


LAURENT LE BO



Baudel-A/r



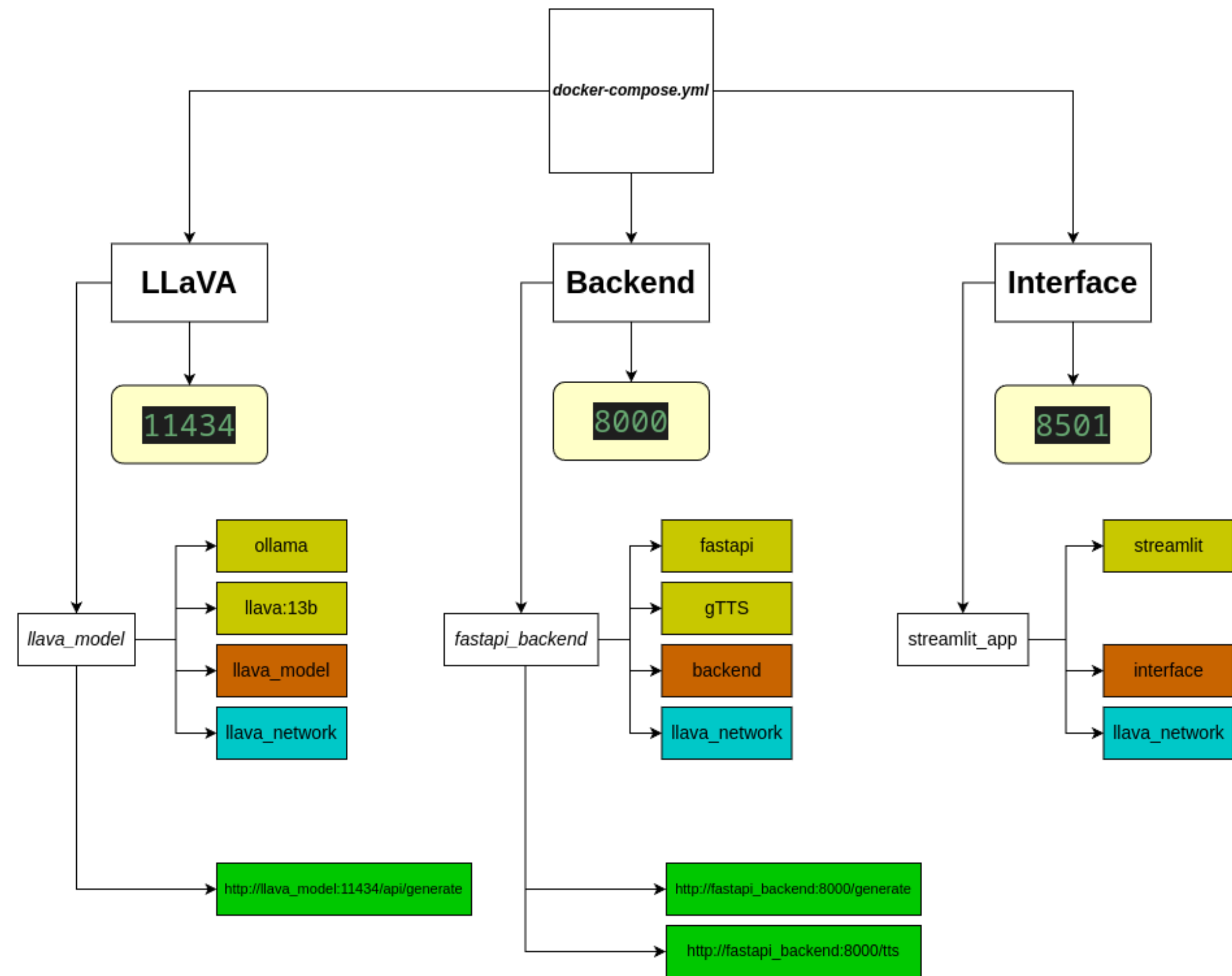
YOUR PERSONNAL ASSISTANT FOR
LOVE LETTER GENERATION



18/12/2024

CLOVIS LECHIEN

LAURENT LE BO



Légende :

ports

features

volumes

networks

api routes

used now:

LLaVA

*Large Language and
Vision Assistant*

- Multimodal LLM (**MLLM**)
- LLaVA:13b
 - Size: ~**8 GiB**
 - LLM : **LLaMA v2**
 - quantization : **Q4_0**
 - Projector: **CLIP**
 - quantization : **F16**
- ollama

before :

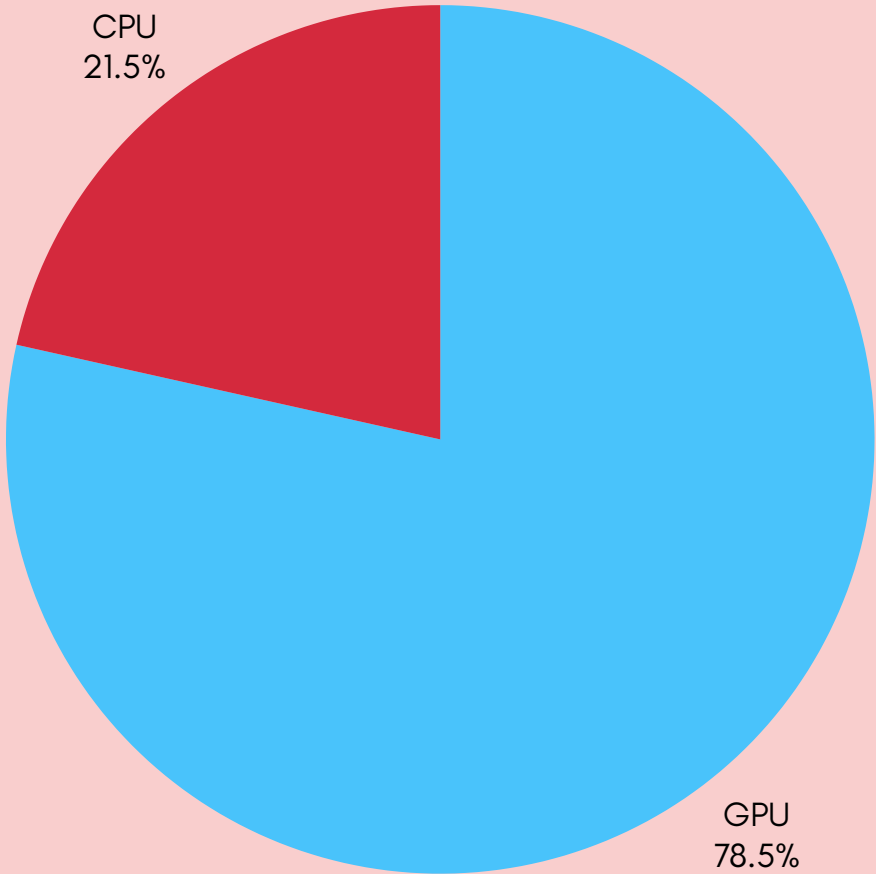
GPT-2 + BLIP2

fine-tuning (legacy)

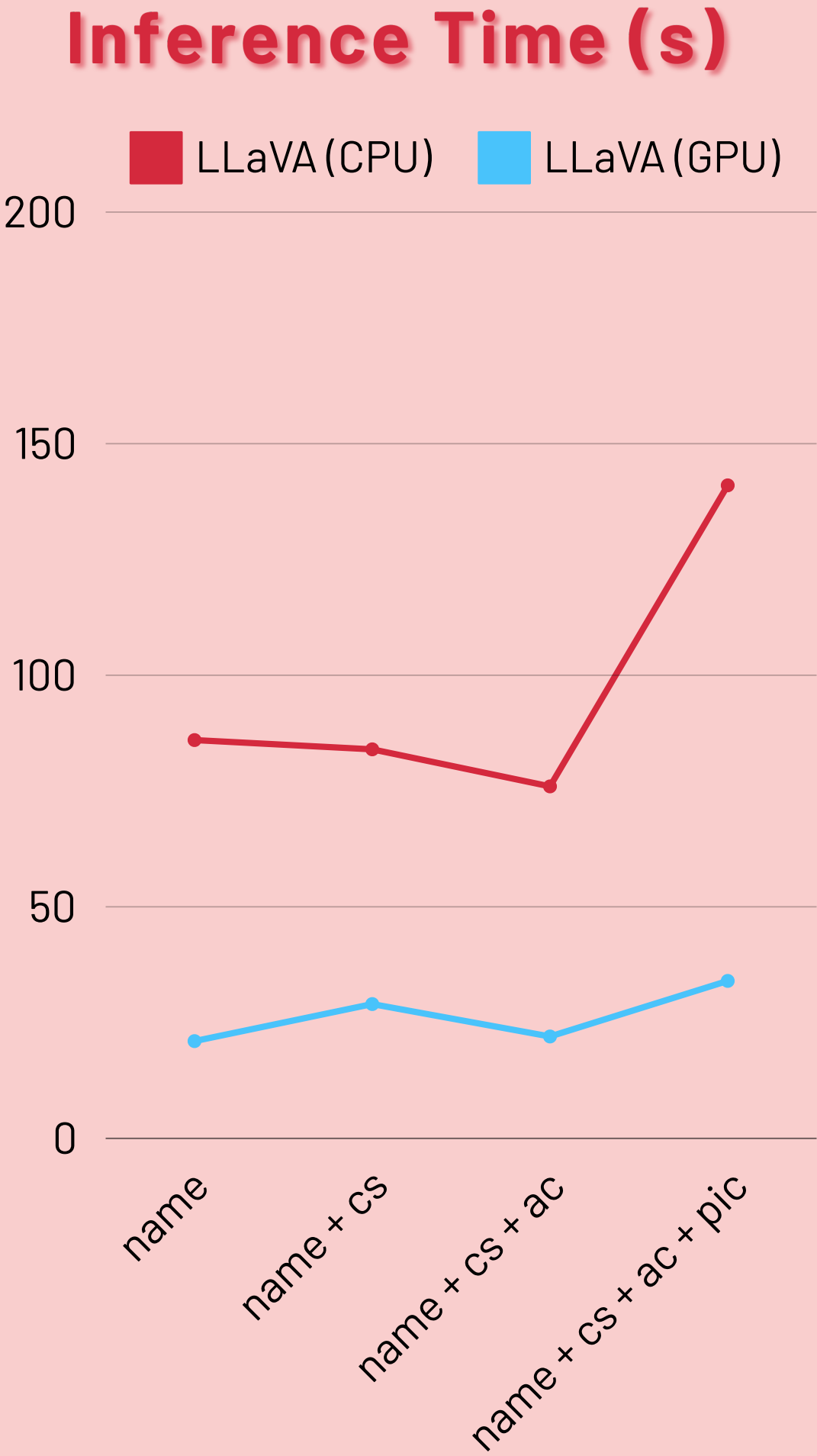
- Large Language Model (**MLLM**)
- Gpt2-medium:380M
 - Size:
 - pre-finetune: ~**1.4 GiB**
 - post-finetune: ~**7 GiB**
 - quantization : **F32**
- Blip2-Salesforce:3.74B
 - Size: ~**14 GiB**
 - quantization : **F32**
- huggingface

	CPU Usage (%)	CPU Memory Usage (GiB)	GPU Usage (%)	GPU Memory Usage (GiB)	Inference Time [cumsum](s)
LLaVA CPU	~90.1%	~27.6 GiB	/	/	387s
LLaVA GPU	~4.07%	~1.25 GiB	~87.5%	~7.17 GiB	106s

- GPU inference **~4x faster** than CPU
- consistent** inference times on GPU



	LLaVA (CPU)	LLaVA (GPU)
name	86s	21s
name + cs	84s	29s
name + cs + ac	76s	22s
name + cs + ac + pic	141s	34s





Thanks :)



CODE AVAILABLE AT :

 **ClovisDyArx/baudel-ai.r**

