



ä½?æ??ã•®ã??ã?ã?°ã?©ã??ã?³ã?° â?? Inference Pipeline PoCã?¢ã??ã?ªã??ã•
ã??ã•£ã•½ã•ã•?ã??

Description

Inference Brokerã•®ç?´é??ã??å°?ã•?ã??ã•?ã•?ã•®ã?¢ã??ã?ªã??æ?jå¼µã•?ã•!ã•?ã•?ã•ã•
?å®?´é?´ã??ã•?ã•¼ã•?ã??

å?•å??ã•¼ã•§ã•ã?ã?ã•£ã??ã??ã?¹ã?–ã??ã??å?´ä½?ã??æ,jã•?ã•!ã•?Stable Diffusionç?´ã•®ã??ã?-
ã?³ã??ã??ç??æ?ã??ã•?ã•?ã•?ç?»å?ã??ç??æ?ã•?ã??ã•?ã•?ã•?ã•?ã•®ã•§ã•?ã•?ã•?ã•?é?ã•
?ã?¹ã?–ã??ã??ã•§ã??ã•ã??ã??ã•?ã•?ç?»å?ã??ç??æ?ã•§ã•ã??ã??ã•?ã•«ã?•
ã?¹ã?–ã??ã??ã?jã??ã?»ã?¼ã?,ã??ã•ã•®ã•¼ã•¼LLMã•«æ,jã•?ã•®ã•§ã•ã?ã•ã?¹ã?¿ã?¼ã?³ã•?ã•ã•
«æ??è??ã??æ?½å°?ã•?ã?ã?¹ã?–ã??ã??å? å•®æ??è??æ?ã??æ,jã•?ã??ã•?ã•«ã•?ã•¼ã•?ã•?ã??

ã•?ã??ã•ªã??ã?ç?¶æ³•ã•«ã??ã??ã•¼ã•?ã•?ã•?ã?³ã?³ã??ã?–ã?¹ã??ã?!ã?£ã?³ã??ã?!ã??æ?°å??ã•®1ã•
«å?§ç,®ã•§ã•ã??ã•ã?ã•§ã•?ã??

å??ã•?ã•?æ§?å•ã•?ä,?ã•®ã?¹ã?–ã?ªã?¼ã?³ã?ã•§ã??ã??ã§ã•?ã??



ã•ã??ã•ã??ã•ã•®é??ã°|ã•ã°ã??ã•ã•ã•ã°¹ã??ã°ªã°¼ã??ã°³ã°½ã°¾ã°¿ã°Àã°Áã°Âã°Ãã°Äã°Åã°Æã°Çã°Èã°Éã°Êã°Ëã°Ìã°Íã°Îã°Ïã°Ðã°Ñã°Òã°Óã°Ôã°Õã°Öã°×ã°Øã°Ùã°Úã°Ûã°Üã°Ýã°Þã°ßã°àã°áã°âã°ãã°äã°åã°æã°çã°èã°éã°êã°ëã°ìã°íã°îã°ïã°ðã°ñã°òã°óã°ôã°õã°öã°÷ã°øã°ùã°úã°ûã°üã°ýã°þã°ÿ

è©|ă•ă«Qwen 3.8 27Bă??ă?•VRAMă«32GBă?ă??ă½?ă|ă•?ç?¶æ ?ă•Să??ă?ă?ă|ă•¿ă•¼ă•ă?ă•
ă??
ă?³ă?³ă??ă?¬ă?¹ă??ă?µă?¶ă?°ă¬ă?¶ă??ă?«ă®ă,?é?•256Kă«è"ă®?ă?ă?ç?¶æ ?ă•Să?ă??
ă??ă•ă«ă¬ă??ă•ă¾ă?ă?ă?ç??æ?é??ă°|ă¬1ætoken / secă•Să?ă?ă??
ă?ă?ă•ă•ă??ă?ă•ă®æ?Sè?½ă•ă®MoEă?ă??ă?«ă?ă°ă?ă??ă?ă•ă¾ă?è©|ă•ă?ă?ă?ă•æ?ă?ă•ă¾ă•
ă??
ă?ă•ă®é•ă?ă•ă¬ă?iGPUă•ă®æ?Sè?½ă•ă?
ă??
ă?ă•¶ă?ă?ă?•æ??ă??æ?¬éç??ă•ă•ă®ă¬ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?
Strix Haloă??DGX Sparkă•Să?? 256Gbps +ç"?ă°|ă•Să?ă??
ă??ă•£ă•ă°?ă?ă?ă¬ă??ă?æ¬ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?
é?•è|ă•ă•ă®ă¬ă?è"?ç®?ă??ă??ă??ă??ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?ă?
1TBè¿ă?
é?»ă??PCă??
ă•ă?ă?ă?ă?ă?ă?ă®ă?æ¬ă?ă?ă?ă?ă?ă?ă?

Category

1. æ?¥ă?ă•ă®æ'»ă??

Tags

1. AI
2. Contextual Computing
3. Inference Broker
4. LLM
5. MoE
6. Stable Diffusion
7. ä,?é??ă??ă?
8. â±±æç??

Date Created

2026ă¹9æ??17æ?¥

Author

kazuo-tsubaki