Support Basis: Fast Attention Beyond Bounded Entries

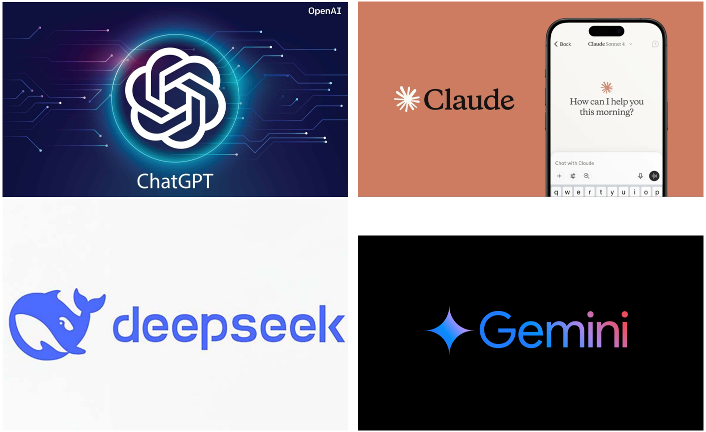
Maryam Aliakbarpour[1,2] Vladimir Braverman[3] Junze Yin[1] Haochen Zhang[1]

[1]Department of Computer Science, Rice University
[2]Ken Kennedy Institute, Rice University
[3]Department of Computer Science, Johns Hopkins University

May 2, 2026

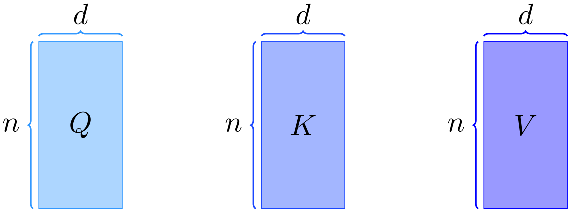
Large Language Models (LLMs)

- Large Language Models (LLMs) can perform a wide range of tasks, like question answering, summarization, paraphrasing & translation, and inference & logic.



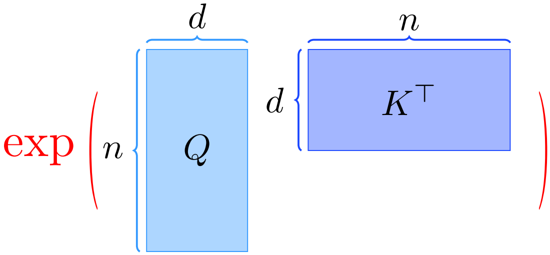
Attention Computation

- **Input**: $Q, K, V \in \mathbb{R}^{n \times d}$.



Attention Computation

- **Input**: $Q, K, V \in \mathbb{R}^{n\times d}$.
- **Compute**:
  - $A = \exp(QK^\top/d) \in \mathbb{R}^{n\times n}$.
    - exp is applied entrywisely.
    - $O(n^{2+o(1)})$ time.



$$\exp\left(\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}\right) = \begin{bmatrix} e^1 & e^2 \\ e^2 & e^4 \end{bmatrix}$$

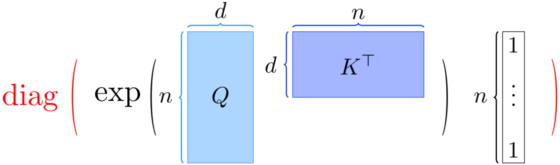
Attention Computation

- **Input**: $Q, K, V \in \mathbb{R}^{n \times d}$.
- **Compute**:
  - $A = \exp(QK^\top / d) \in \mathbb{R}^{n \times n}$.
    - exp is applied entrywisely.
    - $O(n^{2+o(1)})$ time.
  - $D = \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
    - $\mathbf{1}_n := \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^\top \in \mathbb{R}^n$.
    - $O(n^2)$ time.

$$\mathrm{diag}\left(\exp\left(n\left\{ \overbrace{\boxed{\;Q\;}}^{d} \right. \; d\left\{ \overbrace{\boxed{\;K^\top\;}}^{n} \right. \right) \; n\left\{ \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right. \right)$$

Attention Computation

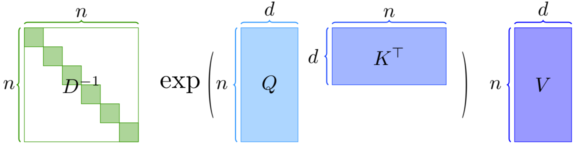
- **Input**: $Q, K, V \in \mathbb{R}^{n \times d}$.
- **Compute**:
  - $A = \exp(QK^\top / d) \in \mathbb{R}^{n \times n}$.
    - exp is applied entrywisely.
    - $O(n^{2+o(1)})$ time.
  - $D = \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
    - $\mathbf{1}_n := \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^\top \in \mathbb{R}^n$.
    - $O(n^2)$ time.
  - $\mathrm{diag} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$



$$\mathrm{diag}\left(\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Attention Computation

- **Input**: $Q, K, V \in \mathbb{R}^{n \times d}$.
- **Compute**:
  - $A = \exp(QK^\top / d) \in \mathbb{R}^{n \times n}$.
    - exp is applied entrywisely.
    - $O(n^{2+o(1)})$ time.
  - $D = \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
    - $\mathbf{1}_n := \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^\top \in \mathbb{R}^n$.
    - $O(n^2)$ time.
  - $\mathrm{diag} : \mathbb{R}^n \to \mathbb{R}^{n \times n}$
- **Output**: $D^{-1}AV$.



A Prior Work on Attention Optimization

[AS23][1] shows:

- $Q, K, V \in \mathbb{R}^{n\times d}$, $d = O(\log n)$.
- **Upper bound:** $\|Q\|_\infty, \|K\|_\infty, \|V\|_\infty \leq o(\sqrt{\log n})$ $\rightarrow$ subquadratic algorithm.
- **Lower bound:** Otherwise, no subquadratic algorithm with small error.

[1] Josh Alman and Zhao Song. Fast attention requires bounded entries. NeurIPS'23.

A Prior Work on Attention Optimization

[AS23][1] shows:

- $Q, K, V \in \mathbb{R}^{n \times d}$, $d = O(\log n)$.
- **Upper bound:** $\|Q\|_\infty, \|K\|_\infty, \|V\|_\infty \leq o(\sqrt{\log n}) \to$ subquadratic algorithm.
- **Lower bound:** Otherwise, no subquadratic algorithm with small error.

Pros

- Time efficient algorithm.

Cons

- Requires bounded-entry assumption.
- Ignores structural properties of $Q, K, V \in \mathbb{R}^{n \times d}$.

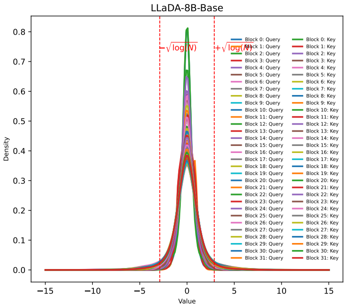
[1] Josh Alman and Zhao Song. Fast attention requires bounded entries. NeurIPS'23.

Our Result

- With high probability, we can accurately and efficiently approximate attention

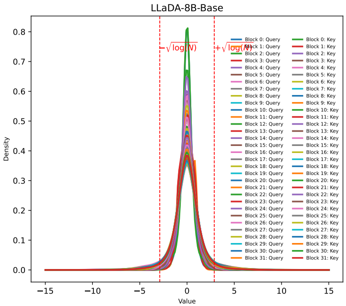
Our Result

- With high probability, we can accurately and efficiently approximate attention
- by using structural properties of $Q, K, V \in \mathbb{R}^{n \times d}$
  - Sub-Gaussianity of entry distribution.



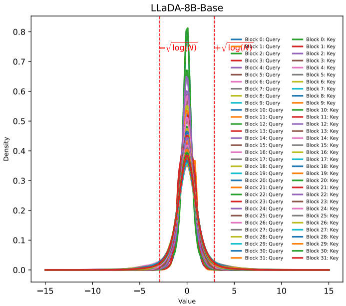
Our Result

- With high probability, we can accurately and efficiently approximate attention
- by using structural properties of $Q, K, V \in \mathbb{R}^{n \times d}$
    - Sub-Gaussianity of entry distribution.
- without bounded-entry assumption



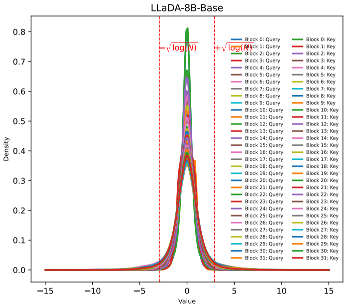
Our Result

- With high probability, we can accurately and efficiently approximate attention
- by using structural properties of $Q, K, V \in \mathbb{R}^{n \times d}$
  - Sub-Gaussianity of entry distribution.
- without bounded-entry assumption
- comparable downstream accuracy with exact computation



Our Result

- With high probability, we can accurately and efficiently approximate attention
- by using structural properties of $Q, K, V \in \mathbb{R}^{n \times d}$
    - Sub-Gaussianity of entry distribution.
- without bounded-entry assumption
- comparable downstream accuracy with exact computation
- adaptively adjust the accuracy-efficiency trade-off.



Divide: Support Basis Decomposition

- $A = \exp(QK^\top/d)$.
- $D = \mathrm{diag}\left(A\mathbf{1}_n\right)$.
- Goal: $D^{-1}AV$.

Disjoint matrices:

$$A^{(L)} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A^{(s)} = \begin{bmatrix} 0 & c & d \\ e & 0 & f \\ g & h & i \end{bmatrix}$$

$$A \to \exp\left(A^{(s)}/d\right) + \exp\left(A^{(L)}/d\right) - \mathbf{1}_{n\times n}.$$

$$AV \to \exp\left(A^{(s)}/d\right)V \quad \text{and} \quad \left(\exp\left(A^{(L)}/d\right) - \mathbf{1}_{n\times n}\right)V$$

AV is similar as $A\mathbf{1}_n$.

Conquer: approximate the attention in $A^{(L)}$ sparsity time.

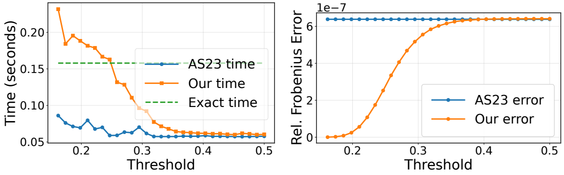
Via support basis decomposition,

$$\underbrace{\exp\left(A^{(s)}/d\right)V}_{\text{via [AS23]}} \quad \text{and} \quad \underbrace{\left(\exp\left(A^{(L)}/d\right) - \mathbf{1}_{n\times n}\right)V}_{\text{via exact computation}}$$

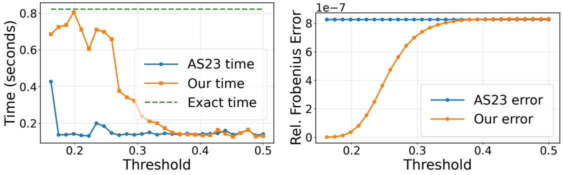
We have

	Error	Time
$\exp\left(A^{(s)}/d\right)V$	$1/\operatorname{poly}(n)$	sub-quadratic
$\left(\exp\left(A^{(L)}/d\right) - \mathbf{1}_{n\times n}\right)V$	0	sub-quadratic

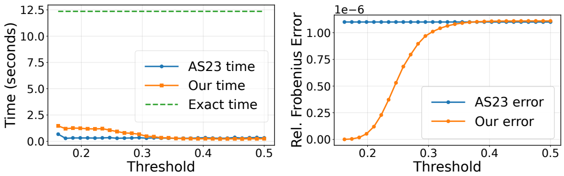
Experiment



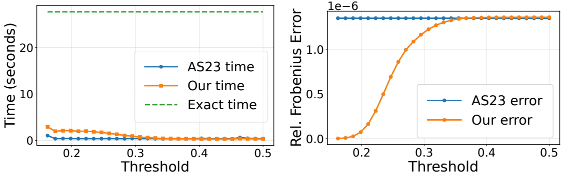
(a) $n = 8,192$



(b) $n = 16,384$



(c) $n = 24,576$



(d) $n = 32,768$

Thank you!

Thank you for your time!