

ã?¢ã??ã•¥ã•ã??å¼ã•§ã??ã??ã•?ã•?ã•?æ?•ã•£ã•|ã•?ã??Alã•®ç ?ç©¶¶

Description

ç ?ç©¶¶ã•?æ?,ã•ã•?ã°?ã•?ã•?Sèç?è£?ã•ã•®ã•§ã??ã¼ã•?æ?,ã•?ã•?æ?¹ã•?è?¬ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
¼ã•?ã•?ã•?

[ã?¢ã??ã•¥ã•ã??å¼ã•?ã•?ã¼ã•?ã•?ã•?æ??ç?¹ã•?§ã•?¬ç?æ?¼ã•?è?ã°?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
ã•?](#)

å•?Sè|æ?jè?è?ã•?ã•?ã•?ã•?«i¼?LLMi¼?ã•?é?¢ã•?ã•?ã•?ã•?ã•?ã•?çµ?ã•?ç

RAGi¼?Retrieval Augmented Generationi¼?ã•?é?•å•?ã•?i¼?Quantizationi¼?ã•?ã•?©ã•
®æ??æ³?ã•?ã¼ã•?£ã•?|ã•?ã•?¹ã•?ã•?æ§ç?¬ã•?ã•?ã•?ã•?ã•?ã•?ã¼ã•?ã•?ã•?ã•?

- ç§ã•®ã•?ã•?ã•?è?¬ã°?ã•?ã•?ã•?REST APIã•?ã¼ã•?£ã•?|æ??ç?«ã•?ã•?ã¼ã•?ã•?ã•?
|ã•?ã•?¼ã•?çã•?ã•?¼ã•?¹ã•?ã•?ã•?
- ã•?ã•?¼ã•?çã•?ã•?¼ã•?¹ã•?®æ??ç?«ã•?ã•?ã•?ã•?çã•?¼ã•?ã•?¼ã•?¹ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
- RAGã•?ã¼ã•?£ã•?|ã•?ã•?ã•?è?¬ã°?ã•?ã•?ã•?ã•?ã•?ã•?ã•?«èç¼å?ã•?ã•?æ??æ?,ç??æ?ã•?ã•?
ã•?
- LLMã•?ã•?ã•?GPUã•?®VRAMã•?§ã•?ã•?ã•?ã•?ã•?«é?•å•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
- åç è|ã•?ã•?ã•?ã•?ã•?jã•?ã•?³ã•?ã•?¥ã•?¼ã•?ã•?ã•?ã•?ã•?èjã•?ç??æ?æ??ã•?®èjç?
¼ã•?èæç?ã•?ã•?
- é«é??ã•?ã•?ã•?ã•?ã•?«CTranslate2ã•?ã¼ã•?

1ã•?2ã•?cronã•?ã•?©ã•?§ã•?æ??ç?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
®ã•?ã•?¼ã•?çã•?ã•?¼ã•?£ã•?|æ?è«ã•?ã•?§ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?

ã•?ã•?ã•?ã•?ã•?çç?ã•?ã•?ã•?ã•?ç??æ?Alã•?ã•?æ¥ã•?ã•?«é©ç?¬ã•?ã•?é?ç?ã•?æ?¢ã•?ã•?|ã•?ã•?
ã•?

ç?»ã•?è?è?ã•?ç?©ã¼?æçç?¥ã•?«é?¢ã•?ã•?ã•?ã•?çµ?ã•?ç

ã•?ã•?ã•?Alã•?®ã•?é??ã•?§ã•
ã•?ã•?ã•?ã•?ã•?ã•?«ã•?ã•?¼ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
ã•?ã•?ã•?ã•?ã•?é??ç?°ã•?®ç ?ç©¶¶ã•?§ã•?ã•?ã•?é©ç?¬ã•?é??ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
¬é?ã•?ã•?ã•?ã•?æ?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?

Transformerã•?Diffusionã•?¢ã•?ã•?ã•?ã•?ã•?ã•?ã•?è?ã¼?ã•?ã•?

è«æ??ã•?ã•?ã•?ã•?ã•?ã•?è£ ä¼ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
«ã•?¹ã•?¬ã•?©ã•?
ã•?ã•?ã•?ã•?ã•?é??ç?°ã•?®ç ?ç©¶¶ã•?§ã•?ã•?ã•?é©ç?¬ã•?ã•?é??ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
¬é?ã•?ã•?ã•?ã•?æ?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
[ã?¢ã??ã•¥ã•ã??å¼ã•?ã•?ã¼ã•?ã•?ã•?æ??ç?¹ã•?§ã•?¬ç?æ?¼ã•?è?ã°?ã•?ã•?ã•?ã•?ã•?ã•?ã•?
ã•?](#)

