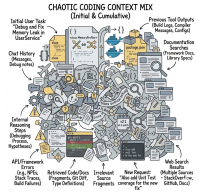
Oblivionis: Intrinsic Entropy of Context Length Scaling in LLMs

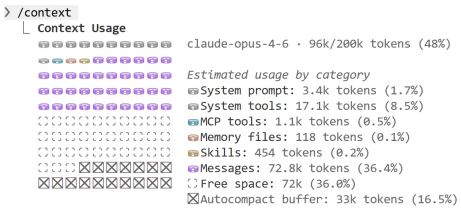
Jingzhe Shi*, Qinwei Ma*, Hongyi Liu*, Hang Zhao^, Jeng-Neng Hwang, Lei Li^

ICLR

1. Long Context Capability is Important...

e.g. Agentic Paradigm need Long Contexts:





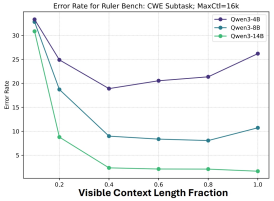
2. Longer Context is NOT Always Better

In Previous work:

Lost in the Middle
Context Rot
Multi-Instance Collapse
Irrelevant Context Distraction
...

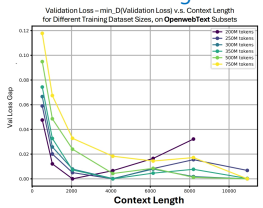
We further conducted experiments showing relevant context might hurt:

Scenario: Testing on downstream task with various visible context

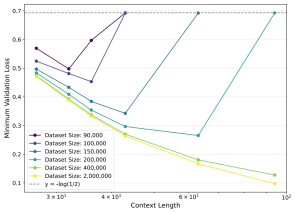


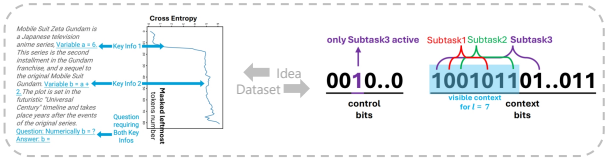
Error Rate (Left)
v.s. Visible Context Length Friction,
On RulerBench with fixed MaxCtl
across different model sizes
=>
More context might HURT for downstream task
(hurt less for larger models)

Scenario: Pretraining with various fixed context length

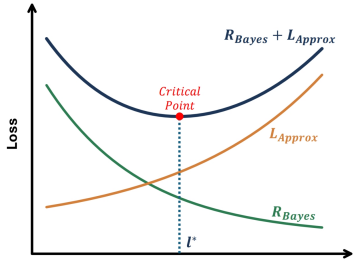


Error v.s. Training context length,
On Openwebtext (Left) &
On Synthetic Dataset (Right),
across different training set sizes
=>
More context may HURT for pre-training
(hurt less for larger training sets)





3. Analysis: Trade-off behind Longer Context

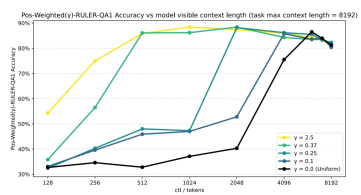


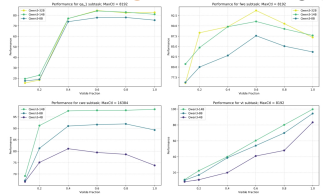
Loss Decomposition:

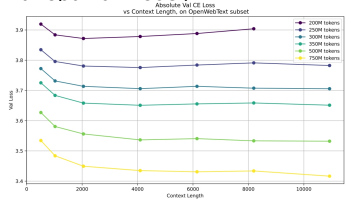
$$Loss = R_{Bayes} + L_{Approx}$$

R_{Bayes}: Bayes Risk
Loss given certain context for Bayes Model (Optimal Model possible)
smaller Bayes Risk if:
longer visible contexts (more information!)
more local tasks (shorter-range dependencies)

L_{Approx}: Approximation Loss
Gap between Actual Model and Bayes Model
smaller Approximation Loss if:
stronger models (less gap with optimal model)







Acc. v.s. Ctl for **task with different locality**
Acc. v.s. Ctl for **different models**
Loss v.s. Ctl for **different training data**

Larger γ = More Local Task
= Smaller Bayes Risk
= Smaller Optimal Context

Larger Model
= Smaller Approximation Loss
= Larger Optimal Context

Larger Training Data
= Smaller Approximation Loss
= Larger Optimal Context

4. Measure Context Information: Intrinsic Entropy

Question: How much information (for next token prediction) Is contained for certain Ctl?
Answer: Measuring by **Intrinsic Entropy**

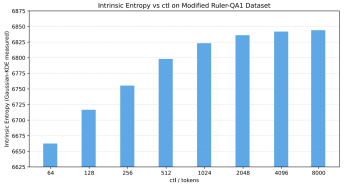
Intrinsic Entropy Calculation

$\{h_i^{(l)}\}_{i=1}^N$: final layer final token hidden state
for context length l across N samples
$\hat{q}_l$: density estimation by Gaussian-KDE

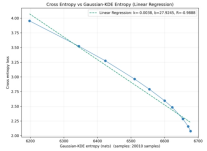
$$S(\hat{q}_l) = -\int \hat{q}_l(z) \log \hat{q}_l(z)\, dz$$

$$\hat{S}(P_l) \approx -\frac{1}{N}\sum_{i=1}^{N} \log \hat{q}_l(h_i^{(l)})$$

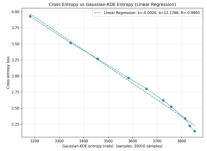
Intrinsic Entropy
measured for different Ctl
(saturation plateau)



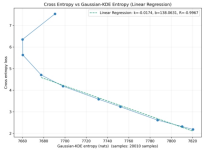
Intrinsic Entropy
v.s. Cross Entropy Loss



Llama-3.1-8B



Qwen-3-8B-Base



RecurrentGemma-9B

5. Looking Forward: Potentials of Discoveries:

1. Better Agent Designs
2. Better 'Entropy'-guided Exploration for RL training.

Scan for Paper