



CONTEXT OVER FORMULA

物理推理

Physics In-Context Learning Data

不是让模型背更多公式，
而是让它读懂当前上下文：
为长文本推理，构造可训练、可评测的 ICL 数据。

立场 · READ THIS PHYSICS, NOT THE TEXTBOOK

结构化物理 ICL 数据， 训练模型真正利用上下文。

上下文不只是把长文本塞进去那么简单。它关乎模型如何理解、记忆并利用信息，这才是拉开应用差距的关键。

从「记住」到「理解」：长上下文不是越长越好。关键信息如果被埋在中间，模型很容易忽略，也就是 Lost in the Middle；竞争点在于能否精准定位和理解关键信息。

从「对话」到「工作流」：长任务需要任务状态管理，而不是无限拉长上下文；模型需要持续知道当前目标、已完成动作和待办事项。

从「塞进去」到「长进心里」：部分技能也应被内化为稳定的条件反射式能力。

物理题天然适合训练这类能力：答案明确，但必须服从当前材料中的定义、边界、规范、sector 与符号约定。

答案对不对是结果，
读没读懂这段 context 才是能力。

CONTEXT FAITHFUL · OBJECT SELECTION · VERIFIABLE

100 道题的一级覆盖， 既有物理广度，也有推理深度。

覆盖结构控制单一理论方向占比，补足量子、凝聚态、数值、实验与数学物理，让数据能测到真实科研阅读中的多种失败机制。

● 凝聚态与统计物理	CONDENSED MATTER	16
相变 · 强关联 · 无序系统 · 量子气体 · 有效模型与统计极限		
● 高能理论与弦论	HEP-THEORY	14
AdS/CFT · CFT · brane · anomaly · sector decomposition · 符号推导		
● 引力·黑洞·宇宙学	GRAVITY	14
度规 · 黑洞热力学 · 宇宙学扰动 · 半经典引力 · 边界与规范约定		
● 量子物理与量子信息	QUANTUM & QI	14
量子态 · 测量基 · 纠缠 · 噪声与开放系统 · Hilbert space 识别		
● 计算与数值物理	COMPUTATIONAL	14
稳定性 · 收敛阶 · 误差指标 · neural operator · 低标量误差 vs 真实正确性		
● 应用与实验物理	APPLIED PHYSICS	12
实验参数读取 · 误差传播 · 仪器设定 · 数据归约与条件恢复		
● 数学物理与形式结构	MATH PHYSICS	8
谱理论 · 拓扑量 · index · 可积结构 · 长证明链与形式约束		
● 粒子现象学与核物理	PHENOMENOLOGY	8
QCD · flavor · collider observable · neutrino · 核结构与现象学约束		

覆盖不是堆方向——每个方向都对应一类真实的 **ICL 失败机制**：定义漂移、物理对象选错、多 sector 漏项、branch 走错。

合计 100 道物理长文本 ICL 样本 · 每条均带论文式 context、标准解、评分点与多模型运行记录。

对齐主流评测方向： 科学问答 · 科研阅读 · 长上下文 · 步骤复核。

这套数据面向现有 benchmark 暴露的共同缺口——模型在**短题式科学问答**进步很快，但在**论文式 context**、**物理对象选择**、**跨段落证据装配**与**稳定推理**上仍不可靠。

CL-bench CONTEXT LEARNING	最直接对齐：评估模型能否从给定上下文中学习并应用新知识，而不只是检索长文本或套用预训练记忆。	上下文学习
HLE EXPERT-LEVEL	补充物理专家级 context-dependent 样本，提升科学条件读取、局部定义保持与答案校准。	专家级
GPQA Diamond GRADUATE QA	把短题科学推理延展到论文式 context，降低模型靠记忆模板拿高分的可能。	研究生级
PaperBench REPLICATIONBENCH	提供可训练的科研阅读子任务：先让模型在物理材料中完成确定性推理。	科研复现
ProcessBench PRM800K	把错误定位从数学题扩展到物理长文本、边界条件、sector 与 branch 选择。	首误定位
MMMU · MathVista MULTIMODAL	可扩展到论文插图、实验表格、公式混合输入，训练科学多模态 ICL。	多模态
LongBench · L-Eval LONG-CONTEXT	从「找到信息」升级为「基于信息完成物理推理并输出可验证答案」。	长上下文



avg@3 测评：强模型在物理 ICL 上仍明显失分。

同一 context、同一 prompt、同一专家 Rubric。每个模型独立运行三次、按 0-10 分取平均，并由专家复核关键失分位置。

CASE · 主要 ICL 失分点	GPT-5.5 PRO	GEMINI 3.1 PRO	CLAUDE OPUS 4.8	DEEPSEEK R1	QWEN3 MAX
19 AdS 光前坐标 Definition Tracking	2.4	1.8	3.1	1.3	1.9
27 强场能量核 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 带电球面对偶场 Boundary Context	2.9	2.1	3.6	1.7	2.2
21 拓扑 selector Branch Selection	3.2	2.4	3.8	1.9	2.7
13 CFT 双迹递推 Recursive Consistency	4.0	2.8	4.5	2.5	3.1

六道样例满分 10，均分集中在 1.3 – 4.5——即便最强模型也远未及格。分数用于展示稳定表现区间；完整记录含每次运行原文、逐项评分与专家复核。

评测样本 · 长文本定义保持

Sample 19

高能理论 / AdS

归一化敏感

AdS 光前坐标与残余系数

卡住模型的不是式子长，而是要在 **ODE**、**residual coefficient**、**affine cutoff** 与**指数幂**之间保持定义一致——每个量都由当前 context 定义。

SETUP · 取 $D=6$, $MR_{\text{AdS}}=3$

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

◆ 关键结构 STRUCTURE

归一化解令 $a(r)$ 为精确一阶 ODE 的**归一化解**——归一化条件只写在 context 里。

剥离解 $b(r)$ 、 $c(x^-)$ 是 context 指定的剥离解， $d(r) = F[a](r)$ 由算子 F 定义。

目标量 W 由上述量在 $\xi_\star$ 处装配，指数幂与 affine cutoff 全部依赖局部约定。

△ 模型如何失手

- 多数模型套用**通用 AdS 教材公式**，忽略 context 里改写过的 normalization。
- 在 residual coefficient 与指数幂之间**丢失定义一致性**，中途换了约定。
- Claude Opus 4.8** 相对最稳，仍在 affine cutoff 末步需人工复核。

✓ AVG@3 · 单题稳定表现 · 满分 10 · 每题 3 次独立运行

Claude Opus 4.8		3.1
GPT-5.5 Pro		2.4
Qwen3 Max		1.9
Gemini 3.1 Pro		1.8
DeepSeek R1		1.3

失分几乎全部落在「**服从 context 局部定义**」这一步——不是不会算 AdS，而是没读住这段 AdS。

评测样本 · 变换保真与因子敏感

Sample 18

数学物理 / 变换

因子敏感

奇偶 sector 的 inverse Hadamard 反演

模型能识别 Hadamard 结构，却常漏掉 $1/4$ 因子或写反 sector 符号——保真度失败，而非不懂变换。

DEFINITION · 有序四元组解释为原始样本

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

◆ 关键结构 STRUCTURE

奇偶分解四个系数 $X_{ee}, X_{eo}, X_{oe}, X_{oo}$ 按 σ_1, σ_2 的奇偶性组合。

反演用 **inverse Hadamard transform** 从有序四元组恢复四个系数。

因子约束反演**必须包含因子** $1/4$ ，且各 sector 符号需与 context 约定一致。

△ 模型如何失手

- 识别出 Hadamard 结构，却**漏掉归一化因子** $1/4$ 。
- 在奇偶 sector 上**写反符号**，把 X_{eo} 与 X_{oe} 弄混。
- **Claude Opus 4.8** 因子与符号最完整，仍有个别 sector 需复核。

✓ AVG@3 · 单题稳定表现 · 满分 10 · 每题 3 次独立运行

Claude Opus 4.8		4.1
GPT-5.5 Pro		3.8
Qwen3 Max		2.9
Gemini 3.1 Pro		2.6
DeepSeek R1		2.3

局部计算都能做对，失分落在 **transformation fidelity · sign control · normalization awareness** 三个上下文保真点上。

一条物理 ICL 数据，长什么样。

不只记录模型答对没有，更刻画它如何在给定材料中读取定义、选择物理对象、装配中间量——开箱即用，可直接接入训练 / 评测管线。

context · prompt	论文式科学阅读材料 + 题目，携带局部定义、边界条件与符号约定。 核心
golden_response	专家撰写的标准解与关键中间结论（Golden Response），作为最高质量锚点。 核心
rubrics · checkpoints	加权评分细则：定义追踪、物理对象选择、变换保真、装配判据，逐条可机检。 核心
domain	8 一级物理方向 + 子方向标签，支持按能力维度精准索引。
source_fields	材料出处、局部符号系统与 normalization / sector / branch 约定。
model_runs	多款前沿模型每题 3 次独立运行的 avg@3 跑分与原文输出。
first_error_review	专家首误复核：标定最早失效位置与 ICL 失败类型（定义 / 对象 / 变换 / 装配）。 核心

交付 · PUT PHYSICS REASONING INTO YOUR MODELS

把科学材料里的专家推理， 沉淀为模型的ICL 能力。

100

物理长文本样本

8

一级方向覆盖

5×3

强模型 × 独立运行

avg@3

+ 首误复核

context 依赖

论文式材料 + 局部定义 / 符号约定标注

过程级评分

定义追踪 · 对象选择 · 变换保真 · 装配判据

多模型 avg@3

关心「为什么失分」而非总分

专家复核

物理研究者定标 + 首误逐步标定

真正有价值的科学数据，应当可训练、可评测、可复核、可追错。告知我们您的训练 / 评测方向，我们将快速组建对口的物理 ICL 数据与评测管线。

索取完整样本 · 预约 walkthrough →

contact@sotalab.ai · 微信 / 飞书联系对接人

Context over Formula

让模型从「积累数据」，
走向「学会学习」。



BY ZHIHU