ICLR 2026 | Oral

Microsoft
Research Asia

LoongRL: Reinforcement Learning for Advanced Reasoning over Long Contexts

Siyuan Wang*[1,2] Gaokai Zhang*[1,3] Li Lyna Zhang[1] Ning Shang[1] Fan Yang[1] Dongyao Chen[2] Mao Yang[1]

[1]Microsoft Research Asia [2]Shanghai Jiao Tong University [3]Carnegie Mellon University

* Equal contribution

arXiv: 2510.19363

7B/14B models rival o3-mini & DeepSeek-R1 on long-context reasoning

The Gap: Long-Context Retrieval vs. Reasoning

- Modern LLMs support 128K+ context windows and excel at **retrieval tasks**
- But they struggle with **multi-hop reasoning** over long documents
- Existing RL methods (DeepSeek-R1, DAPO) focus on math/code with short contexts
- Key challenge: **How to enable small models to reason over long contexts via RL?**

Two Key Challenges

1. Hard to find suitable data for long-context reasoning training

2. RL on long contexts is prohibitively expensive

LoongRL: Two Key Innovations

Data Synthesis: KeyChain

Transform **short** multi-hop QA into **high-difficulty long-context** tasks with UUID key-value chains

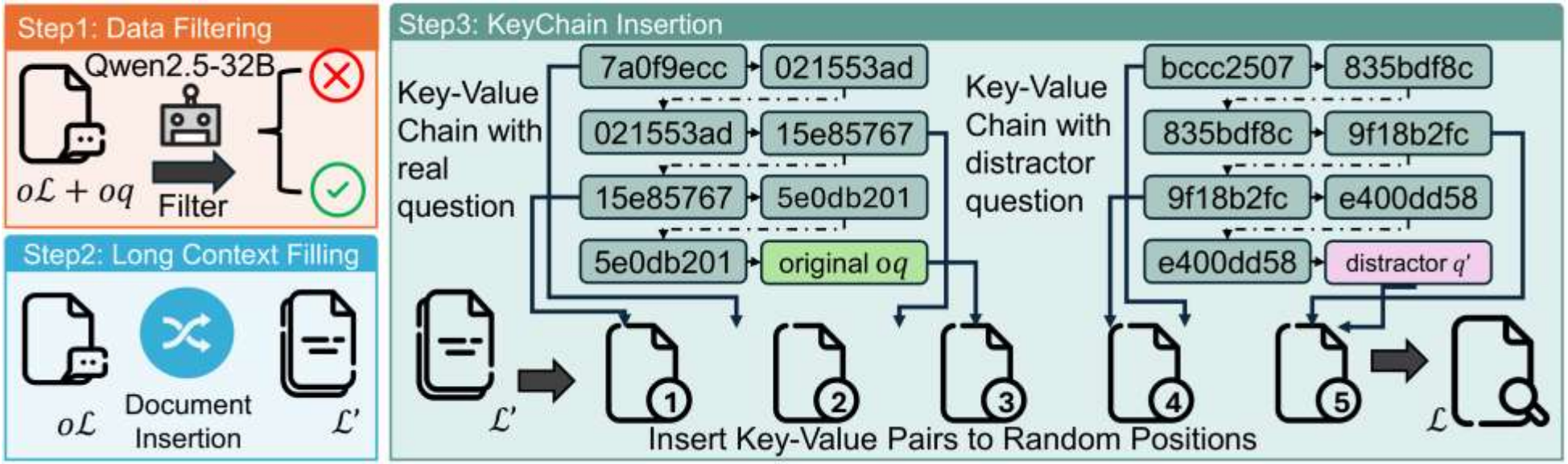
Training Recipe: GRPO

Multi-stage curriculum training, introducing warmup and hard-mining stages.
Train at **16K**, elicit model's retrieval and reasoning capabilities, which generalizes to **128K**

Result: 7B model achieves **72.4** avg on LongBench v1 — rivaling **o3-mini (74.5)** and **DeepSeek-R1 (74.9)**

KeyChain: Data Synthesis Pipeline

Transforming short QA into high-difficulty long-context reasoning tasks



Step 1: Data Filtering

Use Qwen2.5-32B to remove overly easy/hard instances from HotpotQA, MuSiQue, 2WikiMQA

277K → 72K

Step 2: Long Context Filling

Insert distractor documents to extend context to 16K

Shuffle & pad

Step 3: KeyChain Insertion

Embed UUID key-value chains for real + distractor question at random positions

72K → 7.5K

KeyChain: Task Example

The model must trace UUID chains to find the hidden question, then reason to answer

Please read the following text.
Document 0:
...
Document 3:
Who's Who? is a studio album by American jazz musician John Scofield. It features two different bands, one acoustic and one electric. The acoustic group, featuring Scofield's then-employer Dave Liebman on saxophones, Eddie G\u00f3mez on bass, and Billy Hart on drums, recorded "The Beatles" and "How the West Was Won". ...
{"bdd640fb-0667-4ad1-9c80-317fa3b1799d": "23b8c1e9-3924-46de-beb1-3b9046685257"}.
...
Document 10:
...
The university is one of the smallest of the 23 CSU campuses in California. Sonoma State offers 92 Bachelor's degrees, 19 Master's degrees, one Doctoral degree (Doctor of Education), and 11 teaching credentials. {"972a8469-1641-4f82-8b9d-2434e465e150": "Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?"}.
...
Document 47:
Neil Affleck
{ "23b8c1e9-3924-46de-beb1-3b9046685257": "972a8469-1641-4f82-8b9d-2434e465e150"}.
Neil Affleck (born 1953) is a Canadian animator, director, and former actor. He has worked as an animator on "The Simpsons" and "Family Guy", and as an actor appeared in a leading role in the 1981 film "My Bloody Valentine". {"9a1de644-815e-46d1-bb8f-aa1837f8a88b": "b74d0fb1-32e7-4629-8fad-c1a606cb0fb3"}.
...
In the context above, there is one correct question to answer. The correct question can only be found by following the correct consecutive chain of key:value pairs encoded with UUID strings (e.g., f81d4fae-7dec-11d0-a765-00a0c91e6bf6), starting from "bdd640fb-0667-4ad1-9c80-317fa3b1799d".
Find the correct question first, then answer it.

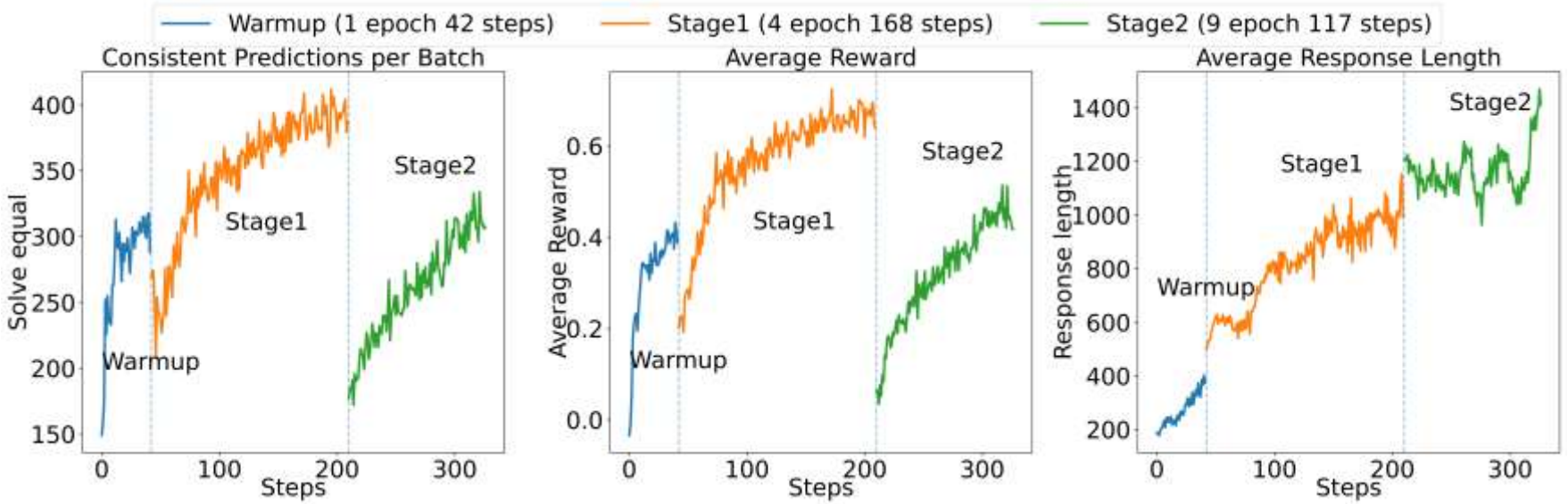
- Model must: **(1)** follow the correct UUID chain to find hidden question **(2)** retrieve relevant facts among distractors before **(3)** reasoning over the supporting facts to answer

Multi-Stage RL Training Recipe

Warmup
Non-KeyChain

Stage I
Normal Training

Stage II
Hard Mining



Key Hyperparameters

Algorithm: **GRPO (G=8)** | LR: **1e-6** | Context: **16K tokens** | Max Output: **4096 tokens** | KL: **0.001** | HW: **16xA100 / 8xMI300X**

Emergent Reasoning Patterns

RL training induces a human-like plan-retrieve-reason-recheck loop

(a) RL-Generated Trajectory with KeyChain Data

{Plan: decompose problems to multi-steps}

To determine when the last time Peter Till's sports team beat the winner of the 1894-95 FA Cup, we need to follow these steps:

1. Identify the winner of the 1894-95 FA Cup.
2. Determine the teams Peter Till played for.
3. Find the last time any of Peter Till's teams beat Aston Villa.

{Retrieve relevant document}

From Passage 6: The 1894-95 FA Cup was won by Aston Villa.

From Passage 5: Peter Till played for several teams, but we need to find the teams he played for profession

{Step-level reasoning}

Now, we need to find the last time any of these teams beat Aston Villa from the information in Passage 5:

{Retrieve relevant document}

{Retrieve relevant document}

From Passage 5: The most recent relevant encounter between Aston Villa and Birmingham City (one of the teams Peter Till played for, Walsall) is mentioned …

\boxed{1 December 2010}

Correct

(b) RL-Generated Trajectory with KeyChain Data

{Mixed reasoning with retrieval}

First, we need to identify who won the 1894-95 FA Cup. From the passage about the 1894-95 FA Cup, we can see that Aston Villa won this competition by defeating West Bromwich Albion 1-0 in the final. So, the winner of the 1894-95 FA Cup was Aston Villa.

Next, we need to find when Peter Till's team, which is Everton, last beat Aston Villa. Looking through the passage about Everton's history and rivalries, we find that the most recent victory of Everton over Aston Villa in the Premier League was mentioned: "Villa finally beat Blues in the Premiership, thanks to a Kevin Phillips goal.

However, we are asked about Everton specifically. Continuing to search the text, we find: "Everton's biggest rivalry is with neighbours Liverpool, against whom the club contests the Merseyside derby.

Therefore, the last time Peter Till's team (Everton) beat the winner of the 1894-95 FA Cup (Aston Villa) was in the 2005-2006 season.</think> \boxed{2005-2006 season}

Wrong

Plan
Decompose problems into specific sub-steps

Retrieve
Find supporting facts from the passages

Reason
Step-level logical inference

Recheck
Verify retrieval, correct if needed

Main Results: Long-Context Reasoning

Model	Size	LC Avg	HotpotQA	2WikiMQA	MuSiQue	NarrQA	QASPER
o3-mini (medium)	--	74.5	83.0	89.0	64.0	60.7	60.5
DeepSeek-R1	671B	74.9	82.7	91.3	72.2	66.9	61.4
GPT-4o	--	64.7	82.5	78.0	54.0	60.5	48.5
QwQ-32B	32B	69.6	78.5	87.4	62.7	61.1	58.5
R1-Distill-LLaMA-70B	70B	65.4	76.1	85.0	61.9	53.4	50.5
Qwen2.5-7B-Instruct	7B	48.9	69.5	50.5	34.0	44.5	46.0
R1-Distill-Qwen-7B	7B	31.2	40.2	53.3	11.1	8.9	42.5
LoongRL-7B	**7B**	**72.4**	**83.1**	**91.1**	**65.6**	**58.4**	**63.6**
Qwen2.5-14B-Instruct	14B	53.1	74.0	60.5	36.5	48.5	46.0
R1-Distill-Qwen-14B	14B	64.9	77.5	87.0	58.0	51.0	51.0
QwenLong-L1-32B	32B	70.1	80.7	89.1	65.2	58.6	56.7
LoongRL-14B	**14B**	**74.2**	**82.2**	**93.3**	**67.5**	**63.4**	**64.5**

LoongRL-7B: +23.5% over baseline (48.9 → 72.4)

LoongRL-14B: +21.1% over baseline (53.1 → 74.2)

Rivals o3-mini & DeepSeek-R1 at 7B/14B scale

Preserving General Short-Context Abilities

Long-context RL training does NOT degrade general capabilities

Model	Gen. Avg	MMLU	MATH-500	IFEval
Qwen2.5-7B-Instruct	73.5	73.4	76.0	71.2
R1-Distill-Qwen-7B	69.9	62.3	92.8	54.7
LoongRL-7B	**75.0**	**76.2**	**78.0**	**70.9**
Qwen2.5-14B-Instruct	81.3	79.4	83.4	81.0
R1-Distill-Qwen-14B	81.0	76.6	93.9	72.6
QwenLong-L1-32B	84.1	78.5	95.2	78.6
LoongRL-14B	**80.7**	**80.5**	**83.2**	**78.4**

- **LoongRL-7B:** +2.8 MMLU over base | preserves IFEval (-0.3%)
- **LoongRL-14B:** +1.1 MMLU over base | R1-Distill drops -8.4% on IFEval
- Proper data mixing preserves general reasoning while improving long-context

Ablation: Each Component Matters

Effect of KeyChain Data

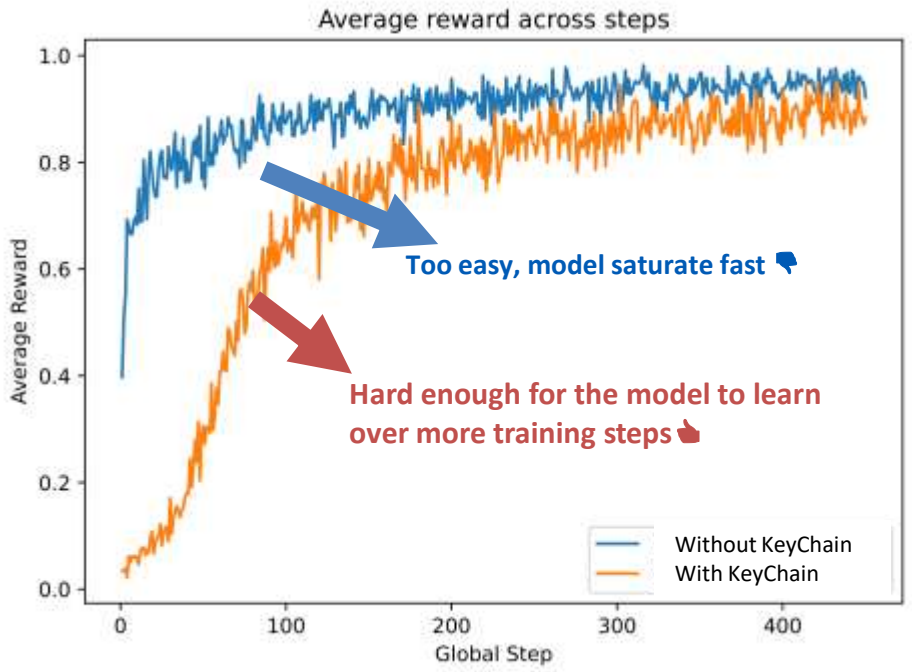
Condition	HotpotQA	2WikiMQA	MuSiQue	NarrQA	QASPER	Avg
Qwen2.5-7B (baseline)	69.5	50.5	34.0	44.5	46.0	48.9
+ RL (w/o KeyChain)	80.3	84.7	58.5	53.0	54.5	66.2
+ RL (w/ KeyChain)	**83.1**	**91.1**	**65.6**	**58.4**	**63.6**	**72.4**

KeyChain adds +6.2% on top of standard RL

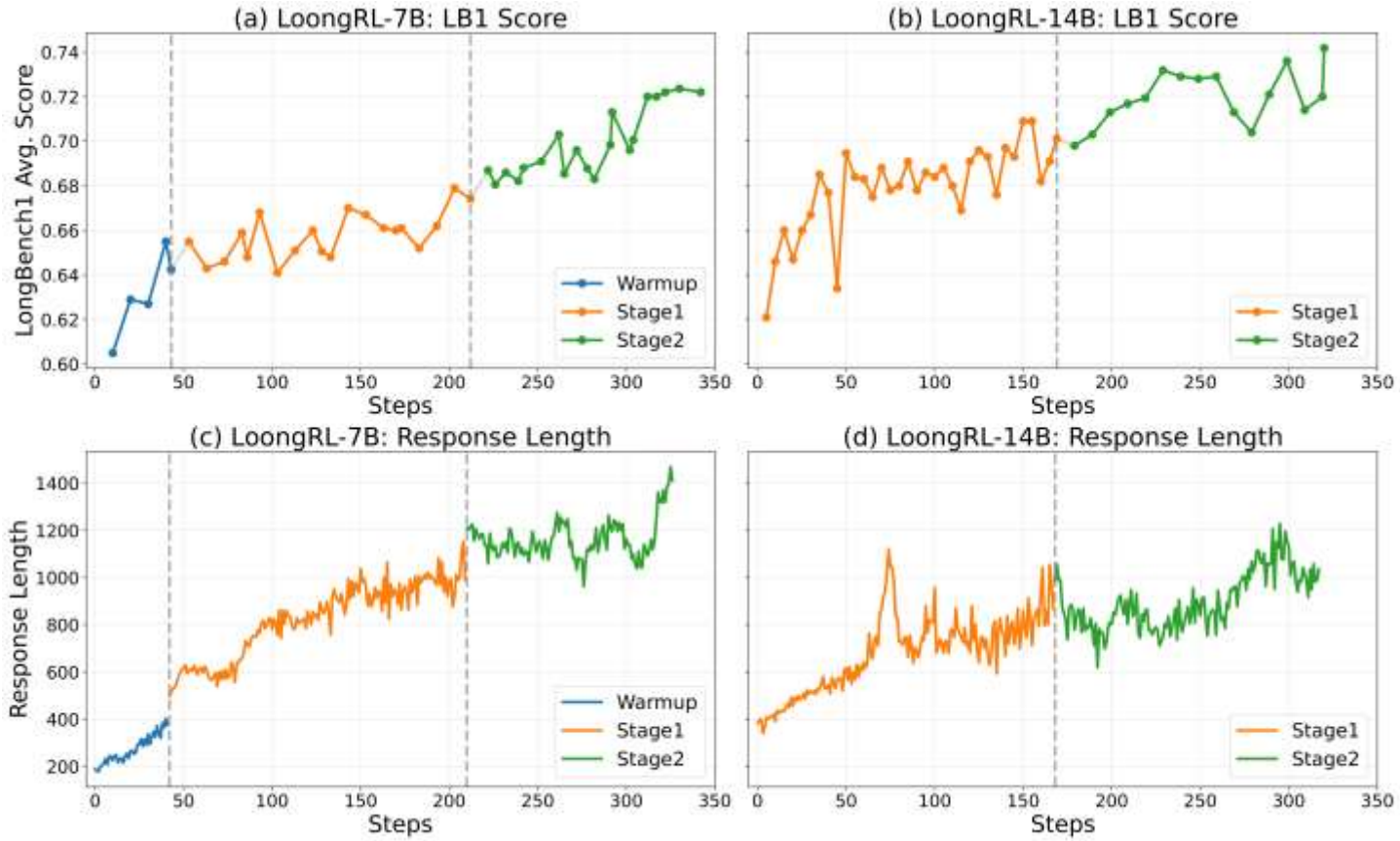
Training Data Composition (~17K examples)

Dataset	Size	Purpose
KeyChain (HotpotQA + MuSiQue + 2Wiki)	**7,500**	**Hard long-context reasoning**
Standard QA (3 datasets)	7,500	Retrieval baseline
Book RULER (multi-key / multi-value)	1,024	Long-context retrieval
Math (DAPO + Multiple-choice)	5,000	General reasoning

Hard-enough data to keep the RL going!



Training Dynamics: 7B and 14B

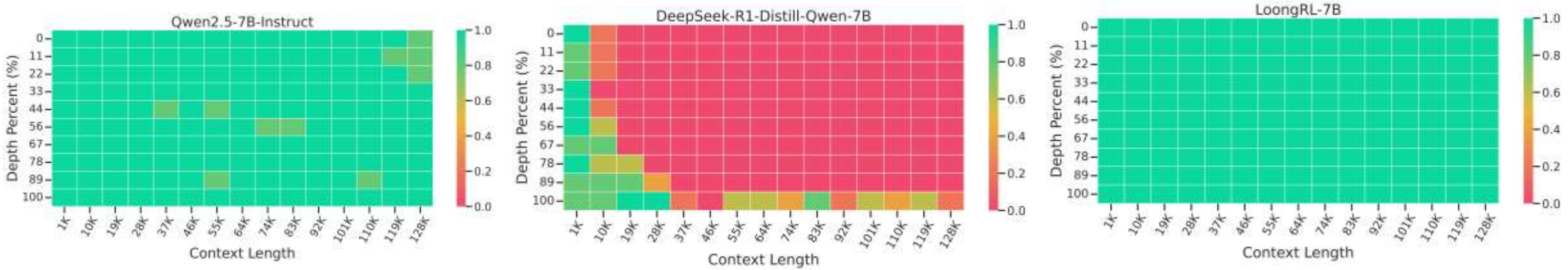


- Both models show consistent LongBench improvement across all training stages
- Response length grows naturally **(200 → 1400 tokens)** — model learns to think more deeply

Length Generalization: Train at 16K, Generalize to 128K

NarrativeQA length-split and RULER benchmark performance

Model	NarrQA 0-16K	NarrQA 16-32K	NarrQA 32-64K	RULER 16K	RULER 32K	RULER 64K	RULER 128K
Qwen2.5-7B-Instruct	55.7	35.2	42.4	92.3	89.5	81.8	69.4
QwenLong-L1-32B	65.9	48.1	60.0	87.6	86.8	80.6	70.2
LoongRL-7B	**69.8**	**47.4**	**57.2**	**93.4**	**91.4**	**86.2**	**76.8**
LoongRL-14B	**69.5**	**55.2**	**64.3**	**95.4**	**95.1**	**87.1**	**79.9**



Thank You!

Questions & Discussion

Paper
https://arxiv.org/abs/2510.19363

Webpage
https://loongrl.github.io/

wsy0227@sjtu.edu.cn

Open to work!

☜Siyuan Wang

Gaokai Zhang☞

gaokaiz@andrew.cmu.edu