



ç??æ?•Alã??a^{1/2}çã•£ã|arXivè«?æ??ã??è^aã??æ?¹æ³?

Description

ZIKUUã•®è«?æ??æµ?ç´ ¢ã??ã??ã??ã•§ã??ã•£ã•|ã•?ã??ã•®ã•¨ã¼¼ã•?æ?¹æ³?ã•§ã•?ã•-
ã?¼ã•?«ã•?«LLMã??ã½½ã•£ã•|arXivè«?æ??ã??ã•?ã¼ã•?ã•?ã•|æ|?è|ã•?è|çµã•?ã??æ?¹æ³?ã•?ç´¹ã»?ã
•ã•¾ã•?ã??

å??ä½?ç?°å¢?ã•-ã?

- 13th Core i5/RTX 2000 Ada
- llama.cpp
- Open WebUI

ã•§ã•?ã??

Open WebUI•§ă??ă?©ă?«ă??ă?¼ă??ă½ă??ă?ă?•ă?•ă?®ă??ă?©ă?«ă??ă?¼ă?®ă?ă?¹ă??ă?ă?ă?-
ă?³ă??ă??ă??ă?æ¬ă?®ă??ă?ă?«ă?ă?ă?¼ă?ă?ă?ă?

a? | ã?¼ã?¶ã?¾ã•®ã? | ä??ä?»ä?½ä?, ä.«é?ϕä¿?ã•®ã•ä??ã••ã•?ã•ªè«?æ??ã???arXivã.
 ?ä??æπ?ç'ϕã•?ã• | ä?·è | ·ç'?ä•?ä• | ä••ã• ä•?ä•?ä??

æɱʔç'ɸã•ʔãʔʔè«ʔæʔʔã•æʔʔåɱʔ5ã•ɱãʔʔ

$$a^3 \otimes a \cdot a \cdot a \cdot a \cdot a \cdot a :$$

- è«?æ??ã?¿ã?¤ã??ã?«ã·?ã·¨ã·«ã?³ã?¼ã??ã??ã??ã?¨ã·«ã?¥ã??ã??
- ã?³ã?¼ã??ã??ã??ã??ã?¨ã·@ä,ã·¨ã?·ã¼?è¿°ã·@ã?°ã??ã??ã??ã?©ã?¼ã??ã??ã??ã?·«°?æ?ã
·?ã·?yاملå¼¤å¼·ã·«
- Summaryã·¨Thought_Extractã·¨è|·ç´?æ??ã·?æ?¥æ?-èª?ã·§ã·ªã·?ã·¨ã·?ã·
¨æ?¥æ?-èª?ã·«ç¿»è¨³ã·?ã?·ã??ã?@¹ã·«æ¼·ã??ã·@ã·ªã·?ã??ã·?ã·«ã?
è¤?æ?°è|?è¨?è¿°ã·¨è?½
- Sourceã·@URLã·¨ã¿ã·?ã?¥ã??ã??
- Tagã·¨è?±èª?ã·§
- Idã·¨è«?æ??ç?ªã··ã??ã½¿ã·£ã·|ã?|ã??ã?¼ã?¨ã·ªã?ã?¼ã??ç??æ?·ã·
?ã??ã??ã¼?:arXiv-1111.22222

ä½?è"ã•ªæ??ç« ä??ç??æ?•ã•?ã•?ã•?ä»¥ä, ?ã•?ã?°ã??ã??ã?©ã?¼ã??ã??ã??ã•sã?°ã??ã
•?ã??ã??

- Title: "{è«?æ??ã?¿ã?ªã??ã?«}"
Id: {ã?|ã??ã?¼ã?¬ã•ªIDã??ç??æ?•}
Year: {æ??ç?¿ã¹´}
Authors: [è??è??ã•®ã?ªã?¹ã??]
Source: {è«?æ??ã•®URL}
Tags: [ã?¿ã?°ã?ªã?¹ã??]
Summary:
- "{è|•ç´?}"
Thought_Extract:
- "{ç ?ç®æ?•æ??}"
ZIKUU_Interpretation: ""
Quote: ""

ä½?æ?ã•?ã•?ã??ã?©ã?«ã??ã?¼ã•®ä,ã•sã?•ã??ã?ã?³ã??ã??ã?«è³ªã?•ã??ã ¥ã??ã?ã•é?•ä¿ã•?ã•¾ã
•?ã??

Attention is all you need

ã•?ã??ã•ª??ç?ã•?è¿ã•£ã•|ã•ã•¾ã•?ã??

- Title: "Attention Is All You Need"
Id: arXiv-1706.03762
Year: 2017
Authors: ["Ashish Vaswani", "Noam Shazeer", "Niki Parmar", "Jakob Uszkiewicz"
Source: <https://arxiv.org/abs/1706.03762>
Tags: ["transformer", "self-attention", "NLP", "deep-learning"]
Summary:
- "ã•?ã•®è«?æ??ã¬ã?•ª¼?æ•¥ã•®ã?ªã?«ã¬ã?³ã??ã??ã?¥ã?¼ã?©ã?«ã?•
ã??ã??ã?¬ã?¼ã?¬ã??CNNã??ã?¼ã?¹ã•®ã?•ã?¼ã?±ã?³ã?¹ã?£ã??ã?«ã??ç½®ã••æ?•ã•?ã?•
ã®?ã?¬ã•«æ³´æ?•æ®?æs?ã•«ã?°ã•¥ã••Transformerã?£ã?¼ã?ã??ã?¬ã?ã?£ã?æ••æ¿ã?
?ã??ã??Encoderã¬Decoderã•®ä, ¿æ?¹ã•¬ã??ã?«ã?¬çµ•ã•?ã•?ã??ã?
?Multiâ??Head Attentionã¬æ®?ã•®çµ•ã•?ã?•LayerNormã??ç?¬ã•?ã?•ã½?
ç½®ã?¬ã?³ã?³ã?¼ã??ã?£ã?³ã?°ã•sã½•ç½®æ??ã ±ã??è£?è¶³ã•?ã??ã??"
- "ã®?é¬?ã•sã¬ã?•Googleç¿»è¬³ã•®è?±èª?ã??ã??ã?©ã?³ã?¹èª?ã?¿ã?¹ã?¬ã?
sæ?£ã?ã?£ã??ã?«ã??ã, ?ã??ã•?ã?•ã??ã?©ã?¿ã?¼ã?¿æ?°ã¬ã•»ã•¼ã•?ç?ã•ªã?
?ã??è"ç®?ã?¹ç??ã?¿ãsã¹?ã•?ä, ?ã•?ã•?ã??"
Thought_Extract:
- "èªª±æ³´æ?•æ®?æs?ã¬é?•è•é?çç?, é?£ã??ã?¹ç??ç??ã•«æ?•æ?•ã•?ã?
ä, |ã??è"ç®?ã??ã¬è?½ã•«ã?ã•?ã•?ã?ã?•æ®?æç°ç¿»è¬³ã??é??ã•«é•®ã?¼ã??ã??ã?
?ã??ã•?ã•?ã??"
ZIKUU_Interpretation: ""
Quote: ""

ã??ã?¼ã?¬ã?³ç??æ?é??ã¶ã¬71tok/secã»ã•®ã•sã•?ã??

GPUã•®VRAMã¬16GBã??ã?ã??ã•ª•ã??ã•sã•?ã??

ã•?ã•?ã•?ã•?ã??ã?¼ã?¿ã??ã??ã?¼ã?¿ã??ã?¼ã?¹ã??ã?ã?|ã•?ã•ã¬ã?•è«?æ??ã??æ?çç´£ã?ã??ã?
®ã•?ã•ã•?ã•¶æ¥½ã•«ãªã??ã¾ã•?ã??

