EntropyLong

Effective Long-Context Training via Predictive Uncertainty

Junlong Jia[1·5] Ziyang Chen[2] Xing Wu[2*] Chaochen Gao[2] Zijia Lin[3] Songlin Hu[2] Binghui Guo[1·4·5*]

[1]Beihang University [2]Institute of Information Engineering, CAS [3]Tsinghua University [4]Beijing Advanced Innovation Center [5]LMIB, NLSDE

ICLR 2026

Outline

1

Background & Motivation

Long-context pretraining needs genuine dependencies—not just long sequences

2

EntropyLong Framework

Model-in-the-loop selection and entropy-based verification of retrieved context

3

Experiments & Results

RULER, LongBench-v2 after SFT, and key ablations

4

Analysis & Conclusion

Empirical verification of dependencies and takeaways

Background: The Long-Context Data Problem

Context windows are getting longer...

- Architectures now support 128K+ tokens
- Yet models still under-use distant content, (e.g., lost-in-the-middle phenomenon)
- **Root cause: training data lacks genuine long-range dependencies**

Current approaches:

- Naive concatenation → long sequence, but no cross-span connections
- Heuristic retrieval → plausible text, but unverified for the target LM
- Task-specific synthesis → useful, but limited coverage

Key Insight: Entropy as Information Deficit Signal

A model's predictive uncertainty (entropy) directly signals where it needs more context.

High entropy → information deficit → model needs help

Entropy reduction after adding context → verified dependency

Predictive Entropy:

$$-\sum_{v\in\mathcal{V}} P_{\boldsymbol{\theta}}(v|x_{<t}) \log P_{\boldsymbol{\theta}}(v|x_{<t})$$

Contextual Information Gain:

$$\frac{H_{\boldsymbol{\theta}}(x_{t_i}|x^{D}_{<t_i}) - H'_{\boldsymbol{\theta}}(x_{t_i}|x^{[C_j;D]}_{<t_i+|C_j|})}{H_{\boldsymbol{\theta}}(x_{t_i}|x^{D}_{<t_i})}$$

→ Principled, model-in-the-loop verification for long-range dependencies

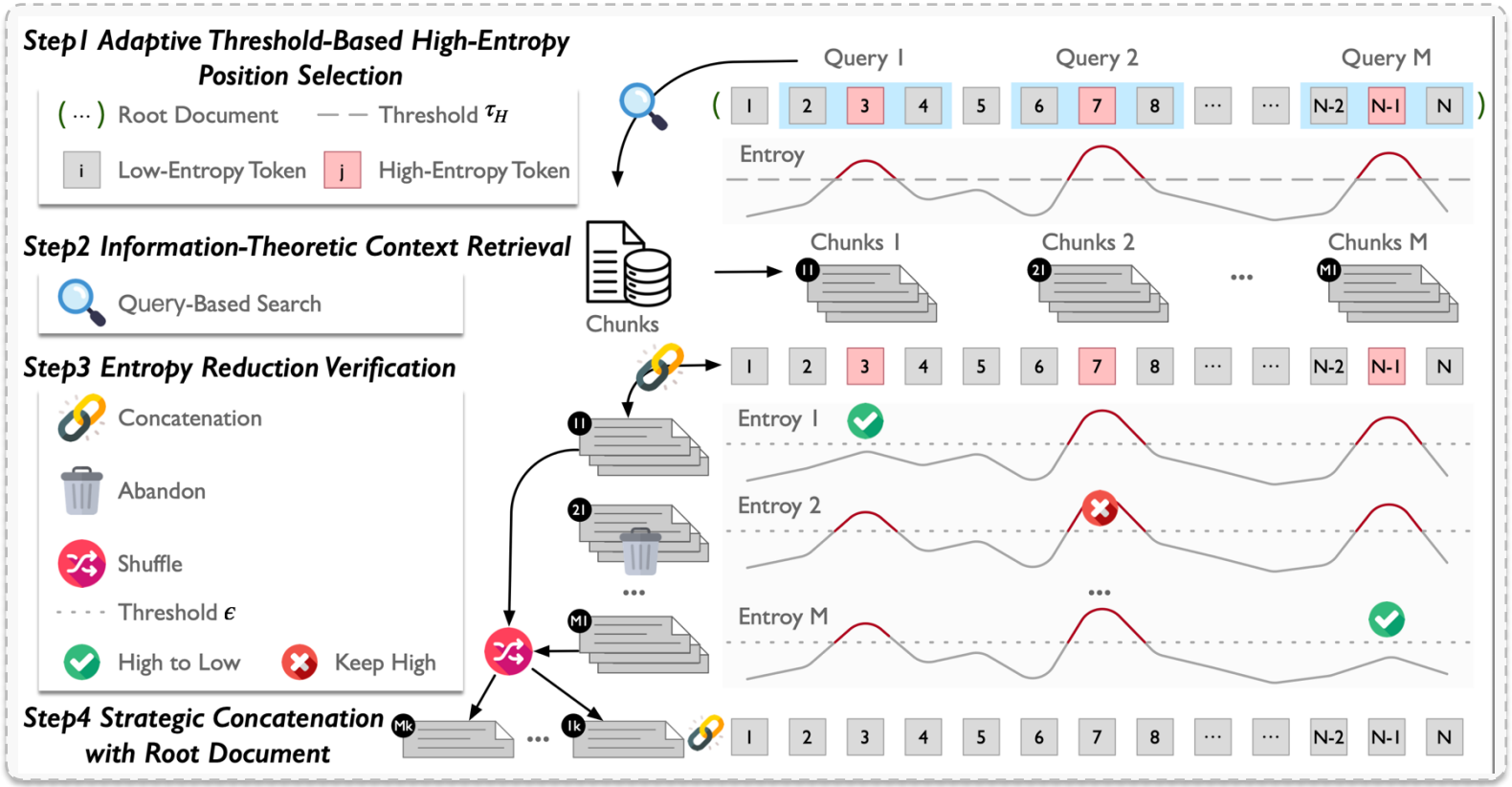
EntropyLong: 4-Stage Pipeline

Step 1: High-Entropy Position Selection
Identify uncertain positions using adaptive threshold: $\tau_H = \mu_H + \alpha \cdot \sigma_H \quad (\alpha = 2.0)$
These mark where context is most needed.

Step 2: Information-Theoretic Retrieval
Extract surrounding context (w=16 words) as query. Retrieve top-K candidates from corpus via dense retrieval (Jina sentence transformer).

Step 3: Entropy Reduction Verification
Prepend retrieved context, re-compute entropy. Retain only if relative reduction > ε (0.4):

Step 4: Strategic Concatenation
Shuffle verified contexts randomly, prepend to root document for 128K training sequences. Random order prevents positional bias.



Experimental Setup

Base Model

- Meta-Llama-3-8B, extended to 128K
- RoPE base=200M, 1000 iters, 4M batch

Dataset

- Source: FineWeb-Edu + Cosmopedia
- Retrieval corpus: 1B+ documents
- Output: 4B tokens of 128K sequences

Baselines

- Quest (coherence-driven)
- NExtLong (discrimination-driven)

Evaluation

RULER Benchmark

- Needle, multi-hop, variable tracking,

  pattern extraction at 8K–128K

LongBench-v2

- Real-world long-context tasks
- After SFT with UltraChat

Results

Table 1: Main results on the RULER benchmark. Best results are in **bold**.

RULER	8k	16k	32k	64k	128k	avg
Quest	91.39	89.72	84.37	77.07	60.11	80.53
NExtLong	89.99	88.58	86.04	83.52	77.99	85.22
EntropyLong	**91.50**	**90.11**	**88.95**	**85.04**	**81.26**	**87.37**

- **Advantage GROWS** with context length: **+21.15** over Quest, **+3.27** over NExtLong at 128K

Table 2: Results on LongBench-v2 after instruction fine-tuning (UltraChat (Ding et al., 2023)).

Model	Easy	Hard	Short	Medium	Long	Overall
Quest	17.70	25.10	25.60	20.00	21.30	22.30
NExtLong	21.40	25.70	27.20	21.90	23.10	24.10
EntropyLong	**25.50**	**28.90**	**30.00**	**23.70**	**31.50**	**27.60**

- **+8.4** on Long tasks over NExtLong
- Verified dependencies translate to real-world downstream gains

Table 8: Performance on short text benchmarks. EntropyLong maintains competitive performance on short texts while excelling at long contexts.

	arc_c	arc_e	hellaswag	piqa	logiqa	winogrande	avg
Llama3-8b-base	50.34	80.18	**60.13**	79.60	**27.50**	**72.85**	61.77
+EntropyLong	**51.45**	**80.98**	59.61	**80.47**	26.27	72.22	**61.83**

- **EntropyLong** maintains competitive short tasks

Analysis

Table 3: The verification step is critical for performance. We report RULER scores across different context lengths and the average.

RULER	8k	16k	32k	64k	128k	avg
EntropyLong-NoVerify	91.44	88.76	86.29	83.85	79.47	85.82
EntropyLong *(Full Method)*	**91.50**	**90.11**	**88.95**	**85.04**	**81.26**	**87.37**

- Removing entropy verification causes a consistent drop across all context

- α = 2.0 selects the right number of positions — not too many (noisy), not too few (insufficient signal).
- ε = 0.4 keeps enough verified dependencies while filtering out weak ones.

Table 4: Performance on RULER benchmark with varying adaptive threshold α values. #Tokens indicates the number of high-entropy tokens selected for context retrieval. Higher α values result in fewer but more selective high-entropy positions.

RULER	#Tokens	8k	16k	32k	64k	128k	avg
EntropyLong ($\alpha = 1.5$)	913	90.25	87.16	82.07	79.49	73.47	82.49
EntropyLong ($\alpha = 2.0$)	292	**91.50**	**90.11**	**88.95**	**85.04**	**81.26**	**87.37**
EntropyLong ($\alpha = 2.5$)	83	90.96	88.29	85.52	82.83	80.02	85.52

Table 5: Performance on RULER benchmark with varying entropy reduction thresholds. #Verified indicates the number of verified dependencies retained after entropy reduction validation. The threshold of 0.4 provides the best empirical performance.

RULER	#Verified	8k	16k	32k	64k	128k	avg
EntropyLong ($\epsilon = 0.2$)	62	91.44	88.44	84.97	83.53	78.84	85.45
EntropyLong ($\epsilon = 0.4$)	46	**91.50**	90.11	**88.95**	**85.04**	81.26	**87.37**
EntropyLong ($\epsilon = 0.6$)	29	90.31	88.64	86.19	84.09	**81.46**	86.14
EntropyLong ($\epsilon = 0.8$)	13	91.39	**90.29**	87.60	83.90	79.19	86.47

Analysis

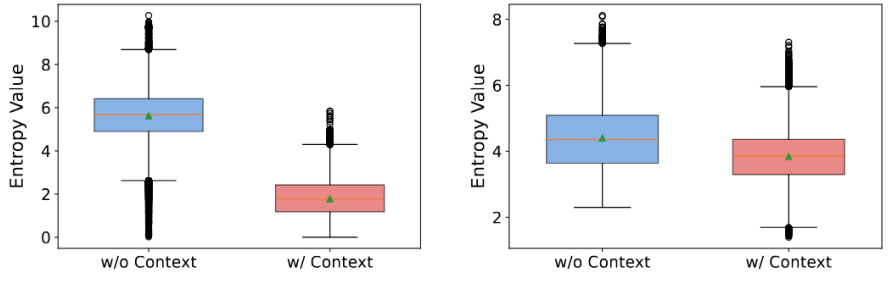


Figure 2: Distribution of entropy before and after adding the supplementary document that maximally reduces entropy for each high-entropy token. (a) shows the entropy distribution for tokens that pass verification before and after adding the retrieved document, (b) shows the corresponding entropy distribution for high-entropy tokens that fail verification before and after adding the retrieved document.

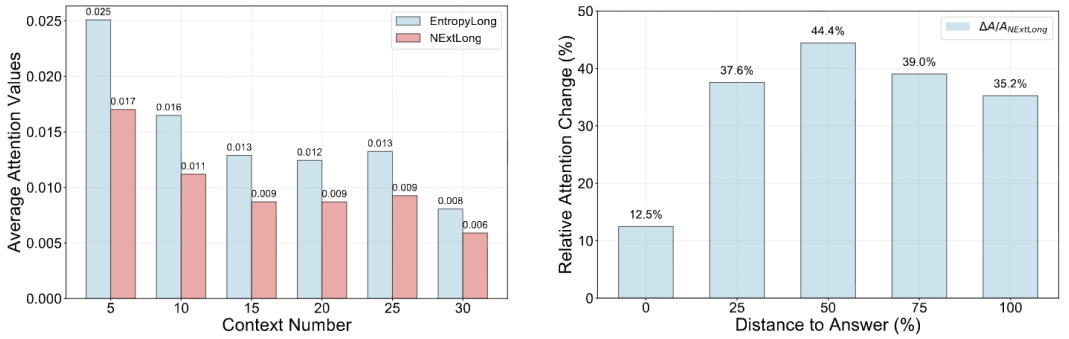


Figure 3: EntropyLong's attention patterns analysis. (a) Attention to correct answers vs NExtLong across different context chunks with answers at front; (b) Relative attention vs NExtLong with answers at different positions, where ΔA represents the attention difference.

Verified context genuinely reduces uncertainty

- Verified tokens: entropy drops from 5.62 → 1.70 after adding retrieved context
- Failed tokens: entropy barely changes (4.40 → 3.84) — verification correctly filters these out

Better attention patterns:

- Consistently higher attention to correct answers across all context lengths
- **Mitigates lost-in-the-middle: relative gains +12.5% to +44.4% at mid positions**

Conclusion & Takeaways

EntropyLong: from heuristic to evidence-based long-context data construction

The model's own uncertainty guides where to add context and whether it truly helps

- ✓ Novel framework: entropy-driven construction with model-in-the-loop verification
- ✓ 128K training corpus with verified long-range dependencies
- ✓ **RULER: 87.37 avg (+2.15 over NExtLong) | LongBench-v2: 27.60 (+3.50 over NExtLong)**
- ✓ Ablations confirm: verification is essential; optimal thresholds (α=2.0, ε=0.4) exist

Thank you! | jiajunlong@buaa.edu.cn