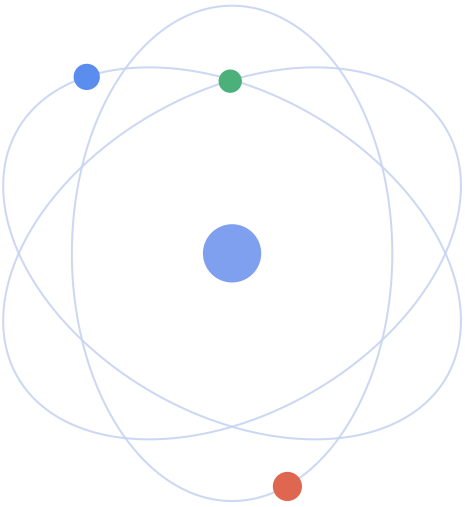
SoTALab

BY ZHIHU

PRODUCT BRIEF
NO.03 / 2026



CONTEXT OVER FORMULA

Physics Reasoning

Physics In-Context Learning Data

**Not more formulas for the model to memorize,
but sharper reading of the current context:
trainable, evaluable In-Context Learning data for long-context reasoning.**

CONTEXT-DEPENDENT REASONING

EXPERT CALIBRATION · AVG@3 REVIEW

POSITION · READ THIS PHYSICS, NOT THE TEXTBOOK

Structured physics In-Context Learning data, training models to truly use context.

Context is not just about stuffing more text into the prompt.It determines how a model understands, retains, and uses information—the real source of application-level differentiation.

From “remembering” to “understanding”: Long context is not better simply because it is longer. When key information is buried in the middle, models often ignore it—the Lost in the Middle problem; the real competition is precise localization and interpretation.

From “chat” to “workflow”: Long tasks require task-state management, not infinitely stretched context; the model must keep track of the current goal, completed actions, and pending steps.

From “putting it in” to “internalizing it”: Some skills should become stable, reflex-like capabilities rather than temporary prompt content.

Physics problems are naturally suited to this training: answers are determinate, but they must obey the current material’s definitions, boundaries, gauges, sectors, and notational conventions.

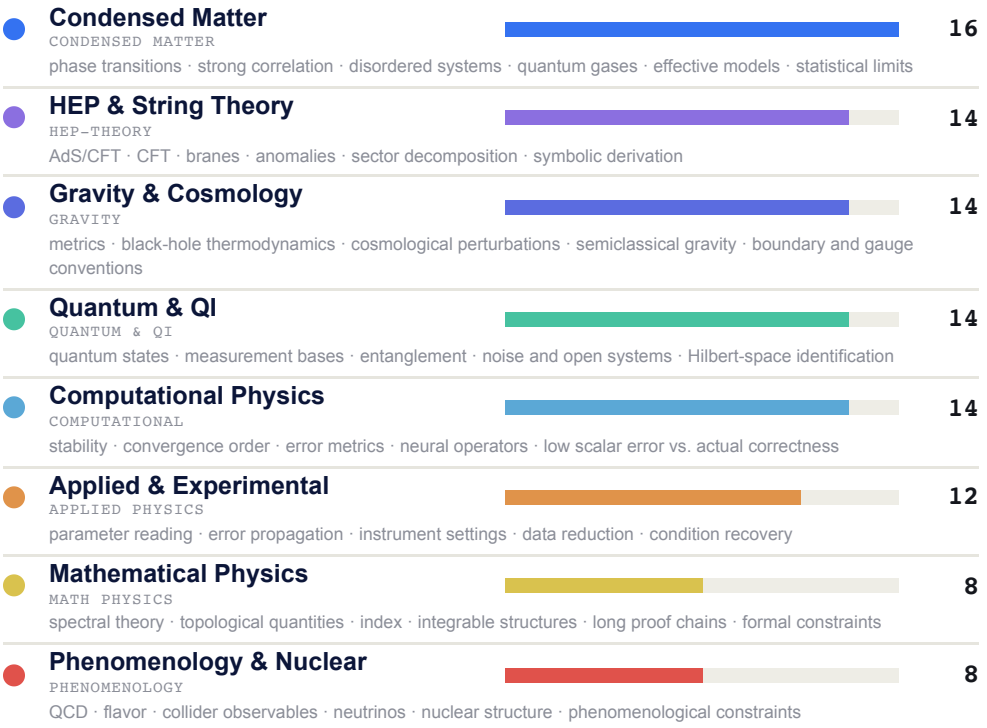
> **Whether the answer is right is the outcome; whether the model understood this context is the capability.**

CONTEXT FAITHFUL · OBJECT SELECTION · VERIFIABLE

COVERAGE · DOMAIN MAP

100 first-level samples, spanning physics breadth and reasoning depth.

The coverage structure limits dominance by any single theory track and fills in quantum, condensed matter, numerical, experimental, and mathematical physics, so the data can expose multiple failure modes.



Coverage is not a pile of topics—each domain maps to a real **In-Context Learning failure mode**: definition drift, wrong object selection, missing multi-sector terms, or branch errors.

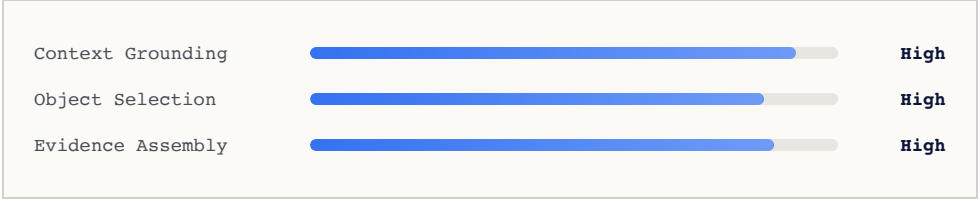
100 long-context physics In-Context Learning samples in total · each includes paper-style context, a reference solution, scoring checkpoints, and multi-model run records.

BENCHMARK ALIGNMENT · EVALUATION MAP

Aligned with mainstream evaluation tracks: scientific QA · research reading · long context · step-level review.

This dataset targets a shared gap exposed by current benchmarks: models improve quickly on short-form scientific QA, but remain unreliable on paper-style context, physics object selection, cross-paragraph evidence assembly, and stable reasoning.

CL-bench CONTEXT LEARNING	Closest alignment: evaluates whether models can learn and apply new knowledge from a given context, rather than merely retrieve long text or reuse pretrained templates.	Context Learning
HLE EXPERT-LEVEL	Adds expert-level physics context-dependent samples to improve condition reading, local definition retention, and answer calibration.	Expert-level
GPQA Diamond GRADUATE QA	Extends short-form scientific reasoning into paper-style context, reducing the chance of high scores from memorized templates.	Graduate-level
PaperBench REPLICATIONBENCH	Provides trainable research-reading subtasks: first make the model perform deterministic reasoning within physics materials.	Replication
ProcessBench PRM800K	Extends error localization from math problems to physics long context, boundary conditions, sector choices, and branch selection.	First Error
MMMU · MathVista MULTIMODAL	Can be extended to paper figures, experimental tables, and formula-mixed inputs to train scientific multimodal In-Context Learning.	Multimodal
LongBench · L-Eval LONG-CONTEXT	Moves from “finding information” to “using information to complete physics reasoning and produce verifiable answers.”	Long Context



EVALUATION EVIDENCE · AVG@3

avg@3 evaluation: strong models still lose substantial points on physics In-Context Learning.

Same context, same prompt, same expert rubric. Each model is run independently three times, averaged on a 0–10 scale, with expert review of the key point-loss locations.

CASE · MAJOR IN-CONTEXT LEARNING FAILURE POINT	GPT-5.5 PRO	GEMINI 3.1 PRO	CLAUDE OPUS 4.8	DEEPSEEK R1	QWEN3 MAX
19 AdS Light-Front Coordinates Definition Tracking	2.4	1.8	3.1	1.3	1.9
27 Strong-Field Energy Kernel Multi-Sector Accounting	3.0	2.2	3.4	1.8	2.5
18 inverse Hadamard Transformation Fidelity	3.8	2.6	4.1	2.3	2.9
02 Charged Sphere Dual Field Boundary Context	2.9	2.1	3.6	1.7	2.2
21 Topological Selector Branch Selection	3.2	2.4	3.8	1.9	2.7
13 CFT Double-Trace Recursion Recursive Consistency	4.0	2.8	4.5	2.5	3.1

Across six 10-point samples, averages cluster between 1.3 – 4.5—far below a passing level even for the strongest models. The scores show stable performance ranges; full records include raw outputs, itemized scoring, and expert review.

EVALUATION SAMPLE · LONG-CONTEXT DEFINITION RETENTION

Sample 19 | HEP Theory / AdS | Normalization-sensitive

AdS Light-Front Coordinates and Residual Coefficients

This example asks the model to read the local context first, then set **d=6** and **mR_{AdS}=3**. The hard part is maintaining definitional consistency across **ODEs, residual coefficients, affine cutoffs, and exponents**—each quantity is defined by the current context.

Question setup · read the context; set d=6 and mR_{AdS}=3

$$r_\star = \frac{3}{5}R_{\mathrm{AdS}}, \quad t_\star = -\frac{9}{20}R_{\mathrm{AdS}}, \quad \xi_\star = \frac{1}{2}(r_\star - t_\star)$$

✦ KEY STRUCTURE

Normalized solution Let $a(r)$ be the **normalized** solution of the exact first-order ODE—the normalization condition appears only in the context.

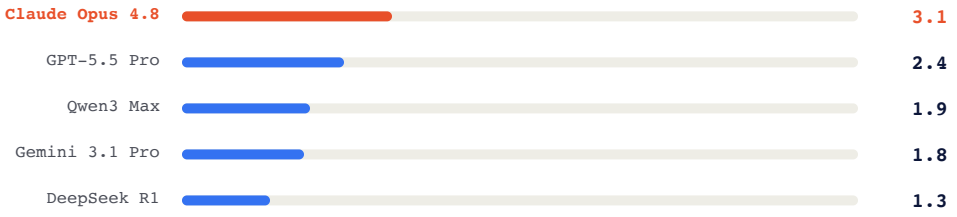
Stripped solutions $b(r)$ and $c(x^-)$ are the stripped solutions specified by the context; $d(r) = F[a](r)$ is defined by operator F.

Target quantity W is assembled from the above quantities at $\xi_\star$; the exponents and affine cutoff all depend on local conventions.

⚠ HOW MODELS FAIL

- Most models apply **generic AdS textbook formulas**, ignoring the normalization rewritten in the context.
- Between the residual coefficient and exponent terms, they **lose definitional consistency** and switch conventions mid-solution.
- **Claude Opus 4.8** is the steadiest, but the final affine-cutoff step still needs human review.

✓ AVG@3 · SINGLE-ITEM STABILITY · 10-POINT SCALE · 3 INDEPENDENT RUNS PER ITEM



Almost all points are lost at the step of **obeying the local definitions in context**—the problem is not that the model cannot compute AdS, but that it did not hold onto this AdS context.

EVALUATION SAMPLE · TRANSFORMATION FIDELITY AND FACTOR SENSITIVITY

Sample 18 | Mathematical Physics / Transform | Factor-sensitive

Inverse Hadamard Inversion over Even/Odd Sectors

> The model can recognize the Hadamard structure, yet **often drops the 1/4 factor or reverses sector signs**—a fidelity failure, not a lack of transform knowledge.

Definition · ordered quadruple as raw samples

$$X(\sigma_1, \sigma_2) = X_{ee} + \sigma_2 X_{eo} + \sigma_1 X_{oe} + \sigma_1 \sigma_2 X_{oo}$$

✦ KEY STRUCTURE

Even/odd decomposition The four coefficients $X_{ee}, X_{eo}, X_{oe}, X_{oo}$ are combined according to the parity of σ_1, σ_2.

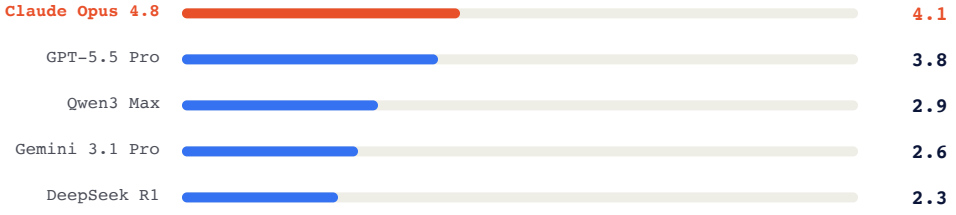
Inversion Use **inverse Hadamard transform** to recover the four coefficients from the ordered quadruple.

Factor constraint Inversion **must include the factor** $1/4$, and every sector sign must match the convention in the context.

⚠ HOW MODELS FAIL

- The model recognizes the Hadamard structure, but **misses the normalization factor** $1/4$.
- Across even/odd sectors, it **reverses signs**, confusing X_{eo} with X_{oe}.
- **Claude Opus 4.8** is most complete on factors and signs, but a few sectors still need review.

✓ AVG@3 · SINGLE-ITEM STABILITY · 10-POINT SCALE · 3 INDEPENDENT RUNS PER ITEM



Local calculations are often correct; the losses fall on **transformation fidelity · sign control · normalization awareness** as three context-fidelity points.

SCHEMA · FIELDS

What one physics In-Context Learning data item looks like.

It records more than whether the model answered correctly; it captures how the model reads definitions, selects physical objects, and assembles intermediate quantities from the given material—ready to plug into training and evaluation pipelines.

context · prompt	Paper-style scientific reading material plus the question, with local definitions, boundary conditions, and notational conventions. core
golden_response	Expert-written reference solution and key intermediate conclusions (Golden Response), used as the highest-quality anchor. core
rubrics · checkpoints	Weighted scoring rules: definition tracking, physical object selection, transformation fidelity, and assembly criteria, itemized for machine checking. core
domain	8 first-level physics domains plus subdomain tags, supporting precise indexing by capability dimension.
source_fields	Source material, local notation system, and normalization / sector / branch conventions.
model_runs	avg@3 scores and raw outputs from three independent runs per item across frontier models.
first_error_review	Expert first-error review: marks the earliest failure point and In-Context Learning failure type (definition / object / transform / assembly). core

DELIVERY · PUT PHYSICS REASONING INTO YOUR MODELS

Turn expert reasoning in scientific materials into the model's In-Context Learning capability.

100	8	5×3	avg@3
long-context physics samples	first-level domains	strong models × independent runs	+ first-error review

Context-dependent

paper-style materials + local definitions / notation annotations

Process-level scoring

definition tracking · object selection · transformation fidelity · assembly criteria

Multi-model avg@3

focuses on why points are lost, not just totals

Expert review

physics researcher calibration + stepwise first-error marking

Truly valuable scientific data should be trainable, evaluable, reviewable, and traceable to errors. Tell us your training or evaluation direction, and we will quickly assemble a matched physics In-Context Learning dataset and evaluation pipeline.

Request full samples · book a walkthrough →

contact@sotalab.ai · WeChat / Feishu contact

Context over Formula

From “accumulating data” toward “learning how to learn”.

BY ZHIHU