BAYESIAN ATTENTION MECHANISM:

A PROBABILISTIC FRAMEWORK FOR POSITIONAL ENCODING AND CONTEXT LENGTH EXTRAPOLATION

MALTA

Machine Learning Theory and Applications Lab

Arthur Sperhacke Bianchessi
Yasmin Cardozo Aguirre
Rodrigo Coelho Barros
Lucas Silveira Kupssinskü

PUCRS

Definitions

Definition 1. For a fixed query vector $\mathbf{q}_i \in \mathcal{R}^{1\times d}$ and key-value matrices $\mathbf{K}, \mathbf{V} \in \mathcal{R}^{L\times d}$, a Bayesian Attention Mechanism computes self-attention as an expectation over its values:

$$\text{self-attention}(\mathbf{q}_i, \mathbf{K}, \mathbf{V}) = \frac{\exp\big(\text{score}(\mathbf{q}_i, \mathbf{k}_j)\big)}{\Sigma_z \exp\big(\text{score}(\mathbf{q}_i, \mathbf{k}_z)\big)}\mathbf{V} = \Sigma_j p_{ij}\mathbf{v}_j = \mathbf{E}_{j|i}[\mathbf{V}]$$

Definition 2. In Bayesian Attention, p_{ij} is a joint probability over the *content* of token j (f_{cont}) and its *position* relative to query $\mathbf{q}_i$ (g_{pos}).

$$p_{ij} = p(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j)|g_{\text{pos}}(i,j)) \times p(g_{\text{pos}}(i,j))$$

Additive Relative Positional Encodings

$$p_{ij} = \frac{\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j) + g_{\text{pos}}(i,j)\right)}{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z) + g_{\text{pos}}(i,z)\right)\right)}$$

$$= \frac{\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j)\right) \cdot \exp\left(g_{\text{pos}}(i,j)\right)}{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z) + g_{\text{pos}}(i,z)\right)\right)} \cdot \underbrace{\frac{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z)\right)\right) \cdot \Sigma_z\left(\exp\left(g_{\text{pos}}(i,z)\right)\right)}{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z)\right)\right) \cdot \Sigma_z\left(\exp\left(g_{\text{pos}}(i,z)\right)\right)}}_{=1}$$

$$= \frac{\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j)\right)}{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z)\right)\right)} \cdot \frac{\exp\left(g_{\text{pos}}(i,j)\right)}{\Sigma_z\left(\exp\left(g_{\text{pos}}(i,z)\right)\right)} \cdot \frac{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z)\right)\right) \cdot \Sigma_z\left(\exp\left(g_{\text{pos}}(i,z)\right)\right)}{\Sigma_z\left(\exp\left(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z) + g_{\text{pos}}(i,z)\right)\right)}$$

$$= \mathbf{p}(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j)) \cdot \mathbf{p}(g_{\text{pos}}(i,j)) \cdot \frac{1}{\Sigma_z\left(p(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_z)) \cdot p(g_{\text{pos}}(i,z))\right)}$$

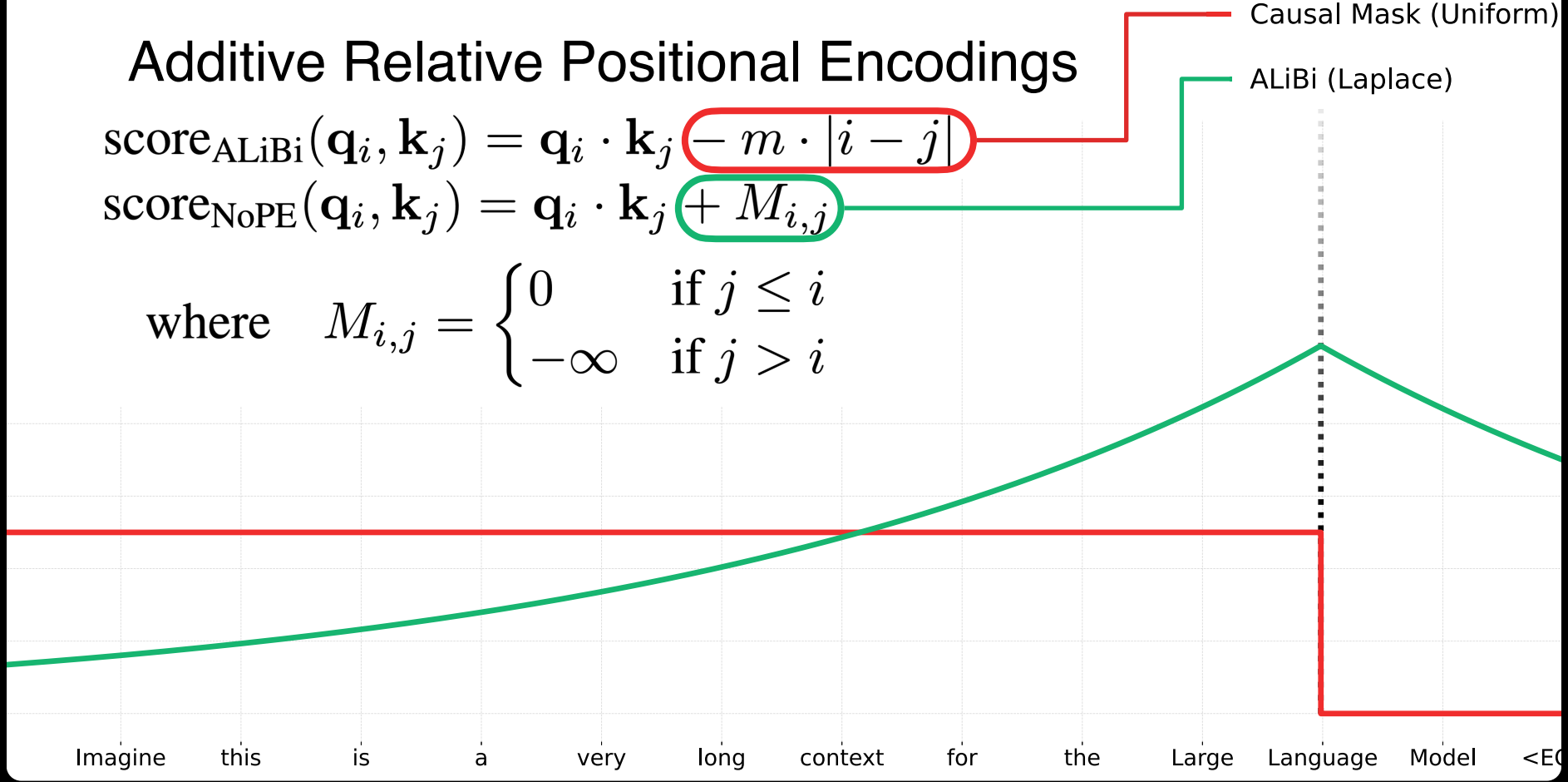
$$= \frac{\mathbf{p}(f_{\text{cont}}(\mathbf{q}_i, \mathbf{k}_j)) \cdot \mathbf{p}(g_{\text{pos}}(i,j))}{Z}$$

Additive Relative Positional Encodings

$$\text{score}_{\text{ALiBi}}(\mathbf{q}_i, \mathbf{k}_j) = \mathbf{q}_i \cdot \mathbf{k}_j - m \cdot |i - j|$$

$$\text{score}_{\text{NoPE}}(\mathbf{q}_i, \mathbf{k}_j) = \mathbf{q}_i \cdot \mathbf{k}_j + M_{i,j}$$

$$\text{where} \quad M_{i,j} = \begin{cases} 0 & \text{if } j \leq i \\ -\infty & \text{if } j > i \end{cases}$$



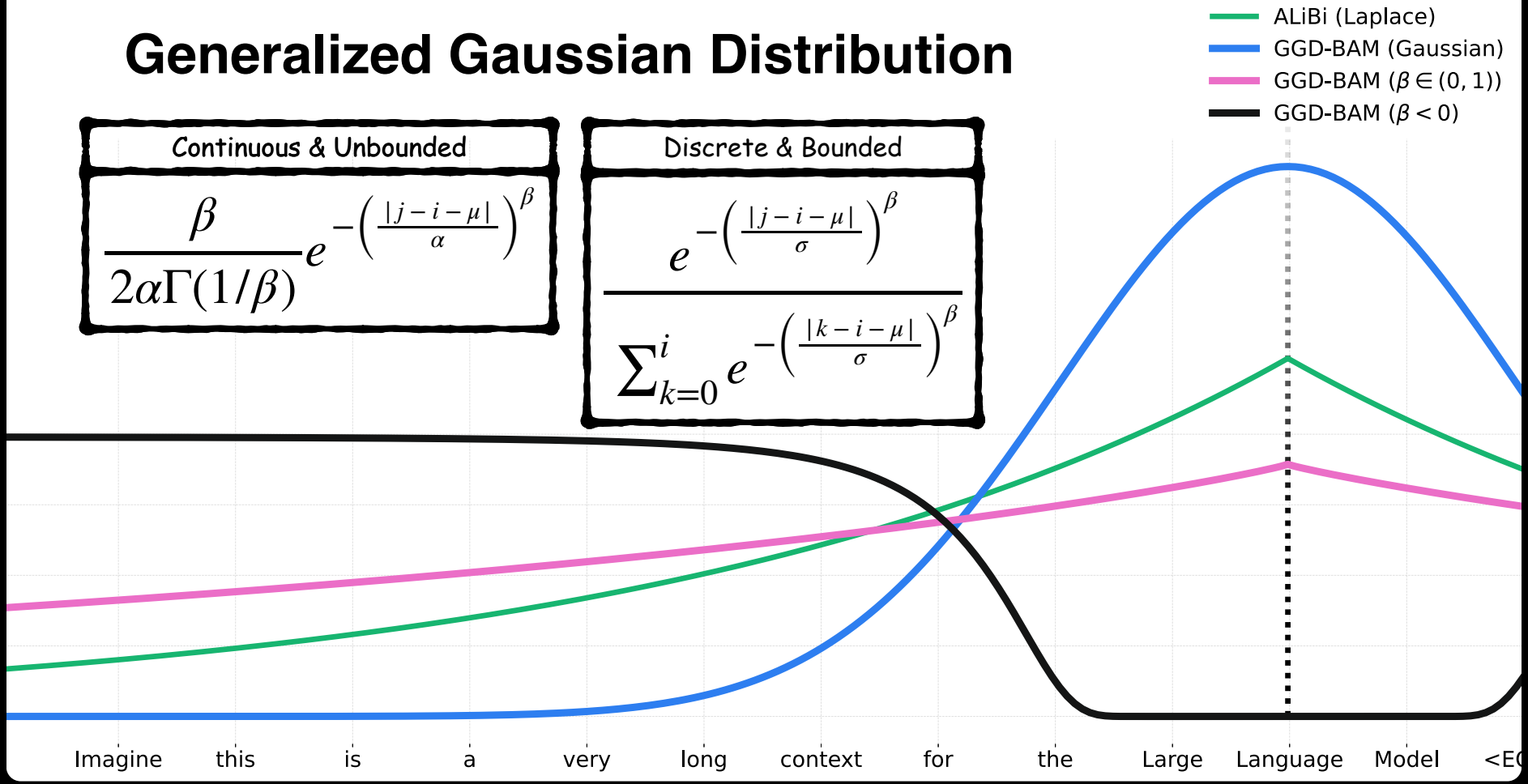
Generalized Gaussian Distribution

Continuous & Unbounded

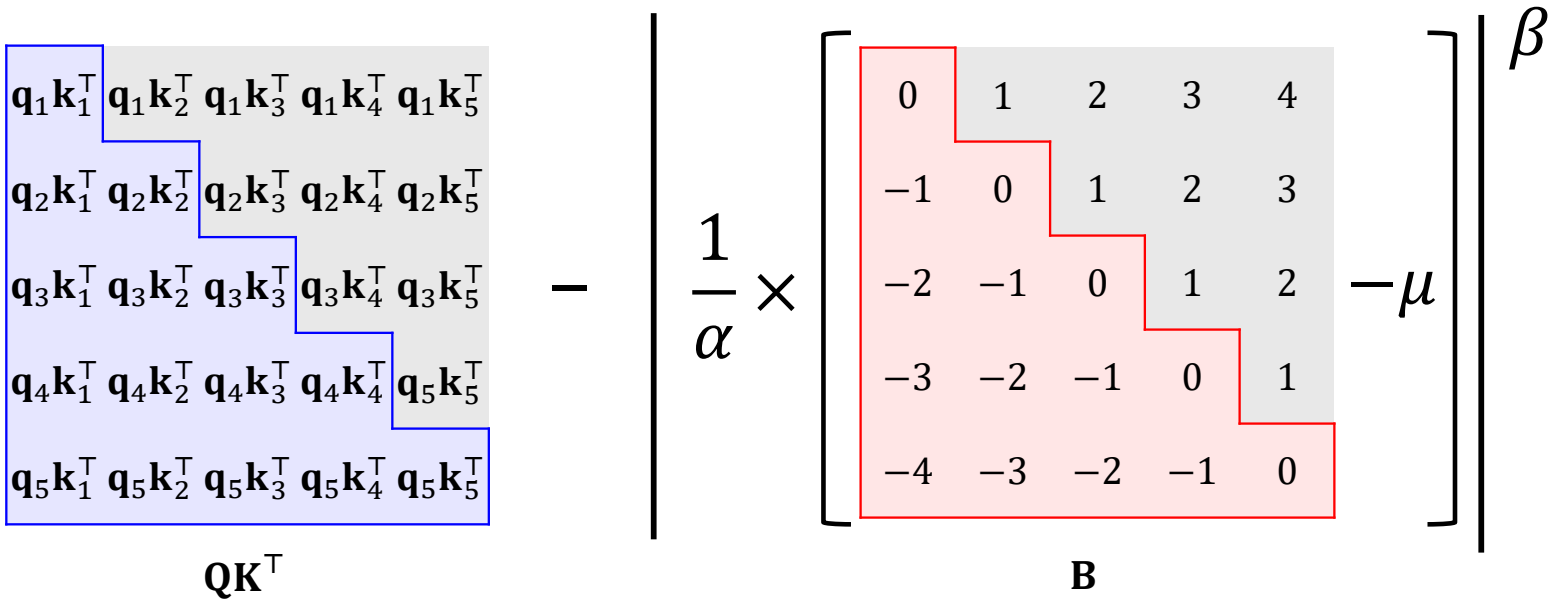
$$\frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-\left(\frac{|j-i-\mu|}{\alpha}\right)^{\beta}}$$

Discrete & Bounded

$$\frac{e^{-\left(\frac{|j-i-\mu|}{\sigma}\right)^{\beta}}}{\sum_{k=0}^{i} e^{-\left(\frac{|k-i-\mu|}{\sigma}\right)^{\beta}}}$$



Generalized Gaussian Distribution



$$\mu = 0 \qquad \alpha = e^{\theta_\alpha} \qquad \beta = \theta_\beta$$

Evaluation I:

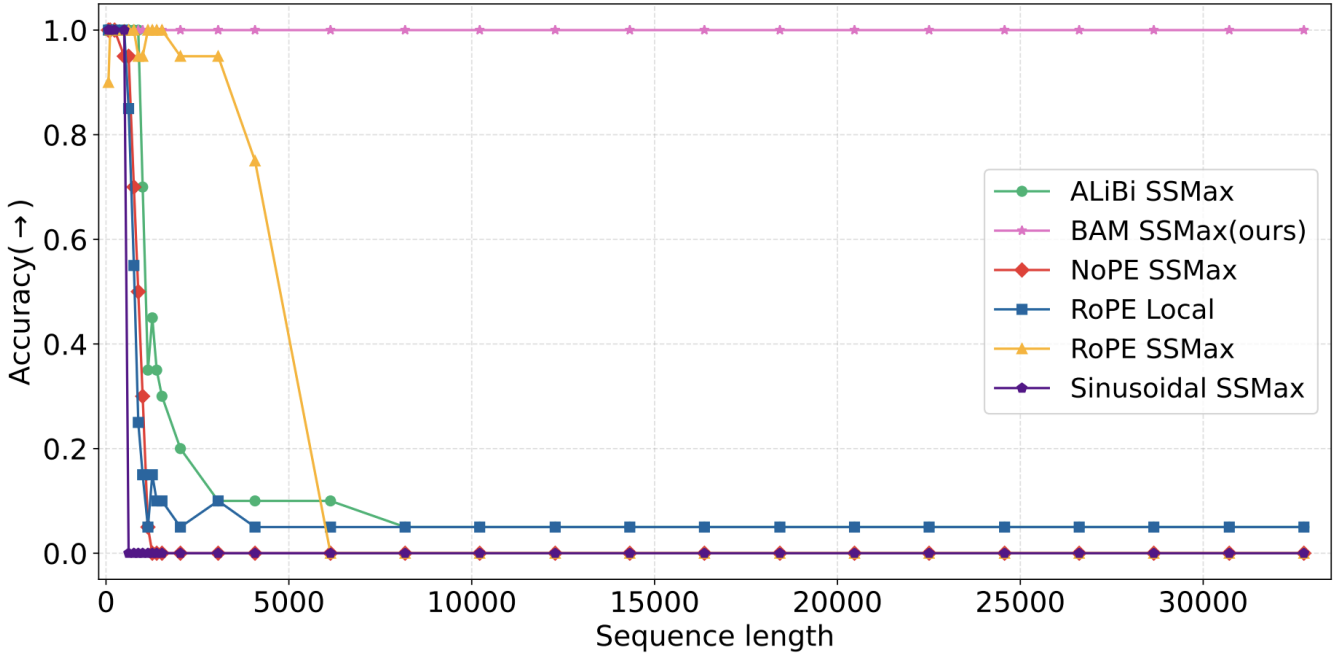
Evaluating Extrapolation

PUCRS

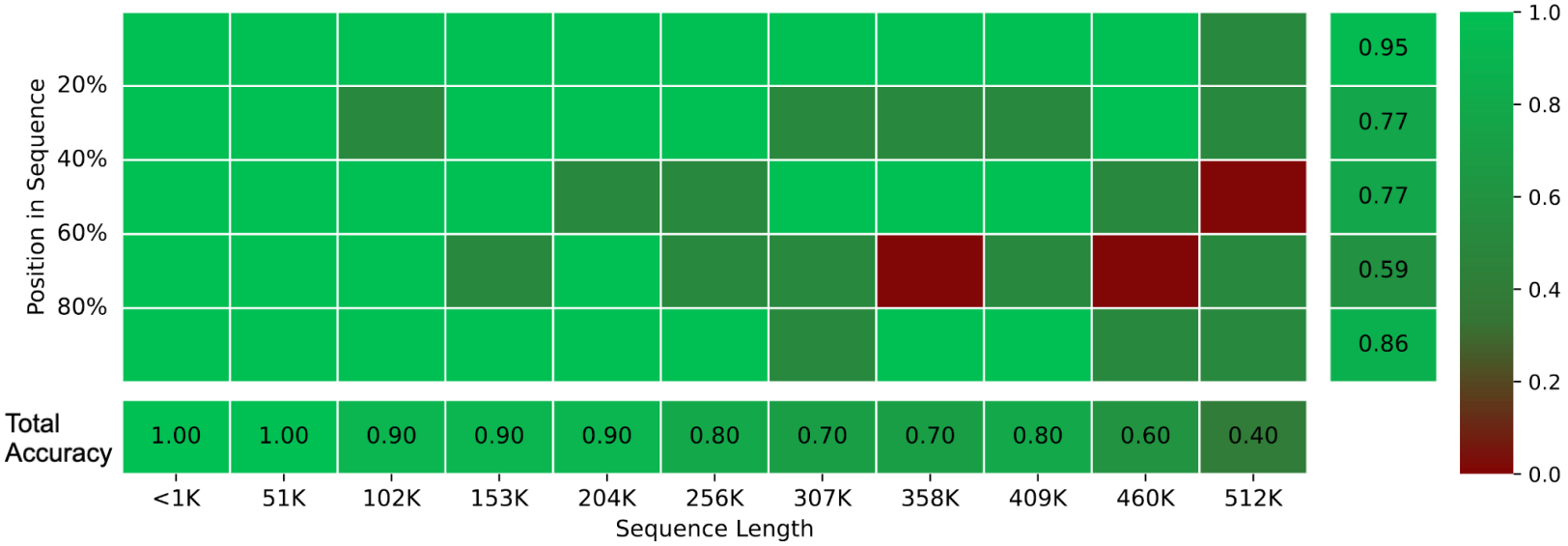
MALTA

Machine Learning Theory and Applications Lab

Passkey Retrieval



Passkey Retrieval (GGD-BAM)



Evaluation II:

Evaluating Extrapolation & **Performance**

1 Billion Parameters Models

+

Instruction Tuning

PUCRS

MALTA

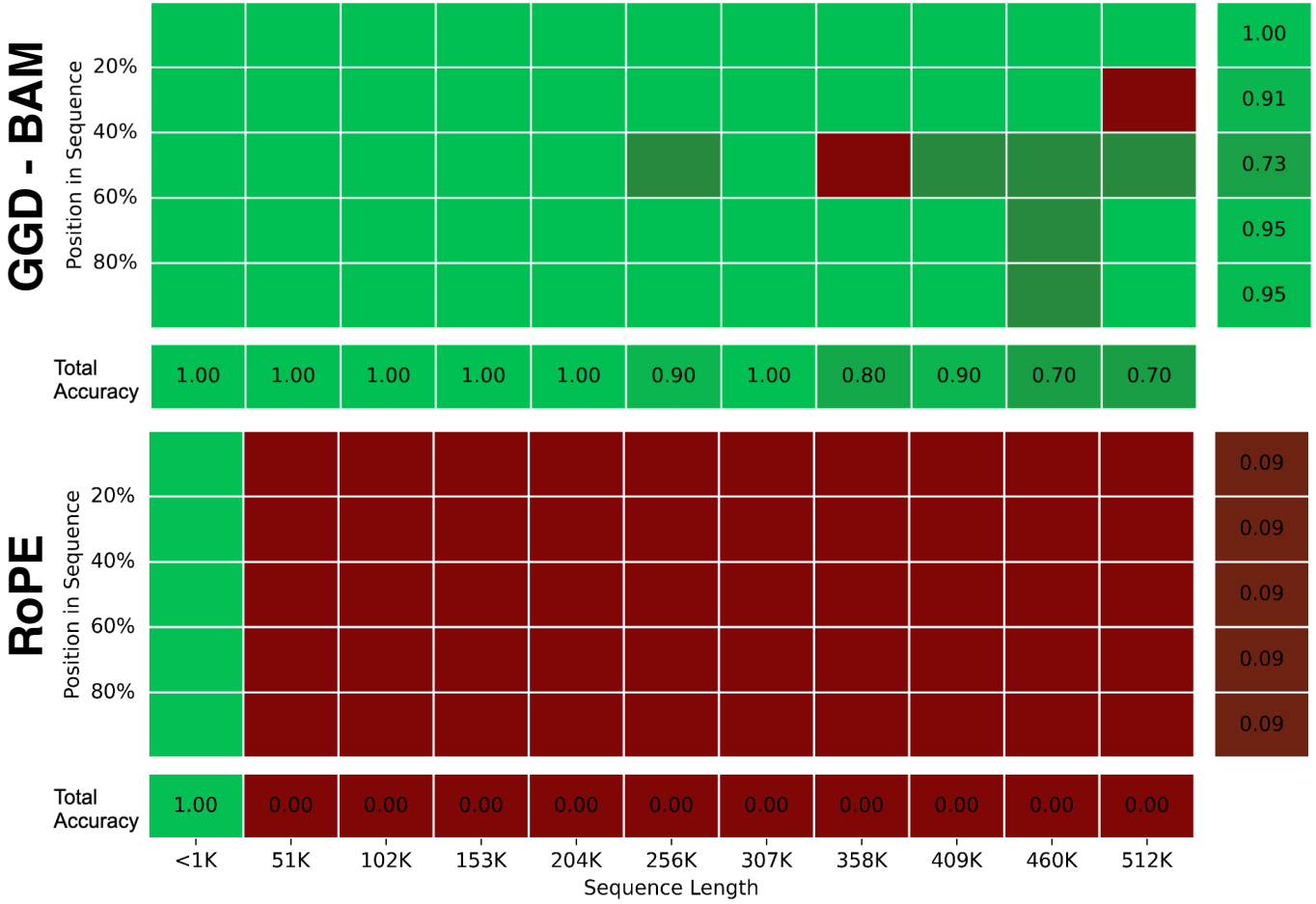
Machine Learning Theory and Applications Lab

Passkey Retrieval

1 Billion Parameter Models

+

Instruction Tuning



Performance Benchmarks

1 Billion Parameter Models + Instruction Tuning

	MMLU	ARC-Easy	**ARC-Challenge**
BAM SSMax	**0.3716**	**0.5770**	**0.4132**
RoPE SSMax	0.3573	0.5715	0.4123

Performance on Long Context

1 Billion Parameter Models + Instruction Tuning

Table 6: LongBenchv2 Benchmark: GGD-BAM vs RoPE, 1B parameter models.

	BAM SSMax	RoPE SSMax
Code Repository Understanding	**41.7**	25.0
Long In-context Learning	**36.4**	30.3
Long-dialogue History Understanding	35.0	35.0
Multi-Document QA	**26.5**	25.3
Single-Document QA	**26.5**	18.8
Overall	**28.6**	24.2

BAYESIAN ATTENTION MECHANISM:

A PROBABILISTIC FRAMEWORK
FOR POSITIONAL ENCODING
AND CONTEXT LENGTH EXTRAPOLATION