Key Trends: 4-bit quantization has emerged as the dominant practical technique with favorable efficiency/accuracy trade-offs and growing hardware support.Combined compression methods consistently outperform single techniques.

lakhasly.com © تم تلخيص النص بواسطة موقع لخصلي