Supplementary Data 1 A list of symbols

Symbol	Meaning
$AND \in \mathbb{R}^{d\times In}$	an embedding matrix
In	the size of vocabulary
d	dimension of embedding vector
$x_t \in \mathbb{R}^{In}$	input token
$p_t \in \mathbb{R}^{d}$	positional encoding
$X \in \mathbb{R}^{n\times d}$	input embedding matrix
$Q \in \mathbb{R}^{n\times d_k}$	query matrix
$K \in \mathbb{R}^{n\times d_k}$	key matrix
$V \in \mathbb{R}^{n\times d_v}$	value matrix
$W_Q \in \mathbb{R}^{d\times d_k}$	weight matrix for query
$W_k \in \mathbb{R}^{d\times d_k}$	weight matrix for key
$W_v \in \mathbb{R}^{d\times d_v}$	weight matrix for value
h	number of heads
$W_O \in \mathbb{R}^{hd_v\times d}$	multi-head weight matrix
$W_1 \in \mathbb{R}^{d_{model}\times d_f}$	weight matrix
$W_2 \in \mathbb{R}^{d_f\times d_{model}}$	weight matrix
$b_1 \in \mathbb{R}^{d_f}$	bias vector
$b_2 \in \mathbb{R}^{d_{model}}$	bias vector
L	number of layers in the transformer
LayerNorm	layer normalization
FFN	Feed-Forward Network
MultiHead	Multi-Head Attention
$Z^{(L)}$	output of the final layer in the transformer
s	state in the RL
$\mathbb{D}^d$	Poincaré ball
$exp_0(\cdot)$	Exponential map from the tangent space at the origin (Euclidean) to $\mathbb{D}^d$
$\log_0(\cdot)$	Logarithmic map from $\mathbb{D}^d$ to the tangent space at the origin (Euclidean).
$\oplus$	Möbius addition in $\mathbb{D}^d$
$\lambda\otimes x$	Möbius scalar multiplication
$<\cdot, >$	standard Euclidean inner product
$\emptyset(v)$	exponential map
$u_t = e_t + p_t$	Input to the transformer
$\tilde{u}_t \in \mathbb{D}^d$	hyperbolic point
$\text{HLayerNorm}(\tilde{z})$	LayerNorm in the hyperbolic space
HMultiHeadAtten	multi-head attention in the hyperbolic space
HFFN	Feed-Forward Network in the hyperbolic space
HL	Hyperbolic Residual & LayerNorm
E_s	state embedding
E_a	action embedding
$\tilde{u}_s \in \mathbb{D}^d$	map the state embeddings in the Euclidean space to the Poincaré Ball
$\tilde{u}_a \in \mathbb{D}^d$	map the action embeddings in the Euclidean space to the Poincaré Ball

$\widetilde{X} \in \mathbb{D}^{N\times(2d)}$	the total state and action embedding in the Poincaré ball
$\widetilde{X}^{(l)} \in \mathbb{D}^{N\times(2d)}$	state and action representation at layer l in the Poincaré ball
$X^{(l-1)}$	state and action representation at layer l in the tangent space
$Q_h, K_h, V_h \in \mathbb{R}^{\boldsymbol{N\times d_h}}$	Queries, Keys, Values in tangent space
$W_h^Q\ , W_h^K, W_h^V$	weight matrices in tangent space
H	number of heads
$W^o \in \mathbb{R}^{Hd_h\times(2d)}$	head weight matrix in tangent space
MAtten	Multi-head Attention in tangent space
HMAtten_i	multi-head attention in hyperbolic, index i denotes the i^{th} episode
HMAtten	total multi-head attention in hyperbolic
HLayerNorm	Hyperbolic layer normalization
$\tilde{z}_i^{(l)}$	hyperbolic residual connection
$\widehat{s'}$	predicted next state
W_s	weight matrix for prediction of next state
$\hat{r}$	predicted next reward
W_r	weight matrix for prediction of next reward
MLA	Multi-head Latent Attention
$X \in \mathbb{R}^{T\times d}$	input sequence to MLA
$h_t \in \mathbb{R}^d$	the attention input of the t^{th} token at alatent attention layer
$q_t, k_t, v_t \in \mathbb{R}^{n_h\times d_h}$	query, key and value vectors in MLA
W^Q, W^K, W^V	linear mapping matrices for projecting h_t to queries, keys, and values
$q_{t,i}, k_{t,i}, v_{t,i} \in \mathbb{R}^{d_h}$	query, key, and value of the i^{th} attention head
$W^O \in \mathbb{R}^{d\times n_h d_h}$	output projection matrix
C_t^{KV}	compressed latent vector for keys and values
$d_c(\ll n_h d_h)$	KV compression dimension
W^{DKV}	down-projection matrix
$W^{UK} \in \mathbb{R}^{n_h d_h\times d_c}$	up-projection matrices for keys
$W^{UV} \in \mathbb{R}^{n_h d_h\times d_c}$	up-projection matrices for values
$c_t^Q \in \mathbb{R}^{d_c^{'}}$	compressed latent vector for queries
$d_c^{'}(\ll n_h d_h)$	query compression dimension
$W^{DQ} \in \mathbb{R}^{d_c^{'}\times d}$	the down-projection matrix for queries
$W^{UQ} \in \mathbb{R}^{n_h d_h\times d_c^{'}}$	up-projection matrices for queries
RoPE	Rotary Positional Embeddings
$q_{t,i}^R \in \mathbb{R}^{d_h^R}$	additional Multihead queries
k_t^R	shared key
d_h^R	per-head dimension of the decoupled queries and key
$W^{QR} \in \mathbb{R}^{n_h d_h^R\times d_c^{'}}$	Matrix to produce the decouples queries
$W^{KR} \in \mathbb{R}^{n_h d_h^R\times d}$	matrix to produce the decouples keys
u_t^l	FFN input of the t^th token
N_s	numbers of shared experts
N_r	numbers of routed experts

$\text{FFN}_i^{(s)}(\cdot)$	the i^{th} shared expert
$\text{FFN}_i^{(r)}(\cdot)$	the i^{th} routed expert
k_r	number of activated routed experts
$g_{i,t}$	gate value for the i^{th} expert
$s_{i,t}$	token to-expert affinity
e_i	centroid of the i^{th} routed expert in this layer
Topk($\cdot$, K)	the set comprising K highest scores among the affinity scores calculated for the t^{th} token and all routed experts
$\mathcal{L}_{\text{ExpBal}}$	auxiliary losses, for controlling expert-level load balance
$\mathcal{L}_{\text{DevBal}}$	auxiliary losses, for controlling device-level load balance
$\mathcal{L}_{\text{CommBal}}$	auxiliary losses, for controlling communication balance
α_1	expert-level balance factor
f_i	fraction of token rooted to the i^{th} expert
p_i	average probability of selecting the i^{th} expert for the entire input sequence
α_2	device-level balance factor
$\tilde{z}_i \in D^d$	a set of hyperbolic token embeddings
$\tilde{l}_j \in D^d$	a set of latent embeddings
$\alpha_{i,x}$	Coefficients for aggregating values in Hyperbolic Multi-Head Latent Attention
$\hat{v}_i$	aggregated value in the tangent space in MLA
$\tilde{v}_i$	aggregated value in the Poincaré ball of hyperbolic MLA
$\tilde{z}_i \in \mathbb{D}^d$	each token in Router, the Poincaré ball.
$z_{\boldsymbol{i}}$	token in Router, the tangent space.
$\tilde{y}_i$	output in HFNN
s	state in Hyperbolic Transformer as a Policy
x_t	token in Hyperbolic Transformer as a Policy
$\tilde{h}$	hyperbolic representation
h	Representation in tangent space
$\pi_\theta(a\|s)$	Policy distribution of action *a* given state *s*.
GRPO	Group Relative Policy Optimization
θ	Parameter in policy distribution model
$\{o_1, o_2, \ldots, o_G\}$	a group of outputs in GRPO
$\pi_{\theta_{old}}$	old policy
$\{s^{(j)}\}_{j=1}^N$	a batch of N states sampled from the environment
$\mathcal{A}(s^{(j)}) = \{a_1^{\{j\}}, \ldots, a_G^{\{j\}}\}$	For each state $s^{(j)}$, a set (or group) of sampled G actions
$r(s^{(j)}, a^{(j)})$	a reward function
$\mu(s^{(j)})$	Group Mean Reward
$\sigma(s^{(j)})$	Group Reward Standard Deviation.
$A\left(s^{(j)}, a_i^{(j)}\right)$	the group-relative advantage for each action $a_i^{(j)}$
$\pi_\theta\left(a_i^{(j)}\|s^{(j)}\right)$	new policy's probability using our hyperbolic Transformer
$\rho_1\left(s^{(j)}, a_i^{(j)}\right)$	probability ratios relative to the old policy

$\rho_2\left(s^{(j)}, a_i^{(j)}\right)$	probability ratios relative to reference
$D_{KL}(\pi_\theta \| \pi_{ref})$	K-L distance between the policy distribution and reference
$L(s^{(j)}, \theta)$	surrogate objective for each state $s^{(j)}$ in GRPO
$L(\theta)$	total surrogate objective over the batch
η	learning rate
CoT	chain-of-thought
$y = (y_1, ..., y_T)$	reasoning chain
$R(s, y, a)$	Reward function
$J(\theta)$	expected reward
$\nabla_\theta J(\theta)$	gradient of expected reward
$\Delta\theta$	Parameter update
$\mathcal{L}$	Loss between true and predicted state and reward
λ	a balancing hyperparameter
$\mathcal{L}_{Euclidean}$	Loss between true and predicted state and reward for using Standard Transformer in Euclidean Space
$\mathcal{L}_{Hyperbolic}$	Loss between true and predicted state and reward for using Hyperbolic Transformer in the Poincaré Ball
ε	clipping parameter