ID de Contribution: **15** Type: **Non spécifié**

Accelerating Large Langage Model Inference with Self-Supervised Early Exits

jeudi 19 septembre 2024 11:20 (15 minutes)

Orateur: VALADE, Florian (LAMA)