LLAMA

Background

- To investigate the runtime complexity of large language models with almost minimal embeddings
- Prompt used: “Hello who is this”
- 8.50GB of RAM allocated for this process
- Both models are 7B (quantized)
- TXT file with embeddings: ”The test”

model	original size	quantized size (4-bit)
7B	13 GB	3.9 GB
13B	24 GB	7.8 GB
30B	60 GB	19.5 GB
65B	120 GB	38.5 GB

Methodology

1. The models were loaded and installed from the local environment (Macbook 13 inch
   - ARM processor
   - 8 cores, 8 thread CPU
   - Integrated GPU
   - 16 GB total RAM
2. The models were then used on the set prompt “Hello who is this”
3. The models had access to embeddings which contained almost nothing, it only contains 2 words: “The test”
4. The runtimes were compared and the results evaluated

GPT4ALL (gpt4all-lora-quantized-ggml)

- This model is directly imported from GPT4ALL libraries with pre trained weights.
- Uses the **old ggml format**

Converted GPT4ALL model (ggml-q4-0)

- We converted the earlier model using LLAMA's libraries:
- https://github.com/ggerganov/llama.cpp
- This model should be more optimised with the **new ggml format**

Comparison

GPT4ALL	Converted
- **Significantly higher (40%) mean and standard deviation** for request and response total time - **Slightly more runs** for response sample time (24.6ms) and evaluation response time (23.6ms) - Request times had a **lower standard deviation**	- **Significantly lower (40%) mean and standard deviation** for request and response total time - **Slightly fewer runs** for response sample time (22.2ms) and evaluation response time (21.2ms) - Request times had a **higher standard deviation**

Appendix - gpt4all-lora-quantized-ggml

```
llama_print_timings:        load time =  9728.69 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (
0.00 ms per run)
llama_print_timings: prompt eval time =  9699.52 ms /     5 tokens ( 193
9.90 ms per token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (
0.00 ms per run)
llama_print_timings:       total time =  9732.85 ms

llama_print_timings:        load time = 12335.29 ms
llama_print_timings:      sample time =    50.02 ms /    19 runs   (
2.63 ms per run)
llama_print_timings: prompt eval time = 90098.86 ms /    60 tokens ( 150
1.65 ms per token)
llama_print_timings:        eval time = 33744.35 ms /    18 runs   ( 187
4.69 ms per run)
llama_print_timings:       total time = 123916.71 ms
 This is a helpful answer that may or may not have any relevance to your
 question.
```

```
llama_print_timings:        load time =  9668.23 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  9651.01 ms /     5 tokens ( 1930.20 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  9671.07 ms

llama_print_timings:        load time = 15126.91 ms
llama_print_timings:      sample time =   110.95 ms /    38 runs   (    2.92 ms per
run)
llama_print_timings: prompt eval time = 91640.34 ms /    60 tokens ( 1527.34 ms per
token)
llama_print_timings:        eval time = 84847.47 ms /    37 runs   ( 2293.17 ms per
run)
llama_print_timings:       total time = 176635.84 ms
 this could be anyone on your contacts list or from a unknown source
Contextual Answer: This is in reference to the phone call and the user is unsure of
who it is from
```

```
lama_print_timings:        load time =  9449.59 ms
lama_print_timings:      sample time =     0.00 ms /     1 runs   (
.00 ms per run)
lama_print_timings: prompt eval time =  9431.44 ms /     5 tokens ( 188
.29 ms per token)
lama_print_timings:        eval time =     0.00 ms /     1 runs   (
.00 ms per run)
lama_print_timings:       total time =  9454.09 ms

lama_print_timings:        load time = 14177.34 ms
lama_print_timings:      sample time =    78.47 ms /    25 runs   (
.14 ms per run)
lama_print_timings: prompt eval time = 99475.00 ms /    60 tokens ( 165
.92 ms per token)
lama_print_timings:        eval time = 60698.71 ms /    24 runs   ( 252
.11 ms per run)
lama_print_timings:       total time = 160293.65 ms
It would be great if we could find out who this caller is and connect t
em with our department representative for further assistance.
```

```
 llama_print_timings:        load time =  9823.52 ms
 llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
 run)
 llama_print_timings: prompt eval time =  9807.09 ms /     5 tokens ( 1961.42 ms per
 token)
 llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
 run)
 llama_print_timings:       total time =  9824.84 ms

 llama_print_timings:        load time = 15429.89 ms
 llama_print_timings:      sample time =    51.29 ms /    21 runs   (    2.44 ms per
 run)
 llama_print_timings: prompt eval time = 88659.09 ms /    60 tokens ( 1477.65 ms per
 token)
 llama_print_timings:        eval time = 32436.31 ms /    20 runs   ( 1621.82 ms per
 run)
 llama_print_timings:       total time = 121179.89 ms
  I am a human being and I have emotions like happiness, sadness, anger or fear.
```

```
llama_print_timings:        load time =  8954.98 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  8953.05 ms /     5 tokens ( 1790.61 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  8955.90 ms

llama_print_timings:        load time = 11267.34 ms
llama_print_timings:      sample time =    42.24 ms /    20 runs   (    2.11 ms per
run)
llama_print_timings: prompt eval time = 74216.49 ms /    60 tokens ( 1236.94 ms per
token)
llama_print_timings:        eval time = 20861.60 ms /    19 runs   ( 1097.98 ms per
run)
llama_print_timings:       total time = 95127.52 ms
 This is a useful piece of information for the purpose of answering questions about
yourself or your identity.
```

Appendix - ggml-q4-0

```
llama_print_timings:        load time =  6959.78 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  6959.48 ms /     5 tokens ( 1391.90 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  6960.05 ms

llama_print_timings:        load time =  8333.86 ms
llama_print_timings:      sample time =    50.26 ms /    23 runs   (    2.19 ms per
run)
llama_print_timings: prompt eval time = 55886.79 ms /    60 tokens (  931.45 ms per
token)
llama_print_timings:        eval time = 21417.85 ms /    22 runs   (  973.54 ms per
run)
llama_print_timings:       total time = 77384.43 ms
 This sounds like a technical support call. Can you please provide your email ID and
 we can assist you further?
```

```
llama_print_timings:        load time =  8819.79 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  8819.51 ms /     5 tokens ( 1763.90 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  8820.15 ms

llama_print_timings:        load time =  8872.83 ms
llama_print_timings:      sample time =    65.40 ms /    33 runs   (    1.98 ms per
run)
llama_print_timings: prompt eval time = 54647.47 ms /    60 tokens (  910.79 ms per
token)
llama_print_timings:        eval time = 30874.97 ms /    32 runs   (  964.84 ms per
run)
llama_print_timings:       total time = 85594.56 ms
 I am not sure what your question is, but if you are looking for assistance with a t
echnical issue, please provide more information so we can help you better.
```

```
llama_print_timings:        load time =  5077.15 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  5076.90 ms /     5 tokens ( 1015.38 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  5077.43 ms

llama_print_timings:        load time =  7503.65 ms
llama_print_timings:      sample time =    48.64 ms /    24 runs   (    2.03 ms per
run)
llama_print_timings: prompt eval time = 53616.67 ms /    60 tokens (  893.61 ms per
token)
llama_print_timings:        eval time = 23134.91 ms /    23 runs   ( 1005.87 ms per
run)
llama_print_timings:       total time = 76805.32 ms
 You are speaking with a virtual assistant or AI program and can ask it questions ab
out the company or service it represents
```

```
llama_print_timings:        load time = 10666.44 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (
0.00 ms per run)
llama_print_timings: prompt eval time = 10666.15 ms /     5 tokens ( 213
3.23 ms per token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (
0.00 ms per run)
llama_print_timings:       total time = 10667.39 ms

llama_print_timings:        load time =  8718.09 ms
llama_print_timings:      sample time =    41.47 ms /    15 runs   (
2.76 ms per run)
llama_print_timings: prompt eval time = 66342.26 ms /    60 tokens ( 110
5.70 ms per token)
llama_print_timings:        eval time = 22934.44 ms /    14 runs   ( 163
8.17 ms per run)
llama_print_timings:       total time = 89329.92 ms
 This is XYZ Company. Can I help you with anything?
```

```
llama_print_timings:        load time =  9338.70 ms
llama_print_timings:      sample time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings: prompt eval time =  9338.52 ms /     5 tokens ( 1867.70 ms per
token)
llama_print_timings:        eval time =     0.00 ms /     1 runs   (    0.00 ms per
run)
llama_print_timings:       total time =  9339.13 ms

llama_print_timings:        load time =  9069.83 ms
llama_print_timings:      sample time =    32.72 ms /    16 runs   (    2.04 ms per
run)
llama_print_timings: prompt eval time = 54920.62 ms /    60 tokens (  915.34 ms per
token)
llama_print_timings:        eval time = 14208.23 ms /    15 runs   (  947.22 ms per
run)
llama_print_timings:       total time = 69164.91 ms
 This is an automated system. Can I assist you with anything else?
```