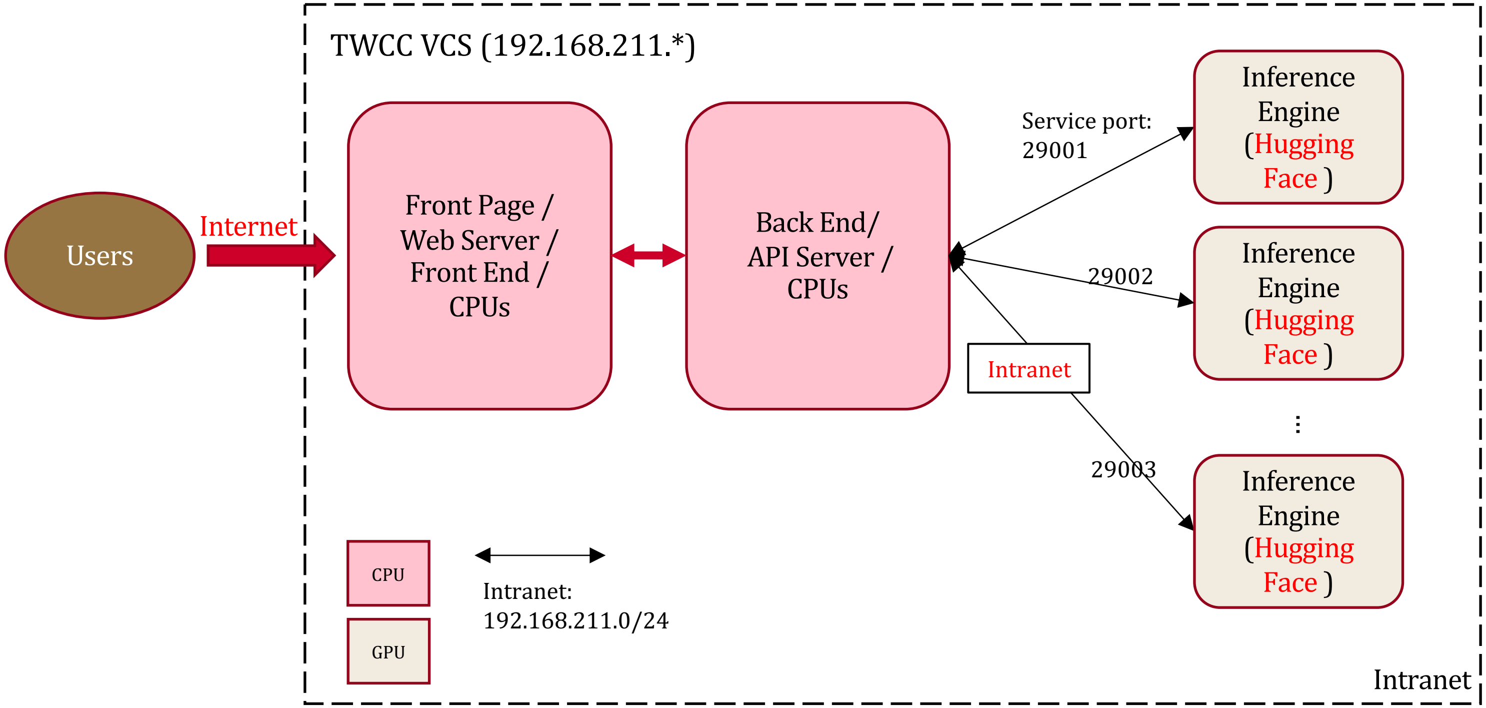
Day 3 –
LLM inference with TensorRT-LLM on NCHC servers

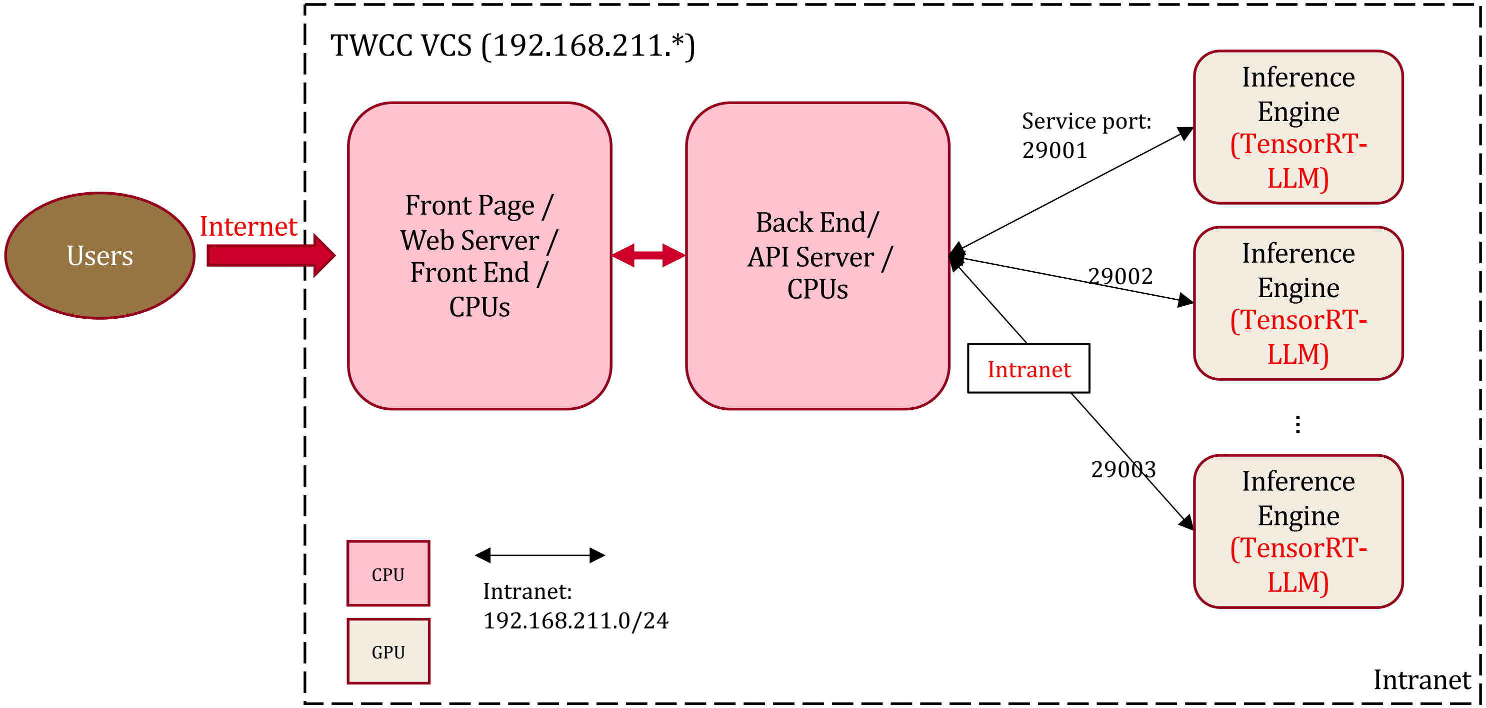
Team member:
Fang-An Kuo, Kuo-Teng Ding, Meng-Chi Huang,
NCHC Speedrunning team

NVIDIA Mentor:
Anthony
Cliff

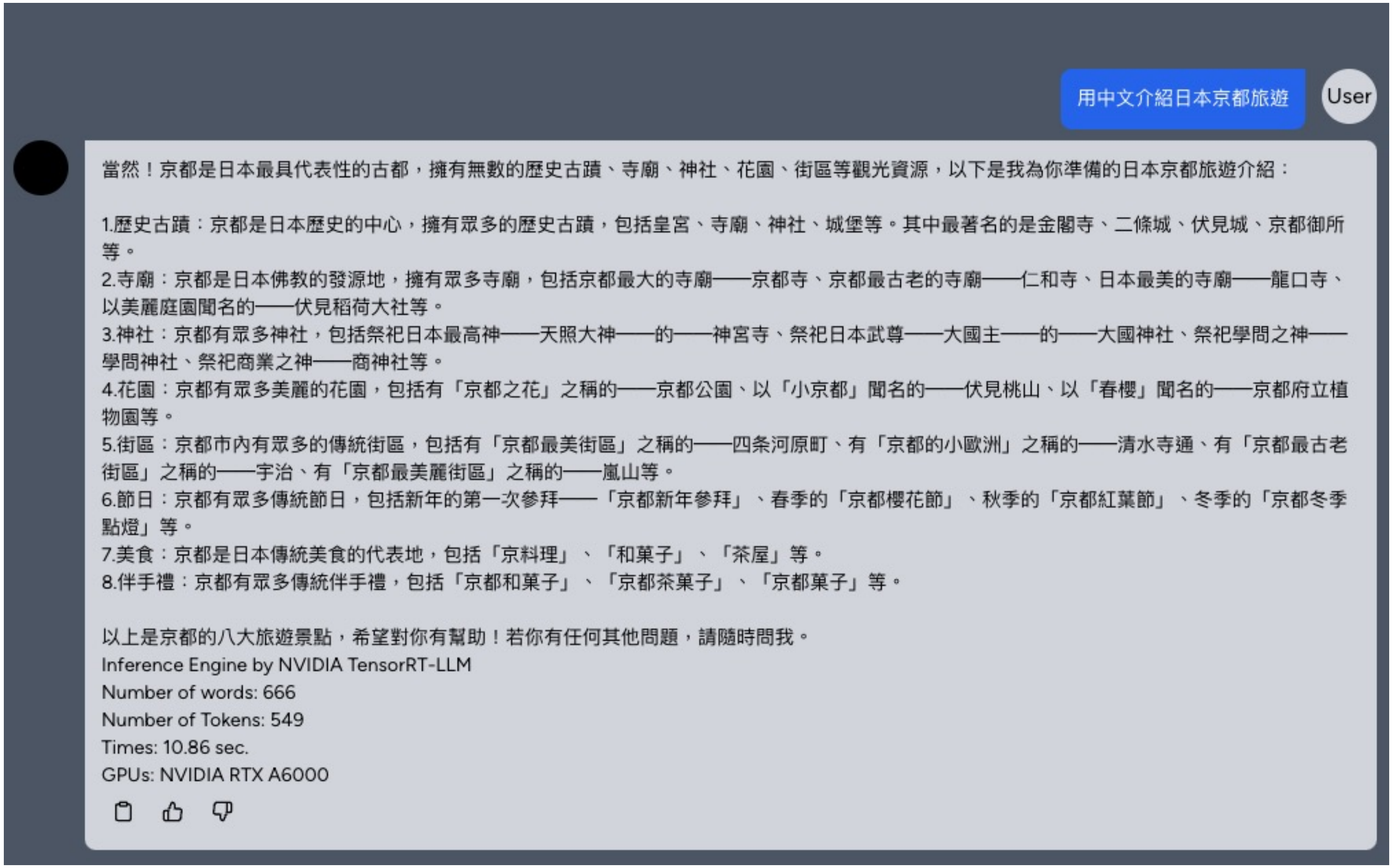
Inference Engine Optimization


TWCC VCS (192.168.211.*)
Users
Internet
Front Page / Web Server / Front End / CPUs
Back End/ API Server / CPUs
Service port: 29001
29002
Intranet
29003
Inference Engine (Hugging Face)
Inference Engine (Hugging Face)
⋮
Inference Engine (Hugging Face)
CPU
GPU
Intranet: 192.168.211.0/24
Intranet

Inference Engine Optimization


TWCC VCS (192.168.211.*)
Users
Internet
Front Page / Web Server / Front End / CPUs
Back End/ API Server / CPUs
Service port: 29001
29002
29003
Intranet
Inference Engine (TensorRT-LLM)
Inference Engine (TensorRT-LLM)
⋮
Inference Engine (TensorRT-LLM)
CPU
GPU
Intranet: 192.168.211.0/24
Intranet

Performance Benchmark on the Demo Website



用中文介紹日本京都旅遊

當然！京都是日本最具代表性的古都，擁有無數的歷史古蹟、寺廟、神社、花園、街區等觀光資源，以下是我為你準備的日本京都旅遊介紹：

1.歷史古蹟：京都是日本歷史的中心，擁有眾多的歷史古蹟，包括皇宮、寺廟、神社、城堡等。其中最著名的是金閣寺、二條城、伏見城、京都御所等。
2.寺廟：京都是日本佛教的發源地，擁有眾多寺廟，包括京都最大的寺廟——京都寺、京都最古老的寺廟——仁和寺、日本最美的寺廟——龍口寺、以美麗庭園聞名的——伏見稻荷大社等。
3.神社：京都有眾多神社，包括祭祀日本最高神——天照大神——的——神宮寺、祭祀日本武尊——大國主——的——大國神社、祭祀學問之神——學問神社、祭祀商業之神——商神社等。
4.花園：京都有眾多美麗的花園，包括有「京都之花」之稱的——京都公園、以「小京都」聞名的——伏見桃山、以「春櫻」聞名的——京都府立植物園等。
5.街區：京都市內有眾多的傳統街區，包括有「京都最美街區」之稱的——四条河原町、有「京都的小歐洲」之稱的——清水寺通、有「京都最古老街區」之稱的——宇治、有「京都最美麗街區」之稱的——嵐山等。
6.節日：京都有眾多傳統節日，包括新年的第一次參拜——「京都新年參拜」、春季的「京都櫻花節」、秋季的「京都紅葉節」、冬季的「京都冬季點燈」等。
7.美食：京都是日本傳統美食的代表地，包括「京料理」、「和菓子」、「茶屋」等。
8.伴手禮：京都有眾多傳統伴手禮，包括「京都和菓子」、「京都茶菓子」、「京都菓子」等。

以上是京都的八大旅遊景點，希望對你有幫助！若你有任何其他問題，請隨時問我。
Inference Engine by NVIDIA TensorRT-LLM
Number of words: 666
Number of Tokens: 549
Times: 10.86 sec.
GPUs: NVIDIA RTX A6000

Performance benchmarking of the both inference engines

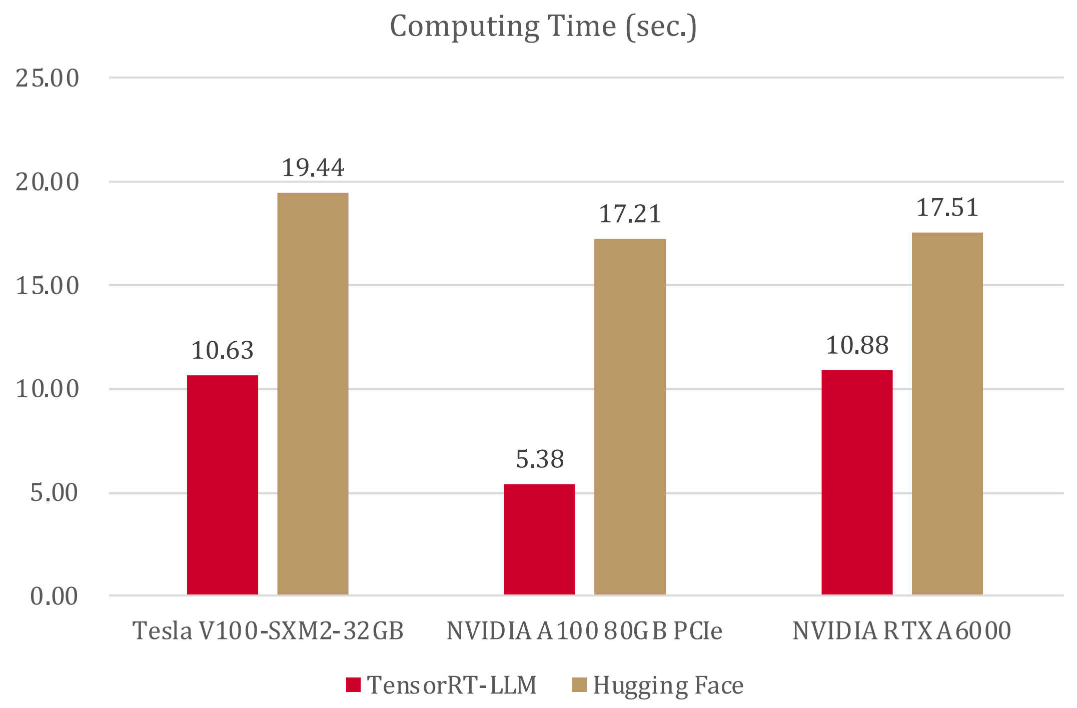
Benchmarks of inferencing Traditional Chinese LLM model on NVIDIA GPUs

- We replace the transformers inference APIs by using TensorRT-LLM APIs
- The benchmarks run on NVIDIA GPUs, including V100/A100/A6000/H100
  - LLM Model: TAIDE Model based on LLaMA2-7B with the version number, b11.
  - Batch size = 1
  - Tensor Parallel = 1 and Pipeline Parallel = 1

Inference APIs	Tokens/sec.			Computing time			Words/sec.		
TAIDE 7B (b11) with FP16	V100	A100	A6000	V100	A100	A6000	V100	A100	A6000
Hugging Face	10.63	5.38	10.88	26.13	29.53	29.01	35.45	40.05	39.35
NVIDIA TRT-LLM	19.44	17.21	17.51	52.30	103.35	51.10	63.59	125.65	62.13
Speedup	1.83	3.20	1.61	2.001	3.500	1.761	1.794	3.138	1.579
Speedup, compared to V100	1.829	3.613	1.787	2.001	3.954	1.955	1.794	3.545	1.753
INT8 model									
Speedup									

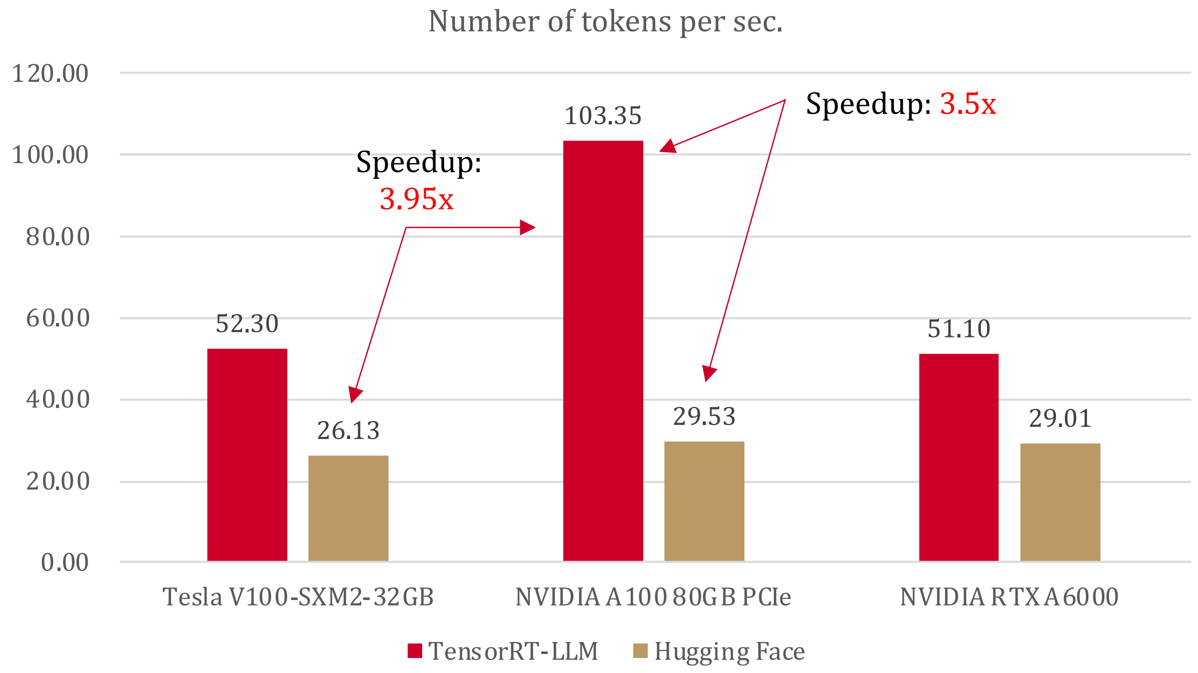
Benchmarks of inferencing Traditional Chinese LLM model on NVIDIA GPUs

- The benchmarks run on NVIDIA GPUs, including V100/A100/A6000/H100
  - LLM Model: TAIDE Model based on LLaMA2-7B with the version number, b11.
  - Batch size = 1
  - Tensor Parallel = 1 and Pipeline Parallel = 1



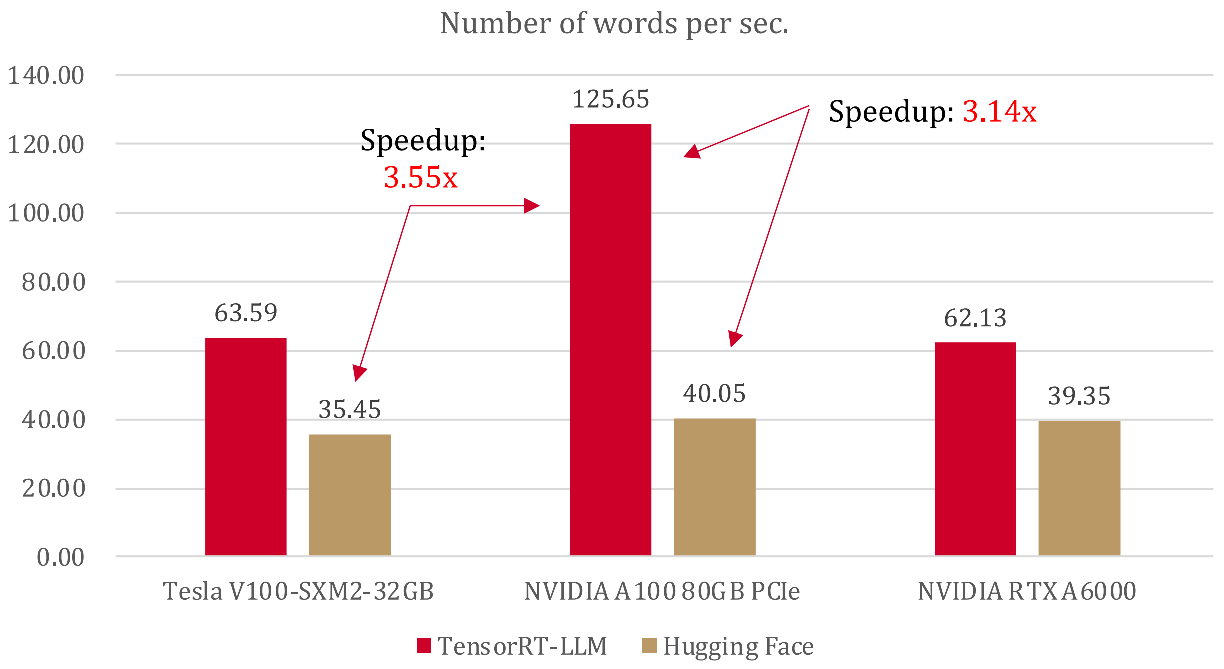
Benchmarks of inferencing Traditional Chinese LLM model on NVIDIA GPUs

- The benchmarks run on NVIDIA GPUs, including V100/A100/A6000/H100
  - LLM Model: TAIDE Model based on LLaMA2-7B with the version number, b11.
  - Batch size = 1
  - Tensor Parallel = 1 and Pipeline Parallel = 1
  - The speedup of generating tokens by using the NVIDIA TensorRT-LLM engines is about 3.5x, which is based on NVIDIA A100

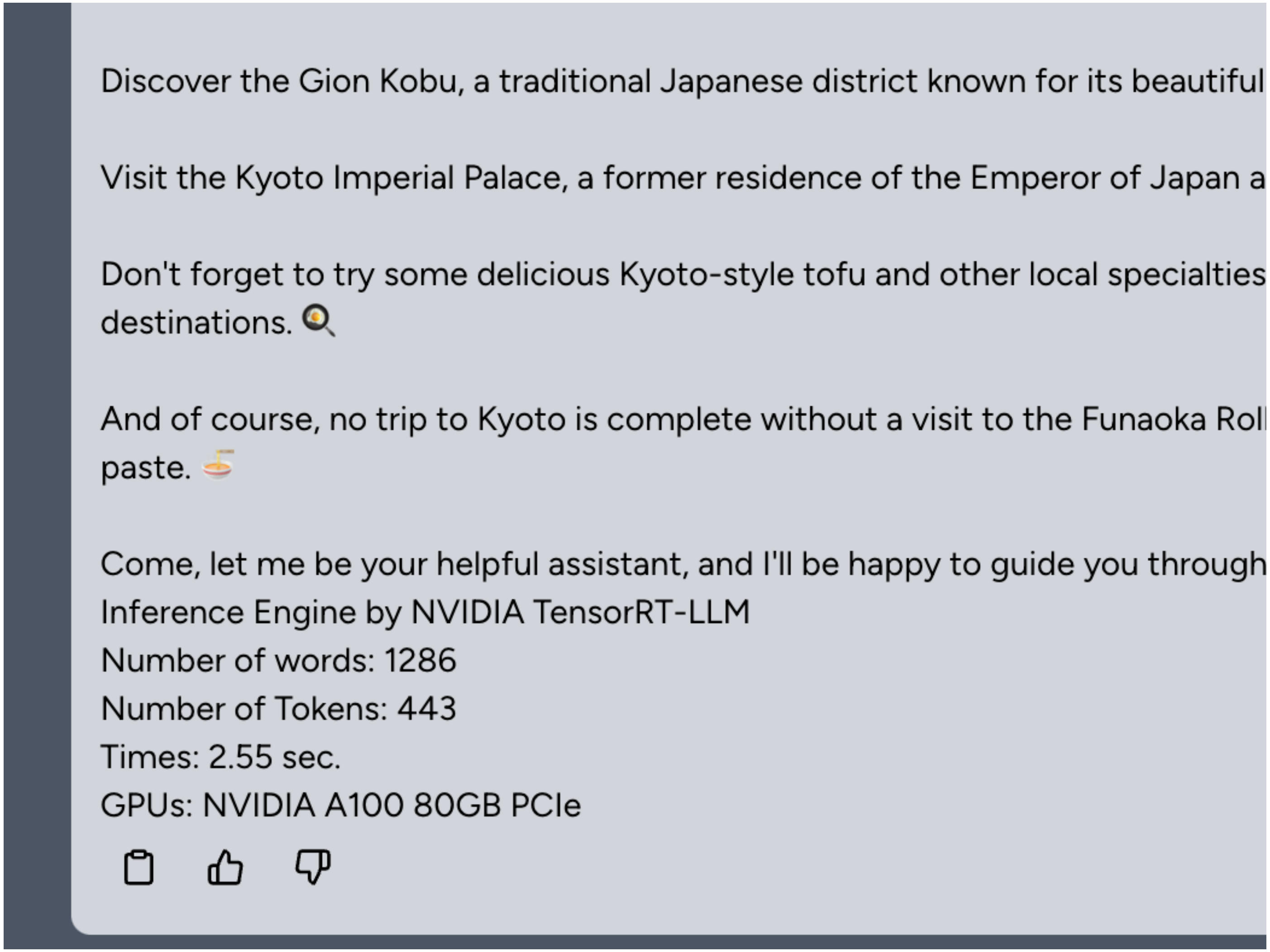


Benchmarks of inferencing Traditional Chinese LLM model on NVIDIA GPUs

- The benchmarks run on NVIDIA GPUs, including V100/A100/A6000/H100
  - LLM Model: TAIDE Model based on LLaMA2-7B with the version number, b11.
  - Batch size = 1
  - Tensor Parallel = 1 and Pipeline Parallel = 1
  - The speedup of generating words by using the NVIDIA TensorRT-LLM engines is about 3.14x, which is based on NVIDIA A100



INT8 model



Discover the Gion Kobu, a traditional Japanese district known for its beautiful

Visit the Kyoto Imperial Palace, a former residence of the Emperor of Japan a

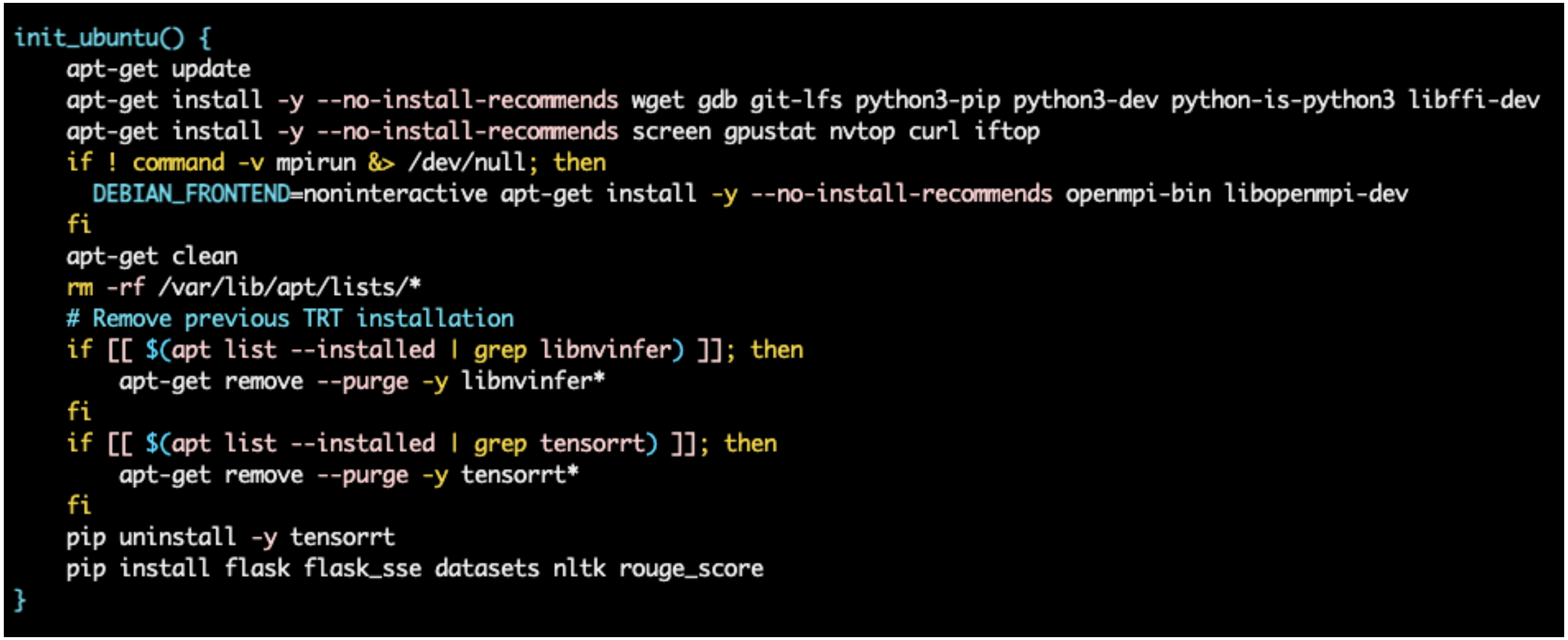
Don't forget to try some delicious Kyoto-style tofu and other local specialties destinations. 🍳

And of course, no trip to Kyoto is complete without a visit to the Funaoka Roll paste. 🍜

Come, let me be your helpful assistant, and I'll be happy to guide you through
Inference Engine by NVIDIA TensorRT-LLM
Number of words: 1286
Number of Tokens: 443
Times: 2.55 sec.
GPUs: NVIDIA A100 80GB PCIe

Problems and Solutions

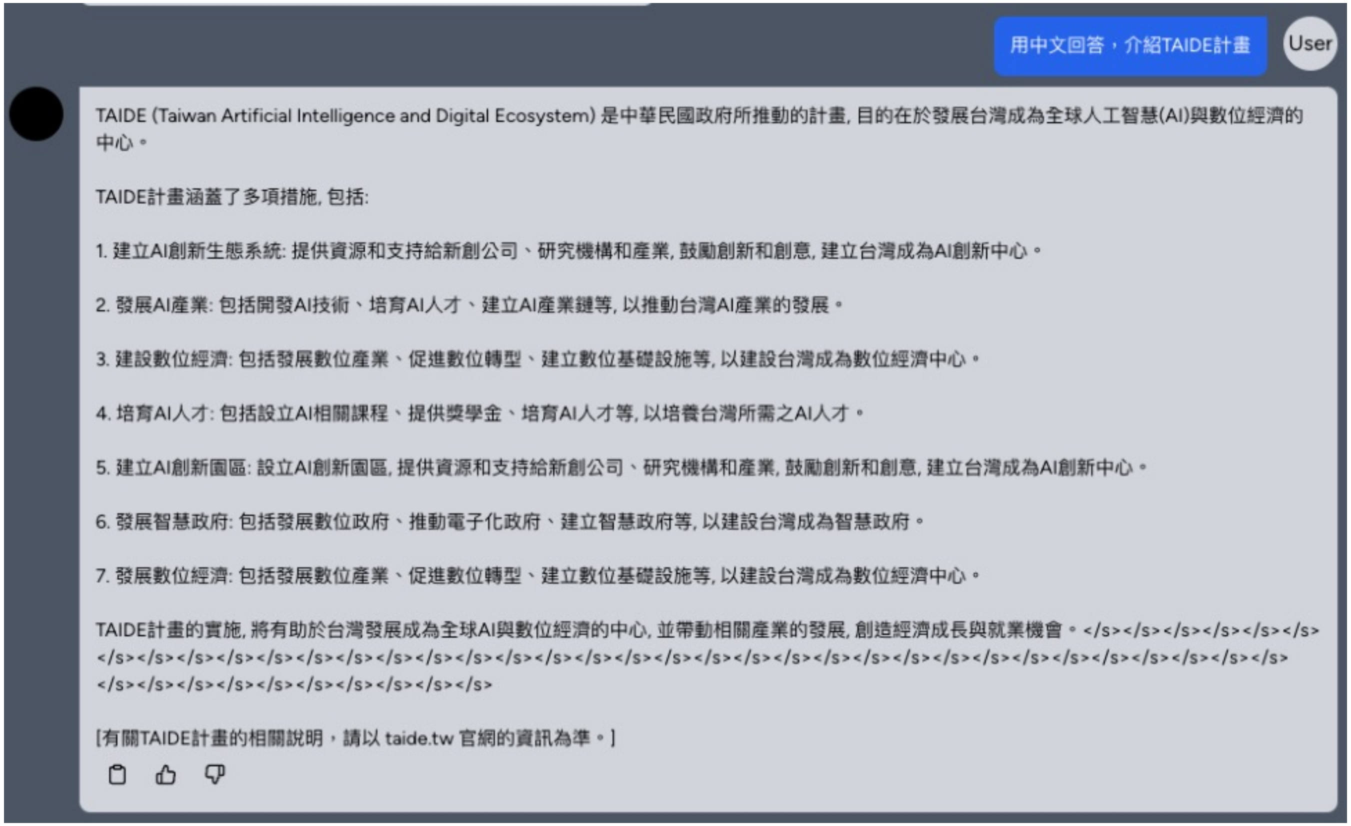
- Modify {TensorRT-LLM}/docker/common/install_base.sh
  - OS: ubuntu 22.04



```
init_ubuntu() {
    apt-get update
    apt-get install -y --no-install-recommends wget gdb git-lfs python3-pip python3-dev python-is-python3 libffi-dev
    apt-get install -y --no-install-recommends screen gpustat nvtop curl iftop
    if ! command -v mpirun &> /dev/null; then
      DEBIAN_FRONTEND=noninteractive apt-get install -y --no-install-recommends openmpi-bin libopenmpi-dev
    fi
    apt-get clean
    rm -rf /var/lib/apt/lists/*
    # Remove previous TRT installation
    if [[ $(apt list --installed | grep libnvinfer) ]]; then
        apt-get remove --purge -y libnvinfer*
    fi
    if [[ $(apt list --installed | grep tensorrt) ]]; then
        apt-get remove --purge -y tensorrt*
    fi
    pip uninstall -y tensorrt
    pip install flask flask_sse datasets nltk rouge_score
}
```

Problems and Solutions

- Generation cannot properly stops at EOS (</s>) and only stops at max output length
  - Solution: output_text.replace("</s>","")
  - Another solution and reference:
    - https://github.com/NVIDIA/TensorRT-LLM/blob/a21e2f85178111fed9812bb88c2cc7411b25f0ba/examples/gpt/run.py#L299



用中文回答，介紹TAIDE計畫

TAIDE (Taiwan Artificial Intelligence and Digital Ecosystem) 是中華民國政府所推動的計畫, 目的在於發展台灣成為全球人工智慧(AI)與數位經濟的中心。

TAIDE計畫涵蓋了多項措施, 包括:

1. 建立AI創新生態系統: 提供資源和支持給新創公司、研究機構和產業, 鼓勵創新和創意, 建立台灣成為AI創新中心。

2. 發展AI產業: 包括開發AI技術、培育AI人才、建立AI產業鏈等, 以推動台灣AI產業的發展。

3. 建設數位經濟: 包括發展數位產業、促進數位轉型、建立數位基礎設施等, 以建設台灣成為數位經濟中心。

4. 培育AI人才: 包括設立AI相關課程、提供獎學金、培育AI人才等, 以培養台灣所需之AI人才。

5. 建立AI創新園區: 設立AI創新園區, 提供資源和支持給新創公司、研究機構和產業, 鼓勵創新和創意, 建立台灣成為AI創新中心。

6. 發展智慧政府: 包括發展數位政府、推動電子化政府、建立智慧政府等, 以建設台灣成為智慧政府。

7. 發展數位經濟: 包括發展數位產業、促進數位轉型、建立數位基礎設施等, 以建設台灣成為數位經濟中心。

TAIDE計畫的實施, 將有助於台灣發展成為全球AI與數位經濟的中心, 並帶動相關產業的發展, 創造經濟成長與就業機會。</s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s>

[有關TAIDE計畫的相關說明，請以 taide.tw 官網的資訊為準。]

рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
kiitos
dankie
dhanyavad
nanni
nandri
thank you
хвала
asante
manana
obrigada
tenki
bedankt
bayarlalaa
gràcie
hvala
maururu
köszönöm
akun
dankon
aciú
djiere dieuf
tau
mochchakkeram
murakoze
enkosi
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
didi madloba
najis tuke
terima kasih
rahmet
shukriya
мерси
kam sah hamnida
rahmat
diolch
dhanyavadagalu
mercé
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci