Cache Memory & Token Scaling