature evaluation for Arctic sea ice classification, what is "mutual information," and how can it be used to assess both feature relevance and feature redundancy?

Explanation
Step 1: Mutual information can be expressed as $I(X;Y) = H(X) - H(X|Y)$, where H(X) is entropy and H(X|Y) is conditional entropy, meaning it measures how much knowing Y reduces uncertainty about X.

Step 2: For relevance analysis, the sea ice class label is treated as one variable and a polarimetric feature as the other; features with higher mutual information have stronger predictive value for sea ice classification.

Step 3: For redundancy analysis, two polarimetric features are compared directly; high mutual information between similarly relevant features indicates that the features contain overlapping information, such as the strong redundancy between H(q) and scattering diversity δ.

Answer
Mutual information is an information-theoretic measure that quantifies how much information two variables share. In polarimetric SAR sea ice classification, it can be used to measure how strongly a feature is related to the sea ice class label and how much information two different features contain in common.
When one variable is the sea ice class and the other is a polarimetric feature, mutual information indicates feature relevance: a higher value means that the feature is more useful for distinguishing ice classes. When both variables are polarimetric features, mutual information

indicates feature redundancy: a high value means that the two features carry similar information, so one may be unnecessary for classification.
This is important because full-polarimetric SAR data can generate many features, some of which are computationally expensive or redundant. Mutual information helps identify useful features and supports pruning the feature set before classification.

Figure 9: **QA Example 1.**

QA Example 2 (Multiple Choice)

Question
In the Neutral Hills Uplands of southern Alberta, the prominent glacitectonic constructional landforms include the Neutral Hills at 120 m high, Misty Hills at 85 m high, Nose Hill at 100 m high, and Mud Buttes at 50 m high. What is the difference between the combined heights of the Neutral Hills and Misty Hills and the combined heights of Nose Hill and Mud Buttes?

A. 35 m
B. 45 m
C. 55 m
D. 65 m

Explanation
Step 1: The combined height of the Neutral Hills and Misty Hills is 120 m + 85 m = 205 m.

Step 2: The combined height of Nose Hill and Mud Buttes is 100 m + 50 m = 150 m.

Step 3: The difference is 205 m - 150 m = 55 m, which corresponds to option C.

Answer
C

Figure 10: **QA Example 2.**

QA Example 3 (True or False)

Question
True or false: In the sedimentary model of the Norwegian Channel Ice Stream, the mid-shelf Zone 2 along western Norway is interpreted as the main depositional zone where sediments eroded from the inner shelf accumulated, rather than as a subglacial transport zone dominated by high sediment fluxes and mega-scale glacial lineations.

Explanation
Step 1: The sedimentary model of the Norwegian Channel Ice Stream distinguishes three process zones: Zone 1 is dominated by erosion, Zone 2 by sediment transport, and Zone 3 by deposition.

Step 2: Zone 2 is interpreted as an area where sediments eroded from the inner shelf were transported subglacially, probably by soft-sediment deformation, rather than mainly deposited.

Step 3: Zone 2 contains mega-scale glacial lineations associated with high sediment fluxes, with only localized sediment accretion linked to individual lineations.

Answer
FALSE

Figure 11: **QA Example 3.**

QA Example 4 (Fill in the Blanks)

Question
In a surface-classification evaluation of ICESat-2 ATL07 data, 584 true thin-ice points were classified as dark leads and 396 true thin-ice points were classified as specular leads out of 18,894 true thin-ice points. What percentage of true thin-ice points were detected as leads?
Explanation
Step 1: The number of true thin-ice points detected as leads is the sum of thin-ice points classified as dark leads and specular leads: 584 + 396 = 980.

Step 2: The total number of true thin-ice points is 18,894.

Step 3: The detected percentage is 980 / 18,894 $\times 100\%$ = 5.1879%, which rounds to 5.19%.

Answer
5.19%

Figure 12: **QA Example 4.**

F Evaluation Prompts

This section provides the exact prompts used to evaluate models on the visual, numerical, and textual components of SeaIce-Bench.

F.1 Sea-Ice Concentration Evaluation Prompt

For sea-ice concentration (SIC) prediction, we first upload the 13 satellite images corresponding to Bands 1–12 and the true-color image (TCI) to the inference server. We then use the following prompt:

```
These are satellite images from Band 1 to Band 12 and the true-colour band (TCI) of a specific
     region. Based on the visual data, predict the Sea Ice Concentration (SIC).
Output only the numerical value in two decimal places representing the percentage (0.00 to
    100.00).
```

F.2 Sea-Ice Extent Evaluation Prompt

For monthly sea-ice extent forecasting, each model receives 40 years of monthly sea-ice extent data and the following prompt:

```
You are given 40 years of monthly sea ice extent data in CSV format for a specified polar
    region.

Task:
Predict the sea ice extent for one month after the latest month in the CSV data.

Requirements:
- Use only the provided CSV data and the specified polar region.
- Identify the latest available month in the CSV.
- Predict the sea ice extent for the following month.
- Output only one numerical value.
- The output must be rounded to two decimal places.
- Do not include units, explanations, labels, markdown, or any additional text.

Polar region:
{region}

CSV data:
{csv_content}

Predicted sea ice extent:
```

F.3 LITERATURE-BASED QA EVALUATION PROMPTS

For the literature-based QA evaluation, questions cover six task categories: Term Explanation, Knowledge QA, Fact Verification, Analysis, Relation Extraction, and Calculation. Each question is presented using one of four standardized prompt templates according to its answer format. In each template, the placeholder `{question}` is replaced with the corresponding benchmark question.

F.3.1 MULTIPLE-CHOICE QUESTIONS

```
You are answering a multiple-choice question.

Question:
{question}

Instructions:
- Select the single correct option.
- Respond with only one uppercase letter: A, B, C, or D.
- Do not include the option text.
- Do not include explanations, punctuation, or any additional text.

Answer:
```

F.3.2 TRUE/FALSE QUESTIONS

```
You are answering a true/false question.

Question:
{question}

Instructions:
- Determine whether the statement is correct or incorrect.
- Respond with exactly one word: True or False.
- Use the exact capitalisation: True or False.
- Do not include explanations, punctuation, or any additional text.

Answer:
```

F.3.3 FILL-IN-THE-BLANK QUESTIONS

```
You are answering a fill-in-the-blank question.

Question:
{question}

Instructions:
- Fill in the blank with the exact word, phrase, number, value, or symbol required.
- Respond with only the content that completes the blank.
- Do not include explanations or any additional text.

Answer:
```

F.3.4 FREE-FORM QUESTIONS

```
You are answering an open-ended question.

Question:
{question}

Instructions:
- Provide a clear, accurate, and complete answer.
- Use an appropriate structure based on the question, such as a paragraph, bullet points,
    numbered steps, or a table.
- Include relevant details, explanations, examples, or reasoning when needed.
- Keep the answer directly focused on the question.
- Do not include unrelated information.

Answer:
```

G ADAPTED MMSEAICE BASELINE

MMSeaIce is a state-of-the-art sea-ice mapping method based on a multi-task U-Net architecture (Chen et al., 2024). Because the visual task in SEAICE-BENCH differs from the original formulation, we adapt its architecture rather than using the published model unchanged. The original formulation processes a single image and predicts sea ice concentration (SIC) in 10-percentage-point intervals, whereas our task provides 13 co-registered images at each location—12 spectral bands and one true-color image (TCI)—and requires a continuous numerical SIC prediction. We therefore modify the input layer and replace the original output with a scalar-regression head, then train the adapted model from scratch using our training data. We evaluate this task-specific non-LLM baseline on both the full-scene and cropped-image settings.

Input images. Each instance contains the 13 co-registered images listed in Table 6. The source images have different spatial resolutions depending on the spectral band. The implementation applies the required image-specific preprocessing and spatial alignment before combining the images into the 13-channel tensor consumed by the network.

Table 6: Required source-image dimensions for the Adapted MMSeaIce input.

Bands	Cropped-image dimensions	Full-scene dimensions
B01, B09, B10	183×183	1830×1830
B02, B03, B04, B08, TCI	1098×1098	10980×10980
B05, B06, B07, B11, B12	549×549	5490×5490

Encoder. The adapted model is a contracting U-Net-style encoder with a scalar-regression output. Each `DoubleConv` block consists of a 3×3 convolution with padding 1 and no bias, GroupNorm, and a SiLU activation, followed by a second 3×3 convolution, GroupNorm, and SiLU activation. The number of GroupNorm groups is chosen as the largest divisor of the output-channel count that does not exceed 8. Each subsequent encoder block first applies 2×2 max pooling and then a `DoubleConv` block. With the default filter configuration $(32, 64, 128, 256)$, the channel progression is

$$13 \longrightarrow 32 \longrightarrow 64 \longrightarrow 128 \longrightarrow 256.$$

Regression head. The final feature map is summarized by both global average pooling and global maximum pooling. The two pooled vectors are concatenated and passed through

$$\text{Linear} \longrightarrow \text{SiLU} \longrightarrow \text{Dropout} \longrightarrow \text{Linear} \longrightarrow \text{Sigmoid}.$$

The final sigmoid constrains the prediction to $[0, 1]$, which is converted to a continuous SIC value on the $[0\%, 100\%]$ scale. Detailed training settings, hyperparameters, and source code are documented in the accompanying GitHub repository, which will be released or linked in accordance with the area chair's guidance.

H EVALUATION METRICS

Accuracy (Acc.) For closed-form questions (multiple-choice, fill-in-the-blank, and true/false) with deterministic correct answers, we calculate accuracy as the exact match rate between model outputs and reference solutions.

Semantic Similarity (SS) Devika et al. (2021) For free-response questions, we additionally measure the conceptual alignment between model outputs and reference answers using sentence embeddings. Let $\boldsymbol{h}_i^{\text{pred}}$ and $\boldsymbol{h}_i^{\text{ref}}$ denote the sentence-transformer embeddings of the model prediction and the reference answer for instance i, respectively. We compute semantic similarity using cosine similarity,

$$\text{SS}_i = \cos\left(\boldsymbol{h}_i^{\text{pred}}, \boldsymbol{h}_i^{\text{ref}}\right) = \frac{\boldsymbol{h}_i^{\text{pred}} \cdot \boldsymbol{h}_i^{\text{ref}}}{\left\|\boldsymbol{h}_i^{\text{pred}}\right\|_2 \left\|\boldsymbol{h}_i^{\text{ref}}\right\|_2}. \tag{4}$$

We report the average cosine similarity over all free-response evaluation instances.

Worst-Case-Penalized RMSE (WCP-RMSE) Computing RMSE only over valid model outputs can favor models that fail to produce predictions for difficult instances. To account for both predictive quality and completion coverage, we additionally report a worst-case-penalized RMSE (WCP-RMSE). This evaluation follows the cost-based reject-option principle, in which abstained or rejected predictions incur a predefined loss (Chow, 1970; Franc et al., 2023).

Let the test set contain N instances. For model m, let y_i be the ground-truth target, $\hat{y}_{m,i}$ be the model prediction, and

$$v_{m,i} = \begin{cases} 1, & \text{if the output is valid and scoreable,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

An output is considered valid only if the complete response contains exactly one parseable finite numerical value, contains no additional explanatory text or numerical alternatives, uses the required units, and lies within the predefined physically admissible target range. Missing outputs, refusals, malformed responses, non-finite values, out-of-range values, and responses that embed a candidate prediction within additional prose or calculations are treated as invalid.

Suppose the admissible target interval is $[a, b]$, and define

$$B = b - a \tag{6}$$

as the maximum possible absolute error between two values in this interval. For visual sea ice concentration estimation, the admissible interval is $[0, 100]$ percent, so $B_{\mathrm{SIC}} = 100$ percentage points and each invalid completion is assigned a squared-error penalty of $10{,}000$ squared percentage points. Sea-ice extent is measured in millions of square kilometers. The highest historical extents are 19.76 million km^2 in the Southern region and 16.34 million km^2 in the Northern region. We therefore round these values up to define the conservative region-specific admissible intervals $[0, 20]$ and $[0, 17]$, respectively. Thus,

$$B_{\mathrm{South}} = 20 \quad \text{and} \quad B_{\mathrm{North}} = 17, \tag{7}$$

in units of million km^2. Each invalid completion is consequently assigned the squared-error penalty $B^2_{\mathrm{South}} = 400$ for the Southern region or $B^2_{\mathrm{North}} = 289$ for the Northern region, in squared units of million km^2. Using the value of B corresponding to the evaluated region, the worst-case-penalized RMSE of model m is therefore

$$\text{WCP-RMSE}_m = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \Big[v_{m,i} \left(\hat{y}_{m,i} - y_i \right)^2 + \left(1 - v_{m,i} \right) B^2 \Big]}. \tag{8}$$

Completion Rate (CR) Inspired by answer-rate and valid-rate analyses in prior LLM evaluations (Wang et al., 2024; Zhan et al., 2024), we report completion rate for the visual and numerical tasks to quantify the fraction of evaluation instances that produce a valid, scoreable result. Let $v_i \in \{0, 1\}$ denote whether the prediction for instance i can be successfully parsed into the required answer format under our evaluation protocol. We define completion rate as

$$\text{CR} = \frac{1}{n} \sum_{i=1}^{n} v_i \tag{9}$$

For single-value numerical tasks, the entire response must consist of exactly one parseable finite value that satisfies the prescribed range and formatting constraints; a scalar embedded in prose, calculations, a range, or alternative estimates is not sufficient. Structured numerical tasks such as motion-vector forecasting must instead conform completely to their required output schema. CR complements WCP-RMSE by separately showing the model's ability to return an evaluable visual or numerical prediction. We do not report CR for textual QA because missing, malformed, and incorrect textual responses are retained in the denominator and directly reduce Acc. or SS.

I Cropped-Image Error Analysis

The cropped-image setting requires a model to inspect 12 co-registered spectral-band crops and a true-color image (TCI), integrate their complementary information, and return a single numerical sea ice concentration (SIC). The responses below illustrate that failures can occur before the numerical estimation itself: a model may not recognize the attached images, may refuse the visual task, or may attend to only one image rather than jointly reasoning over all bands. These failure cases reveal weak visual grounding and substantial instability. The response wording is retained below.

False detection of missing images. Kimi K2.6 frequently behaved as though the image inputs had not been delivered. A representative response is:

> **Kimi K2.6:** No images were provided in your message, so I cannot visually assess the satellite bands or TCI to estimate Sea Ice Concentration. Please upload the images and I will provide the numerical prediction.

This response is not an abstention caused by visually ambiguous sea-ice conditions. Instead, it is an input-recognition failure: the model denies the presence of the images that constitute the test instance. It consequently returns no numerical SIC value. This failure pattern helps explain the 0.54% completion rate of Kimi K2.6 in Table 3.

Qwen 3.6 Flash exhibits a related modality-level refusal:

> **Qwen 3.6 Flash:** I cannot view or analyze satellite imagery, so I am unable to calculate or predict Sea Ice Concentration from visual data. Please attach the image or specify the region/data source for accurate assistance.

Here, the model first claims that it cannot analyze satellite imagery and then asks the user to attach an image, although the spectral crops and TCI are already included. The response therefore fails both the visual-interpretation requirement and the numerical-output requirement. Such refusals contribute to the lower 79.08% completion rate of Qwen 3.6 Flash.

Inconsistent image access and visual grounding. DeepSeek V4 Flash shows a particularly unstable pattern. In some trials it acknowledges the spectral bands and gives a confident numerical prediction:

> **DeepSeek V4 Flash, numerical response:** Based on the visual data from the satellite bands, the Sea Ice Concentration (SIC) appears to be nearly complete coverage with very few open water areas visible.
>
> **95.50%**

In other trials, however, it rejects the same type of task because it claims that the images are missing:

> **DeepSeek V4 Flash, refusal response:** I cannot provide a numerical prediction because no satellite images were included in your message. Please attach the images so I can analyze the bands and estimate the sea ice concentration.

The contrast between these responses is important. The first response asserts that the model has inspected the satellite bands, whereas the second asserts that no satellite images exist. This inconsistency suggests that the model's ability to access or retain multiple visual inputs is unreliable. Moreover, the numerical response provides no band-specific evidence for its description of nearly complete ice coverage. The refusal responses reduce the model's completion rate to 82.67%.

Failure to follow the required single-value output format. The prompt instructed each model to output only a single numerical SIC value between 0 and 100. Gemini 2.5 Flash does not follow this instruction. Although its response ends with the value 30.00, it first provides a lengthy explanation containing numerous percentages, intermediate calculations, and alternative estimates. The response also shows that the model largely discards the spectral-band sequence and bases its estimate on the final rendered image:

> **Gemini 2.5 Flash:** To predict the Sea Ice Concentration (SIC) from the provided visual data, I will primarily rely on the last image, which is a highly contrasted processed image showing distinct bluish-grey regions against a white background. The preceding images (bands 1–12 and TCI) are largely black or very dark, making it impossible to visually discern sea ice features or open water. It's assumed that the bluish-grey areas in the final image represent sea ice, and the white background represents open water or an unclassified area within the region of interest.
>
> Upon careful visual inspection and estimation of the area covered by the bluish-grey "sea ice" features:

1. **Distribution:** The sea ice is concentrated predominantly in the right half and upper-right quadrant of the image, with sparser patches in the upper-left.
2. **Density:** The ice features are not a solid sheet but appear as broken floes or textured regions. Some areas on the right are quite dense, while others are more scattered.
3. **Estimation Method:** I will divide the image mentally and estimate the coverage.
   - **Left half (approximately 50% of the total image area):** This part contains very sparse ice, mostly in the upper-left corner. I estimate an average coverage of about 2–3% within this half. Thus, $0.50 \times 0.025 = 0.0125$, or 1.25% of the total image area.
   - **Right half (approximately 50% of the total image area):** This part contains much denser ice. Some sections appear to be 60–80% covered, while others are less. I estimate an average coverage of about 55% within this half. Thus, $0.50 \times 0.55 = 0.275$, or 27.5% of the total image area.

Summing these estimates: Total SIC = 1.25% (from left half) + 27.5% (from right half) = 28.75%.

Another approach, considering a split into 60% left and 40% right:

- **Left** 60% **of the image area:** Very little ice, perhaps 5% coverage within this section. Thus, $0.60 \times 0.05 = 0.03$, or 3% of the total image.
- **Right** 40% **of the image area:** Substantial ice, perhaps 70% coverage within this section. Thus, $0.40 \times 0.70 = 0.28$, or 28% of the total image.

Total SIC = $3\% + 28\% = 31.00\%$.

Given these close estimations, I will choose a value that represents this range. The ice concentration appears to be roughly one-third of the image, possibly slightly less.

Final estimate: Combining visual assessment and these calculations, a value around 30% seems most appropriate.

30.00%

Gemini 3 Flash Preview likewise fails to return only the requested numerical value. It appears to consider more of the band sequence, but embeds its final estimate in a qualitative explanation:

Gemini 3 Flash Preview: Based on the visual analysis of the satellite images across various bands, there is a clear region of concentrated sea ice floes in the upper-right quadrant of the image. This area features the characteristic fractured texture of ice and is significantly brighter in most spectral bands compared to the rest of the image.

While the True Color Image (TCI) appears very bright and mostly white, this is likely due to cloud cover, as the underlying spectral bands (especially the near-infrared and shortwave infrared bands like 8, 11, and 12) show most of the remaining area to be quite dark, which is typical for open water.

Estimating the area covered by the visible ice floes and the more solid ice mass in the top-right corner, they occupy approximately 30–35% of the total image area.

32.14%

Although 32.14 is within the admissible range, the response also contains reasoning; it therefore does not satisfy the instruction to output only one numerical value.

Implications for validity and scoring. These examples expose two distinct error levels. At the completion level, a response is valid only when its entire output consists of one finite numerical value in the physically admissible SIC range $[0, 100]$, with no additional text, intermediate values, ranges, or alternative estimates. Refusals, missing outputs, malformed or non-finite values, out-of-range values, and verbose responses that merely contain an extractable scalar are therefore invalid. At the prediction level, a format-compliant value can still be inaccurate because the images were misinterpreted or not jointly processed. Accordingly, CR measures strict output validity, while WCP-RMSE additionally incorporates the numerical error of valid predictions and assigns the maximum penalty to invalid ones.

J SEA-ICE EXTENT ERROR ANALYSIS

The invalid sea-ice extent completions fall into two distinct categories. For Qwen 3.6 Flash and Kimi K2.6, every failed prediction was returned or recorded as `NaN`, rather than as a finite sea-ice extent. One plausible explanation is that these models interpreted the historical missing-value markers as evidence that the requested forecast month was also missing. In particular, some earlier years in the input series contain missing observations; the models may have propagated this missing-data pattern instead of inferring a numerical forecast. This behavior is especially consequential for Kimi K2.6, whose completion rates are only 60.5% and 63.3% for the Northern and Southern regions, respectively. Qwen 3.6 Flash exhibits the same failure mode less frequently, with invalid outputs confined to a small fraction of the Southern-region cases.

Gemini 2.5 Flash exhibits a different failure mode: it often derives a candidate prediction but does not follow the required single-value output format. The following representative failed completion is reproduced in full:

```
To predict the sea ice extent for one month after the latest month, we will follow these steps:


1. **Parse the data**: Load the provided CSV data.
2. **Identify the latest month and its extent**: Find the last entry in the data.
3. **Determine the target prediction month**: This will be the month immediately following the
      latest month.
4. **Calculate the average historical month-to-month change**: Compute the average change in
     sea ice extent from the identified latest month (May) to the target prediction month (
     June) across all available years in the dataset.
5. **Predict the extent**: Add this average change to the latest observed extent.
6. **Format the output**: Round the result to two decimal places.

**Data Analysis:**

- The latest month in the CSV data is **2019, May (month 5)**, with an extent of **8.85**.
- The target prediction month is **2019, June (month 6)**.
- We need to calculate the average difference between June's extent and May's extent for all
     available years. We'll exclude the `-9999.0` missing values, though they do not occur in
     May or June for any year that would affect this calculation.

Let's list the May and June extents and their differences for all years from 1980 to 2018 (as
     1979 only has June, and 2019 only has May data for this pair):

| Year | May Extent | June Extent | Change (June - May) |
|---:|---:|---:|---:|
| 1980 | 8.86 | 12.49 | 3.63 |
| 1981 | 9.35 | 13.11 | 3.76 |
| 1982 | 10.76 | 13.42 | 2.66 |
| 1983 | 9.53 | 12.64 | 3.11 |
| 1984 | 9.75 | 13.23 | 3.48 |
| 1985 | 10.15 | 13.43 | 3.28 |
| 1986 | 9.30 | 12.69 | 3.39 |
| 1987 | 9.69 | 12.68 | 2.99 |
| 1988 | 9.60 | 13.09 | 3.49 |
| 1989 | 10.30 | 13.58 | 3.28 |
| 1990 | 9.72 | 13.08 | 3.36 |
| 1991 | 10.28 | 13.07 | 2.79 |
| 1992 | 9.92 | 12.81 | 2.89 |
| 1993 | 10.10 | 13.15 | 3.05 |
| 1994 | 10.41 | 13.56 | 3.15 |
| 1995 | 10.42 | 13.29 | 2.87 |
| 1996 | 10.92 | 13.87 | 2.95 |
| 1997 | 10.12 | 13.12 | 3.00 |
| 1998 | 10.33 | 13.31 | 2.98 |
| 1999 | 10.27 | 13.65 | 3.38 |
| 2000 | 10.92 | 13.89 | 2.97 |
| 2001 | 10.33 | 13.42 | 3.09 |
| 2002 | 9.51 | 12.38 | 2.87 |
| 2003 | 10.74 | 13.99 | 3.25 |
| 2004 | 10.47 | 13.78 | 3.31 |
| 2005 | 10.29 | 13.29 | 3.00 |
| 2006 | 9.46 | 13.34 | 3.88 |
| 2007 | 9.65 | 13.29 | 3.64 |
| 2008 | 11.05 | 14.06 | 3.01 |
| 2009 | 10.93 | 13.91 | 2.98 |
| 2010 | 10.64 | 14.41 | 3.77 |
| 2011 | 10.09 | 13.33 | 3.24 |
| 2012 | 10.46 | 13.55 | 3.09 |
| 2013 | 10.92 | 14.15 | 3.23 |
```

```
| 2014 | 11.52 | 14.69 | 3.17 |
| 2015 | 11.72 | 14.47 | 2.75 |
| 2016 | 10.10 | 13.24 | 3.14 |
| 2017 | 9.01 | 12.41 | 3.40 |
| 2018 | 9.32 | 12.89 | 3.57 |

Sum of all "Change (June - May)" values = 122.95

Number of years = 39 (from 1980 to 2018)

Average historical change (June - May) = 122.95 / 39 = 3.1525641025641026

**Prediction Calculation:**

Latest extent (May 2019) = 8.85

Predicted extent (June 2019) = Latest extent + Average historical change

Predicted extent = 8.85 + 3.1525641025641026 = 12.002564102564103

Rounding to two decimal places, the predicted sea ice extent is 12.00.

12.00
```

Although the final line contains a plausible forecast, the complete response also contains explanatory prose and numerous intermediate numerical values. It was therefore marked invalid and assigned the worst-case penalty in WCP-RMSE. Thus, the Gemini 2.5 Flash failures primarily reflect instruction-following and output-format noncompliance, whereas the Qwen 3.6 Flash and Kimi K2.6 failures reflect an inability to return a finite forecast in the presence of historical missing-value markers.

K Motion-Vector Error Analysis

Inspection of unsuccessful generations revealed that low completion rates were often caused by output-format violations rather than by the absence of numerical predictions alone. A representative Gemini 2.5 Flash response, for example, began with Python imports (e.g., `pandas`, `numpy`, and `io`) and placed the tabular data inside a code string. In other words, the model returned its reasoning or a code scaffold instead of the requested output-only, comma-separated prediction records. A shortened excerpt is shown below:

```
import pandas as pd
import numpy as np
import io

data = """year,week,u1,u2,u3,u4
1981,25,1.0,0.8999999761581421,...
```

Because no valid prediction table conforming to the required schema was produced, the response could not be parsed and was counted as incomplete.

Unexpected year and week fields. Gemini 3 Flash Preview frequently returned numerical predictions but prepended the year and week to the final record, although the instruction required only the ordered motion-vector values. A representative shortened output is:

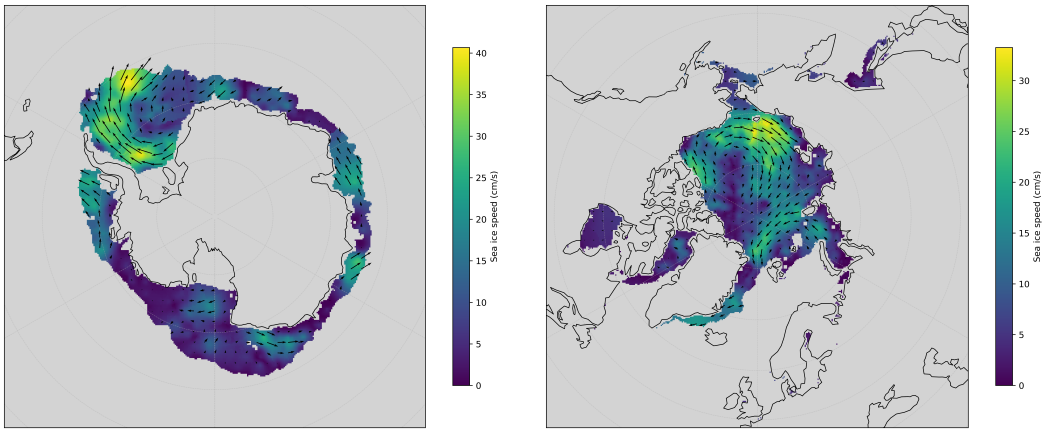
```
2020,8,-1.1,-1.2,-1.2,...
```

Here, `2020` and `8` are temporal metadata rather than predicted motion components. Their inclusion shifts the expected column alignment and produces the wrong number of fields, so the output cannot be parsed under the prescribed schema even though it contains candidate vector values. This recurrent instruction-following error helps explain the particularly low Northern-window completion rate of $0.2 \pm 0.1\%$ for Gemini 3 Flash Preview. Its Southern-window completion rate is higher but still incomplete at $72.7 \pm 15.6\%$, showing that schema compliance also varies across the two evaluation windows.

The inspected outputs from the open-weight models—Qwen 3.6 Flash, DeepSeek V4 Flash, and Kimi K2.6—showed similar tendencies to include reasoning or code rather than only the final records. An additional recurrent failure was delimiter substitution: although the required fields had to be separated

by commas, some generations used vertical bars (|). Such responses could appear structurally plausible to a reader but failed the automated CSV parser and therefore did not satisfy the completion criterion. These observations help explain the low or unstable completion rates of Gemini 2.5 Flash, Gemini 3 Flash Preview, Qwen 3.6 Flash, and DeepSeek V4 Flash in Table 4: their WCP-RMSE reflects not only numerical prediction error, but also the penalty for failing to return a complete, machine-readable output in the prescribed format.

L SEA ICE MOTION VECTOR DATA EXAMPLES



(a) Southern Hemisphere sea ice motion vector example.

(b) Northern Hemisphere sea ice motion vector example.

Figure 13: Examples of the sea ice motion vector data used in SEAICE-BENCH for the Southern and Northern Hemispheres.

M EXPERT VERIFICATION OF THE TEXTUAL QA CORPUS

We randomly sampled and manually reviewed 5,770 QA pairs, corresponding to approximately 10.7% of the complete 53,785-pair textual QA corpus. The reviewed examples are stored in files ending with `_high.csv` and cover all six task families. Sea-ice experts manually reviewed every example in this subset. Table 7 reports the number of reviewed QA pairs by task type.

Task type	Reviewed QA pairs
Term Explanation	980
Knowledge QA	945
Fact Verification	961
Analysis	951
Relation Extraction	956
Calculation	977
Total	**5,770**

Table 7: Distribution of the randomly sampled textual QA pairs manually reviewed by sea-ice experts.

N HUMAN EVALUATION OF THE TEXTUAL BENCHMARK

N.1 SURVEY DESIGN AND PARTICIPANT SAFEGUARDS

We designed and administered a human-participant survey to evaluate the question-answering component of SEAICE-BENCH and examine whether performance varied across participants with different educational and domain backgrounds. To limit cognitive burden and respondent fatigue, we targeted a completion time of approximately 20 minutes. This duration is consistent with empirical evidence on the maximum web-survey length that respondents consider acceptable (Revilla & Ochoa, 2017), while longer questionnaires have been associated with lower participation and response quality (Galesic & Bosnjak, 2009). To cover every benchmark subtask within this time budget, each participant answered two multiple-choice questions from each of the six task families, for a total of 12 benchmark questions. The multiple-choice format provided uniform scoring across participants; established item-writing guidance emphasizes the importance of well-constructed stems and answer options in such assessments (Haladyna et al., 2002). Every evaluated question was sampled from SEAICE-BENCH and

had previously been verified by a domain expert. Representative administered items are provided in Appendix N.5. Participants were instructed to answer without the assistance of LLMs.

The survey began with an informed-consent form and included separate robot and attention checks. Participants were informed that participation was voluntary and that they could discontinue the study at any time, consistent with the Belmont Report's principle of respect for persons and its requirement that consent be voluntary (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The attention check was used to identify participants who did not follow the survey instructions, a standard motivation for instructional manipulation checks (Oppenheimer et al., 2009).

The order of the benchmark questions was randomized independently for each participant. This randomization reduced the possibility that a fixed questionnaire position would systematically favor or disadvantage a particular task family. The design is relevant because response quality can deteriorate as a survey progresses, and randomized question placement can reveal or mitigate fatigue-related position effects (Jeong et al., 2023).

N.2 Sample-size rationale and inferential scope

The human evaluation was designed as a task-oriented reference for the difficulty of a verified subset of the textual benchmark, not as an econometric, epidemiological, or population-level behavioral study intended to estimate a generalizable causal effect or a precise population parameter. Consequently, the original benchmark design did not include a conventional a priori power analysis. The unit of interest was performance on sampled benchmark questions under a controlled survey protocol, and each participant completed a balanced 12-question subset rather than the full 53,785-item textual benchmark.

This use of a relatively small, task-oriented human reference group is consistent with established practice in LLM and multimodal benchmark research. WebArena estimated human performance using five computer-science graduate students on a sample of 170 tasks and reported a human completion rate of 78.24% (Zhou et al., 2024). MMMU recruited 90 senior university students—three for each of its 30 subjects—and each participant answered 30 subject-specific questions rather than the full benchmark (Yue et al., 2024). MMLU reported 34.5% performance from unspecialized Amazon Mechanical Turk participants but did not report a matched expert cohort; its cited expert-level estimate of 89.8% was instead derived from 95th-percentile examination results and informed estimates when examination statistics were unavailable (Hendrycks et al., 2021). These precedents show that human results in AI benchmarking are often used as descriptive or diagnostic reference points without requiring every participant to complete the full benchmark or presenting an a priori statistical power analysis.

The survey was added during the limited rebuttal period in response to a reviewer request for a human reference and is not the paper's principal contribution. Within that period, we used manually verified questions, predefined participant classifications, informed consent, randomized question ordering, robot and attention checks, and explicit response-exclusion criteria. A substantially larger, formally powered human-subject study would require additional recruitment and resources, together with any further institutional human-research ethics review or approval required for an expanded protocol, and could not realistically be completed within the rebuttal period. We therefore present the current 24-participant sample as a transparent exploratory baseline. It provides preliminary evidence about the difficulty of the verified QA subset for the recruited groups, but it is not conclusive evidence of a universal expert gap or a basis for population-level inference.

N.3 Quality control and participant groups

Before computing the results, we retained only responses from participants who satisfied all of the following quality-control criteria:

1. provided full informed consent;
2. passed the robot check;
3. passed the attention check; and

4. exhibited a plausible relationship between their self-reported question difficulty and the time spent answering the questions.

The final criterion used response time as survey paradata rather than as a direct measure of ability. Response latency depends on question complexity and participant characteristics (Yan & Tourangeau, 2008), while survey-quality research recommends combining multiple indicators when screening for careless responses (Meade & Craig, 2012). Accordingly, the timing check complemented, rather than replaced, the robot and attention checks.

We assigned valid participants to the following mutually exclusive groups, applying the sea-ice-expert criterion first and then classifying the remaining participants by degree:

Sea-ice expert: held a PhD degree and rated both their sea-ice knowledge and climate knowledge as at least 3 on a five-point scale, where 1 indicated *Novice* and 5 indicated *Expert*;

PhD: held a PhD degree but did not meet the sea-ice-expert criterion;

Master: held a master's degree; and

Bachelor: held a bachelor's degree.

Applying the sea-ice-expert criterion first prevented participants from being counted in both the sea-ice-expert and PhD groups. After quality-control filtering, the analysis included 24 participants: 3 sea-ice experts, 7 other PhD holders, 9 master's degree holders, and 5 bachelor's degree holders. Each participant answered two questions from each of the six task families, for a total of 12 benchmark questions per participant.

N.4 Human-evaluation results

Table 8 reports accuracy and perceived difficulty for every participant group and task family. Pooled across the six task families, sea-ice experts achieved the highest accuracy at 83.3% (30/36), followed by the PhD group at 66.7% (56/84), the master's group at 65.7% (71/108), and the bachelor's group at 51.7% (31/60). This ordering is preliminary evidence that the verified QA subset differentiates the recruited groups, although performance also varies substantially across task families. Given the modest sample size and the particularly small sea-ice-expert subgroup, the results are descriptive and exploratory; they do not establish a population-level difference or a universal expert gap.

Classification	Question type	N participants	N valid answers	N correct	Accuracy	Mean difficulty (1–5)	N valid difficulty ratings
Sea-ice expert	Term explanation	3	6	6	1.0000	3.1667	6
Sea-ice expert	Knowledge	3	6	5	0.8333	3.5000	6
Sea-ice expert	Fact verification	3	6	6	1.0000	3.1667	6
Sea-ice expert	Analysis	3	6	4	0.6667	3.5000	6
Sea-ice expert	Relation extraction	3	6	4	0.6667	3.3333	6
Sea-ice expert	Calculation	3	6	5	0.8333	1.8333	6
PhD	Term explanation	7	14	10	0.7143	3.8571	14
PhD	Knowledge	7	14	8	0.5714	4.0000	14
PhD	Fact verification	7	14	10	0.7143	3.6429	14
PhD	Analysis	7	14	7	0.5000	3.9286	14
PhD	Relation extraction	7	14	9	0.6429	3.6429	14
PhD	Calculation	7	14	12	0.8571	2.0714	14
Master	Term explanation	9	18	12	0.6667	4.0000	18
Master	Knowledge	9	18	10	0.5556	4.2222	18
Master	Fact verification	9	18	11	0.6111	3.7778	18
Master	Analysis	9	18	14	0.7778	3.9444	18
Master	Relation extraction	9	18	8	0.4444	4.1111	18
Master	Calculation	9	18	16	0.8889	3.0000	18
Bachelor	Term explanation	5	10	8	0.8000	4.4000	10
Bachelor	Knowledge	5	10	4	0.4000	4.4000	10
Bachelor	Fact verification	5	10	1	0.1000	4.1000	10
Bachelor	Analysis	5	10	4	0.4000	4.4000	10
Bachelor	Relation extraction	5	10	7	0.7000	4.6000	10
Bachelor	Calculation	5	10	7	0.7000	2.5000	10

Table 8: Exploratory human-evaluation results by participant group and textual task family. Difficulty was self-reported on a five-point scale, with higher values indicating greater perceived difficulty. Accuracy is computed over valid answers; subgroup differences are reported descriptively and are not population-level estimates.

N.5 Verified questions administered in the human survey

The following questions were sampled from the textual question dataset, verified before use, and administered in the human-participant survey. They are grouped by answer format. For each item, we provide the reference answer, the verification notes used to check the answer, and the supplied verification source when one was available. Calculation answers marked as human-verified were independently checked before survey administration.

N.5.1 Multiple-choice questions

Term explanation: mega-scale glacial lineations

Question. On the bed of the Norwegian Channel Ice Stream, which option best defines "mega-scale glacial lineations" (MSGLs) and explains their significance?

A. Highly elongate subglacial ridges and grooves formed beneath fast-flowing ice, used as indicators of ice-flow direction and dynamic sediment transport or deformation.

B. Circular depressions produced by gas escape from marine sediments, used mainly to identify postglacial fluid seepage.

C. Large transverse sediment wedges deposited at grounding-line stillstands, used to mark pauses during ice-margin retreat.

D. Curved erosional marks cut by high-pressure meltwater vortices, used to identify localized subglacial meltwater erosion.

Reference answer. A.
Verification notes.

- Mega-scale glacial lineations are highly elongate landforms characteristic of soft-bedded ice-stream beds.
- Along the Norwegian Channel Ice Stream, MSGLs are interpreted as features formed beneath fast-flowing ice in a dynamic sedimentary system with high sediment fluxes.
- Their orientation records former ice-flow direction, and their presence supports interpretations of subglacial sediment transport, deformation, and localized accretion.

Verified source. Clark, Tulaczyk, Stokes, and Canals. "A groove-ploughing theory for the production of mega-scale glacial lineations, and implications for ice-stream mechanics." *Journal of Glaciology*, 49(165):240–256, 2003. https://doi.org/10.3189/172756503781830719

Term explanation: CryoSat-2 radar freeboard

Question. In CryoSat-2 processing of Arctic summer sea ice, what does "radar freeboard" mean and why is it important?

A. The horizontal distance between adjacent sea-ice floes, used to estimate lead width.

B. The height difference between the radar-measured sea-ice floe surface and the local sea level inferred from leads, used as a key input for estimating sea-ice thickness.

C. The fraction of the sea-ice surface covered by melt ponds, used to classify optical satellite images.

D. The calibrated radar backscatter coefficient from sea ice, used directly as a substitute for ice thickness.

Reference answer. B.
Verification notes.

- Radar freeboard is the elevation of the sea-ice surface relative to nearby sea level, where sea level is inferred from radar returns classified as leads.
- In CryoSat-2 summer sea-ice processing, freeboard is calculated from the elevation difference between radar returns from classified ice floes and nearby leads.
- Radar freeboard is important because it is the primary measurement converted into sea-ice thickness, although that conversion requires accounting for uncertainties such as snow loading, meltwater, ice density, and radar ranging bias.

Verified source. Ricker, Hendricks, Kaleschke, Tian-Kunze, King, and Haas. "A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data." *The Cryosphere*, 11(4):1607–1623, 2017. https://doi.org/10.5194/tc-11-1607-2017

Knowledge QA: glacier thinning in the Garhwal Himalaya

Question. For 205 glaciers of at least $0.1\ \mathrm{km}^2$ in the Alaknanda, Bhagirathi, and Mandakini basins of the Garhwal Himalaya, what change in mean annual thinning rate occurred between 2000–2015 and 2015–2020?

A. From $0.07 \pm 0.09\ \mathrm{m\,a}^{-1}$ to $0.31 \pm 0.19\ \mathrm{m\,a}^{-1}$.

B. From $0.31 \pm 0.19\ \mathrm{m\,a}^{-1}$ to $0.07 \pm 0.09\ \mathrm{m\,a}^{-1}$.

C. From $1.00 \pm 0.15\ \mathrm{m\,a}^{-1}$ to $1.00 \pm 0.33\ \mathrm{m\,a}^{-1}$.

D. From $0.57 \pm 0.09\ \mathrm{m\,a}^{-1}$ to $0.53 \pm 0.21\ \mathrm{m\,a}^{-1}$.

Reference answer. A.
Verification notes.

- The mean annual thinning rate for the 205 glaciers was $0.07 \pm 0.09\ \mathrm{m\,a}^{-1}$ during 2000–2015.
- The mean annual thinning rate increased to $0.31 \pm 0.19\ \mathrm{m\,a}^{-1}$ during 2015–2020.
- This change indicates an acceleration of glacier thinning in the later observation period.

Verified source. Bhambri et al. "Heterogeneity in glacier thinning and slowdown of ice movement in the Garhwal Himalaya, India." *Science of the Total Environment*, 875:162625, 2023. https://doi.org/10.1016/j.scitotenv.2023.162625

Knowledge QA: Greenland Neoglacial expansion periods

Question. Glacier-margin evidence across Greenland indicates three distinct Neoglacial glacier-expansion periods. Which set of age ranges correctly identifies these three expansion periods?

A. 8,000–6,000 yrs BP; 4,500–3,500 yrs BP; 1,000–500 yrs BP.

B. 2,500–1,700 yrs BP; 1,250–950 yrs BP; 700–50 yrs BP.

C. 5,000–4,000 yrs BP; 3,100–2,400 yrs BP; 600–0 yrs BP.

D. 1,950–1,250 yrs BP; 1,000–700 yrs BP; 250–50 yrs BP.

Reference answer. B.
Verification notes.

- The three Greenland Neoglacial glacier-expansion periods consist of two older advances and one younger advance associated with the Little Ice Age.
- The two older advances occurred at 2,500–1,700 yrs BP and 1,250–950 yrs BP.
- The younger Neoglacial advance occurred at 700–50 yrs BP and corresponds to the Little Ice Age glacier expansion.

Verified source. Kjær et al. "Glacier response to the Little Ice Age during the Neoglacial cooling in Greenland." *Earth-Science Reviews*, 227:103984, 2022. `https://doi.org/10.1016/j.earscirev.2022.103984`

Fact verification: Antarctic Bottom Water exports

Question. In a 1958–2018 ACCESS-OM2-01 ocean–sea-ice simulation, the mean Antarctic Bottom Water exports for four formation regions were: Weddell Sea, 1.5 Sv; Prydz Bay, 2.8 Sv; Adélie Coast, 1.5 Sv; and Ross Sea, 2.9 Sv. Verify the claim: Prydz Bay had the lowest mean export among the four formation regions.

A. Correct; Prydz Bay had a mean export of 0.4 Sv, lower than all other regions.

B. Incorrect; Prydz Bay had a mean export of 2.8 Sv, while the Weddell Sea and Adélie Coast each had lower mean exports of 1.5 Sv.

C. Partially correct; Prydz Bay had the lowest surface water-mass transformation but not the lowest export.

D. Unverifiable; the simulation did not report regional Antarctic Bottom Water export values.

Reference answer. B.
Verification notes.

- The mean Antarctic Bottom Water exports were 1.5 Sv for the Weddell Sea, 2.8 Sv for Prydz Bay, 1.5 Sv for the Adélie Coast, and 2.9 Sv for the Ross Sea.
- Prydz Bay was not the lowest-export region because both the Weddell Sea and Adélie Coast had lower mean exports.
- The claim is therefore incorrect based on the specified regional export values for the 1958–2018 simulation.

Verification. The comparison and answer key were independently checked by a human reviewer.

Fact verification: British–Irish Ice Sheet arrival at Garret Hill

Question. Verify the claim: At Garret Hill on the north Norfolk coast, optically stimulated luminescence dating of glaciofluvial outwash beneath the Holkham Till constrains the arrival of the British–Irish Ice Sheet to immediately after 22.8–21.5 ka. Which verification is accurate?

A. Correct; two OSL ages from the outwash are 22.8 ± 1.8 ka and 21.5 ± 1.3 ka, and the overlying till indicates ice arrival after those deposits formed.

B. Incorrect; the relevant OSL ages are 31.6 ± 2.1 ka and 25.8 ± 2.4 ka from Dogger Bank outwash, not Garret Hill.

C. Incorrect; the Garret Hill age constraint is based only on Holocene radiocarbon ages from marine shells.

D. Unverifiable; no dated sediment underlying the Holkham Till is present at Garret Hill.

Reference answer. A.
Verification notes.

- The Garret Hill section contains glaciofluvial outwash assigned to lithofacies GA 2 beneath the brown massive diamict interpreted as Holkham Till.
- Two optically stimulated luminescence samples from GA 2 yielded ages of 22.8 ± 1.8 ka and 21.5 ± 1.3 ka.
- Because the Holkham Till overlies the dated outwash, ice arrival at the north Norfolk coast is constrained to after deposition of the outwash, immediately after approximately 22.8–21.5 ka.

Verified source. Roberts et al. "Ice marginal dynamics of the last British–Irish Ice Sheet in the southern North Sea: Ice limits, timing and the influence of the Dogger Bank." *Quaternary Science Reviews*, 198:181–207, 2018. `https://doi.org/10.1016/j.quascirev.2018.08.010`

Analysis: NorHAPS sea-ice edge and marginal ice zone accuracy

Question. In an evaluation of sea-ice concentration products against Norwegian Ice Service ice charts for the period 1 January 2021 to 30 April 2022, which evidence best supports the conclusion that NorHAPS with local analytical optimal nudging (LAON) produced the most accurate sea-ice edge and marginal ice zone among the compared products?

A. NorHAPS LAON had the lowest mean integrated ice-edge error of 1.78×10^5 km^2 and the lowest mean integrated marginal-ice-zone error of 3.54×10^5 km^2 among all compared products.

B. AMSR2 sea-ice concentration was assimilated into NorHAPS LAON, so AMSR2 necessarily had lower ice-edge and marginal-zone errors than all model analyses.

C. TOPAZ4 had a larger mean integrated ice-edge error than NEMO but a smaller mean integrated marginal-ice-zone error than neXtSIM, proving TOPAZ4 was the most accurate overall.

D. NorHAPS Free_run had the highest mean integrated ice-edge error and integrated marginal-ice-zone error, showing that data assimilation was unnecessary for accurate sea-ice edge simulation.

Reference answer. A.
Verification notes.

- The relevant performance metrics were integrated ice-edge error and integrated marginal-ice-zone error, both evaluated against Norwegian Ice Service ice charts.
- NorHAPS LAON had mean errors of 1.78×10^5 km^2 for the ice edge and 3.54×10^5 km^2 for the marginal ice zone, lower than the corresponding values for AMSR2, SSMIS, NEMO, TOPAZ4, neXtSIM, and NorHAPS Free_run.
- Lower values of these error metrics indicate better agreement with the ice chart, so the lowest values directly support the conclusion that NorHAPS LAON performed best for both representations.

Verified source. Wang, Ali, and Wang. "Local analytical optimal nudging for assimilating AMSR2 sea ice concentration in a high-resolution pan-Arctic coupled ocean (HYCOM 2.2.98) and sea ice (CICE 5.1.2) model." *The Cryosphere*, 17:4487–4510, 2023. https://doi.org/10.5194/tc-17-4487-2023

Analysis: Type 2 ice-stream lateral moraines

Question. Type 2 ice-stream lateral moraines are interpreted as lateral marginal moraines formed at the boundary between a marine-terminating ice stream and adjacent terrain free of grounded ice. Which observation most directly supports this interpretation?

A. They maintain a constant width, thickness, and cross-sectional shape along their length.

B. They contain internal reflections dipping away from the former trough, indicating sediment progradation away from the ice stream.

C. They occur as discontinuous segments along terrestrial ice-stream margins.

D. They are generally less than 3.5 km wide and less than 60 m thick.

Reference answer. B.
Verification notes.

- Type 2 ice-stream lateral moraines contain internal dipping reflections that indicate sediment progradation away from the former ice stream.
- Progradation away from the trough indicates lateral delivery and accumulation of sediment at an ice margin, consistent with formation beside terrain not occupied by grounded ice.
- The other options describe characteristics of Type 1 lateral shear moraines, interpreted as subglacial features formed between fast-flowing ice streams and slower-flowing grounded ice.

Verified source. Batchelor and Dowdeswell. "Lateral shear-moraines and lateral marginal-moraines of palaeo-ice streams." *Quaternary Science Reviews*, 151:1–26, 2016. https://doi.org/10.1016/j.quascirev.2016.08.020

Relation extraction: resonance in a continuously stratified ocean

Question. In an ice–ocean coupled theory for ice-band formation, resonance occurs when the ice-band propagation speed matches the phase speed of an internal inertia-gravity wave. How does representing the ocean as continuously stratified change this resonance relationship compared with a 1.5-layer ocean model?

A. Continuous stratification prevents resonance because internal-wave phase speeds are always faster than ice-band propagation speeds.

B. Continuous stratification provides infinitely many baroclinic modes, so higher modes with slower phase speeds can satisfy resonance even when ice-band propagation speed decreases.

C. Continuous stratification makes resonance depend only on sea-ice thickness, eliminating any dependence on internal-wave phase speed.

D. Continuous stratification allows resonance only for the first baroclinic mode, producing a stricter cutoff wind speed than in a 1.5-layer model.

Reference answer. B.
Verification notes.
- The resonance condition requires the ice-band propagation speed to match the phase speed of an internal inertia-gravity wave.
- A continuously stratified ocean has infinitely many baroclinic modes, and higher modes have slower phase speeds.
- Because slower higher-mode waves remain available, resonance can persist at lower ice-band propagation speeds, unlike in a 1.5-layer model with only one baroclinic mode and a cutoff speed.

Verified source. Saiki et al. "Mechanism of ice-band pattern formation caused by resonant interaction between sea ice and internal waves in a continuously stratified ocean." *Progress in Oceanography*, 190:102474, 2021. `https://doi.org/10.1016/j.pocean.2020.102474`

Relation extraction: RES-UNET and Conv-CRF in RUF

Question. In the RUF architecture for sea-ice floe segmentation in synthetic aperture radar (SAR) images, which option correctly describes the relationship between the RES-UNET component and the Convolutional Conditional Random Field (Conv-CRF) component?

A. RES-UNET generates coarse pixelwise segmentation maps that serve as unary potentials for Conv-CRF, while Conv-CRF uses those maps with spatial pairwise potentials to refine the final segmentation boundaries.

B. Conv-CRF first removes speckle noise from SAR images, and RES-UNET then classifies entire images into ice-covered and open-water scenes.

C. RES-UNET replaces Conv-CRF during inference because Conv-CRF is used only to initialize model weights before training.

D. Conv-CRF provides labelled training patches to RES-UNET, while RES-UNET only performs land masking before segmentation.

Reference answer. A.
Verification notes.
- RUF combines a residual UNET with a Convolutional Conditional Random Field for pixel-level sea-ice floe segmentation.
- The RES-UNET computes expressive features and produces coarse segmentation maps, which are used as unary inputs to the Conv-CRF.
- The Conv-CRF incorporates spatial pairwise potentials to reduce spurious predictions and refine object boundaries in the final segmentation mask.

Verified source. Nagi, Kumar, Sola, and Scott. "RUF: Effective Sea Ice Floe Segmentation Using End-to-End RES-UNET-CRF with Dual Loss." *Remote Sensing*, 13(13):2460, 2021. `https://doi.org/10.3390/rs13132460`

Calculation: Greenland percolation facies coverage

Question. Satellite L-band radiometry maps perennial firn aquifer areas over $64{,}000\ \mathrm{km}^2$ and ice-slab areas over $76{,}000\ \mathrm{km}^2$ within a Greenland percolation facies extent of $5.8 \times 10^5\ \mathrm{km}^2$. What percentage of the percolation facies is covered by the combined perennial firn aquifer and ice-slab areas?

A. 11%.

B. 24%.

C. 32%.

D. 41%.

Reference answer. B.
Verification notes.
- The combined mapped area is $64{,}000 + 76{,}000 = 140{,}000\ \mathrm{km}^2$.
- The covered percentage is $(140{,}000/580{,}000) \times 100\%$.
- This equals approximately 24.1%, which rounds to 24%.

Verification. The calculation was independently checked by a human reviewer.

Calculation: Martian north polar layered deposits

Question. A glacier-ice volume in the Oudemans crater basin is estimated to be approximately $200{,}000\ \mathrm{km}^3$. If this amount equals one-fifth of the volume of the Martian north polar layered deposits, what is the implied total volume of the north polar layered deposits?

A. $40{,}000\ \mathrm{km}^3$.

B. $200{,}000\ \mathrm{km}^3$.

C. $1{,}000{,}000\ \mathrm{km}^3$.

D. $5{,}000{,}000\ \mathrm{km}^3$.

Reference answer. C.
Verification notes.

- The required quantity is the total volume of the Martian north polar layered deposits.
- If $200{,}000\ \mathrm{km}^3$ represents one-fifth of that total, the total volume is $200{,}000 \times 5$.
- The implied volume is therefore $1{,}000{,}000\ \mathrm{km}^3$.

Verification. The calculation was independently checked by a human reviewer. *Verified source.* Yin, Moon, and Day. "Landform evolution of Oudemans crater and its bounding plateau plains on Mars: Geomorphological constraints on the Tharsis ice-cap hypothesis." *Icarus*, 360:114332, 2021. `https://doi.org/10.1016/j.icarus.2021.114332`

N.5.2 TRUE/FALSE QUESTION

True/false: Rb/K paleosalinity thresholds

Question. True or false: In geochemical analysis of Carboniferous silt/clay rhythmites from Ibaré, Rio Grande do Sul, the Rb/K paleosalinity proxy classifies Rb/K values below 4×10^{-3} as freshwater, values above 6×10^{-3} as marine, and intermediate values as brackish.
Reference answer. True.
Verification notes.

- The Rb/K paleosalinity proxy is based on the substitution of Rb^+ for K^+ in clay minerals such as illite.
- Rb is much more abundant in seawater than in freshwater, so higher Rb/K ratios are interpreted as indicating greater marine influence.
- For these Carboniferous rhythmites, values below 4×10^{-3} indicate freshwater conditions, values above 6×10^{-3} indicate marine conditions, and values between these limits indicate brackish conditions.

Verification. The item and reference answer were verified before survey administration.

N.5.3 FILL-IN-THE-BLANK QUESTIONS

Fill in the blank: NorHAPS assimilation method

Question. In the HYCOM–CICE NorHAPS modelling system, what is the name of the sea-ice concentration assimilation method that updates the model at every time step using an analytically derived optimal nudging coefficient and nudges the model toward an optimal estimate rather than directly toward the observation?
Reference answer. Local analytical optimal nudging.
Verification notes.

- Local analytical optimal nudging is used to assimilate AMSR2 sea-ice concentration in the NorHAPS HYCOM–CICE modelling system.
- The method differs from ordinary nudging because the model state is gradually nudged toward an optimal estimate rather than simply toward the observation.
- The nudging coefficient is derived analytically, and assimilation is performed online at every model time step.

Verified source. Wang, Ali, and Wang. "Local analytical optimal nudging for assimilating AMSR2 sea ice concentration in a high-resolution pan-Arctic coupled ocean (HYCOM 2.2.98) and sea ice (CICE 5.1.2) model." *The Cryosphere*, 17:4487–4510, 2023. `https://doi.org/10.5194/tc-17-4487-2023`

Fill in the blank: SAR tidal-bias correction

Question. In synthetic aperture radar (SAR) differential range offset tracking (DROT), vertical tidal displacement can introduce an artificial ground-range velocity bias. If $\Delta\zeta$ is the double-difference sea level and θ_i is the radar incidence angle, the differential tidal bias ΔS translated to ground range is calculated as $\Delta S =$ _____.
Reference answer. $\Delta\zeta / \tan\theta_i$.
Verification notes.

- DROT uses the side-looking geometry of SAR, in which vertical tidal motion produces an apparent horizontal displacement component in the radar line of sight.
- When two slant-range displacement fields with the same time interval are differenced, steady horizontal ice motion and uniform vertical bias are cancelled.
- The remaining differential tidal bias translated to ground range is $\Delta S = \Delta\zeta / \tan\theta_i$.

Verified source. Wallis, Hogg, Zhu, and Hooper. "Change in grounding line location on the Antarctic Peninsula measured using a tidal motion offset correlation method." *The Cryosphere*, 18:4723–4740, 2024. `https://doi.org/10.5194/tc-18-4723-2024`

Fill in the blank: Last Glacial Maximum monsoon dynamics

Question. During the Last Glacial Maximum, a cool, dry central Core Monsoon Zone indicated weak Indian Summer Monsoon conditions, while the peripheral Core Monsoon Zone showed influence from winter precipitation associated with the Northeast Monsoon. This contrast was interpreted as evidence that the Northeast Monsoon was strengthened by an enhanced _____.
Reference answer. Siberian High.
Verification notes.

- The central Core Monsoon Zone experienced a cool and dry climate during the Last Glacial Maximum, indicating weak Indian Summer Monsoon conditions.
- The peripheral Core Monsoon Zone showed winter-precipitation influence caused by the Northeast Monsoon during the same interval.
- A strengthened Northeast Monsoon influence is linked with an enhanced Siberian High.

Verified source. Quamar, Banerji, Thakur, and Kar. "Hydroclimatic changes in the Core Monsoon Zone of India since the Last Glacial Maximum: An overview of the palynological data and correlation with the marine and continental records." *Palaeogeography, Palaeoclimatology, Palaeoecology*, 633:111844, 2024. `https://doi.org/10.1016/j.palaeo.2023.111844`

Fill in the blank: CFD–DEM fluid-force application

Question. In CFD–DEM simulations of ship interaction with broken ice, some ice pieces in noncontact mode showed unrealistic turning because the fluid forces were applied to the _____ rather than to the boundaries of the DEM particles.
Reference answer. Center of the DEM particles.
Verification notes.

- The numerical algorithm applied fluid forces at a simplified representative point for each DEM ice particle.
- This simplification saved computational resources and allowed many particles to be included.
- Because the boundaries of the DEM particles were not directly captured, some noncontact ice pieces could be pushed upright or overturn unrealistically.

Verified source. Xue et al. "A combined experimental and numerical approach to predict ship resistance and power demand in broken ice." *Ocean Engineering*, 292:116476, 2024. `https://doi.org/10.1016/j.oceaneng.2023.116476`

Fill in the blank: Antarctic Bottom Water contribution

Question. In a global ocean–sea-ice simulation of Antarctic Bottom Water export, the mean exports are 2.8 Sv from Prydz Bay, 2.9 Sv from the Ross Sea, and 8.6 Sv from all four formation regions combined. What is the combined Prydz Bay and Ross Sea contribution to total Antarctic Bottom Water export, rounded to one decimal place?
Reference answer. 66.3%.
Verification notes.

- The combined export from Prydz Bay and the Ross Sea is $2.8 + 2.9 = 5.7$ Sv.
- The fraction of total Antarctic Bottom Water export is $5.7/8.6$.
- Multiplying by 100 gives 66.279%, which rounds to 66.3%.

Verification. The calculation was independently checked by a human reviewer.

Provider	Serving stack	Public alias	Version note
OpenAI	OpenAI API	`gpt-4.1`	latest documented update Apr 2025.
OpenAI	OpenAI API	`gpt-5`	latest documented update Aug 2025.
OpenAI	OpenAI API	`gpt-5.2`	latest documented update Dec 2025.
OpenAI	OpenAI API	`gpt-5.4`	latest documented update Mar 2026.
OpenAI	OpenAI API	`gpt-5.5`	latest documented update Apr 2026.
Google	Gemini API	`gemini-2.5-flash`	latest documented update June 2025.
Google	Gemini API	`gemini-2.5-pro`	latest documented update June 2025.
Google	Gemini API	`gemini-3-flash-preview`	latest documented update Dec 2025.
Alibaba Cloud	OpenRouter	`qwen3.6-flash`	latest documented update Apr 2026.
DeepSeek	OpenRouter	`deepseek-v4-flash`	latest documented update Apr 2026.
Moonshot AI	OpenRouter	`kimi-k2.6`	latest documented update Apr 2026.

Table 9: Provider-hosted inference infrastructure and public model identifiers.

N.5.4 FREE-FORM QUESTION

Free-form QA: normalized ATL07 background rate

Question. In the ICESat-2 ATL07 sea-ice surface-type classification context, what is the normalized background-rate variable, often denoted as `background_r_norm` or `bck_norm`, and why is it important for distinguishing open water, thin ice, and snow-covered or thick sea ice?
Reference answer. The normalized background rate measures photons that deviate from the main surface-return height distribution in an ICESat-2 ATL07 height segment after normalization for solar elevation angle. These background photons are associated with instrument noise, scattered sunlight, shadows from clouds or ridges, atmospheric scattering, and specular reflections from smooth surfaces such as leads.
In sea-ice surface classification, the variable is important because different surface types produce different relationships between surface-return and background photons. Specular returns from dark leads can have high photon rates but very low background rates, whereas snow-covered sea ice shows different background-rate behavior. In neural-network classification of Ross Sea ATL07 data, `bck_norm` was the most important input variable for separating surface types, especially open water and snow-covered or thick sea ice.
Verification notes.

- ATL07 aggregates photons into height segments, and photons outside the primary surface-height distribution are treated as background photons related to noise, sunlight scattering, atmospheric effects, shadows, or specular behavior.
- The background rate is normalized because solar illumination varies with latitude, season, and time of day, making the variable more comparable across observations.
- Feature-importance analysis using SHAP values showed that `bck_norm` contributed more than any other input variable to the neural-network surface classification.

Verified source. Koo, Xie, Kurtz, Ackley, and Wang. "Sea ice surface type classification of ICESat-2 ATL07 data by using data-driven machine learning model: Ross Sea, Antarctic as an example." *Remote Sensing of Environment*, 296:113726, 2023. `https://doi.org/10.1016/j.rse.2023.113726`

O INFERENCE INFRASTRUCTURE AND MODEL VERSIONS

All inference is executed through vendor-hosted APIs; we do not self-host model weights or run local GPU serving for any model. The benchmark driver normalizes the input format across providers by sending plain-text questions for textual tasks, CSV files for numerical tasks, and PNG image sets plus the shared textual instruction header for visual tasks. We keep tool use, web search, and code execution disabled so that each run reflects the native response of the requested model to the benchmark input only. Table 9 records the public model aliases together with the provider-reported dated snapshot when one is exposed in the official documentation. It lists the ten models evaluated in the main experiments and GPT-5.5, which is used only for question–answer generation and excluded from evaluation. The model aliases and snapshot strings were verified against the providers' official documentation as accessed in May, 2026.