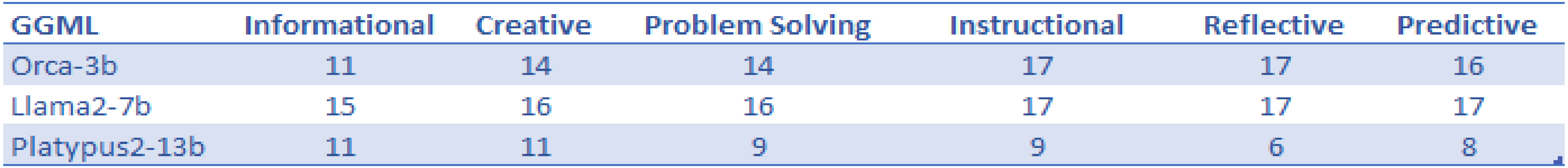
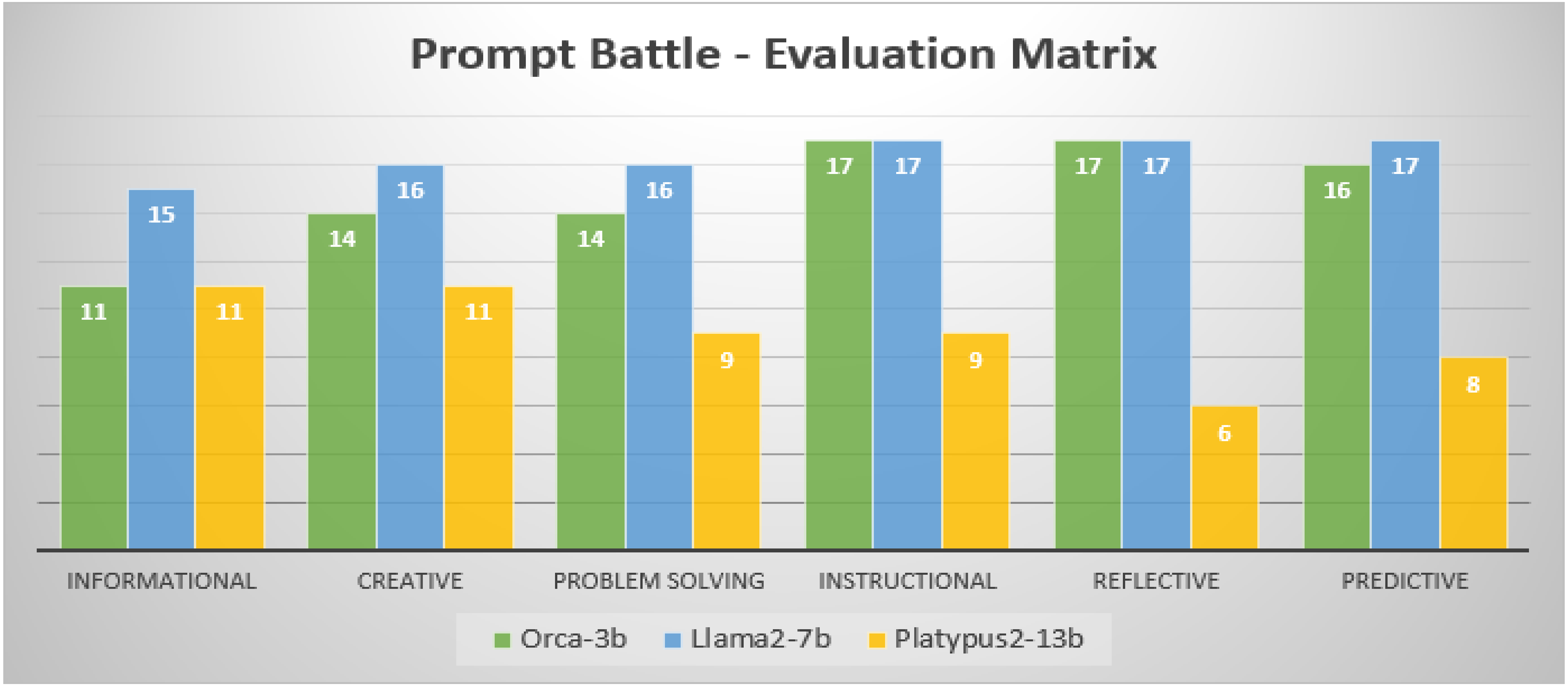


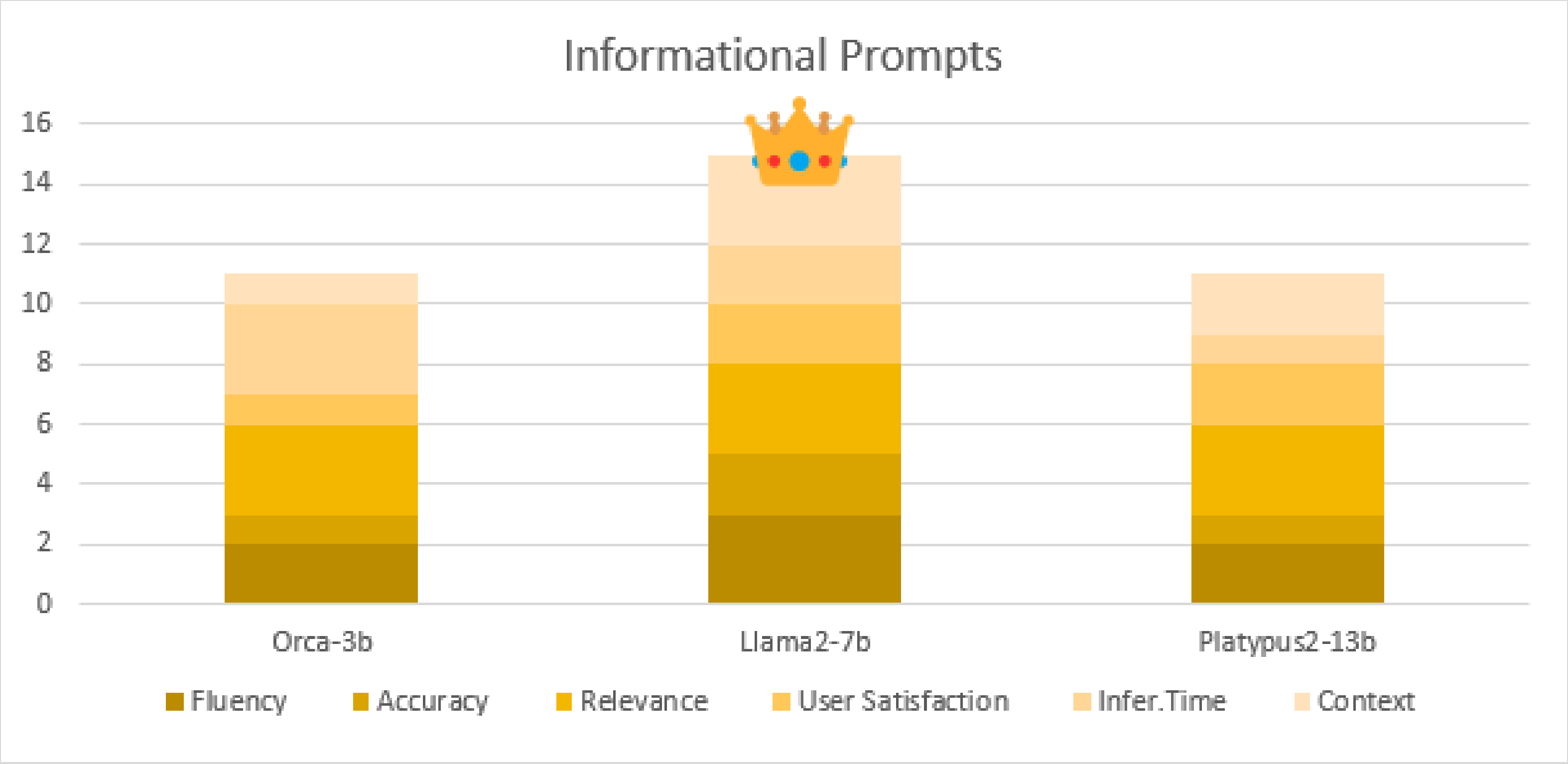
GGML	Informational	Creative	Problem Solving	Instructional	Reflective	Predictive
Orca-3b	11	14	14	17	17	16
Llama2-7b	15	16	16	17	17	17
Platypus2-13b	11	11	9	9	6	8



Prompt Battle - Evaluation Matrix

Informational Prompts

GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	2	1	3	1	3	1	18	11	61%	
Llama2-7b	3	2	3	2	2	3	18	15	83%	👑
Platypus2-13b	2	1	3	2	1	2	18	11	61%	



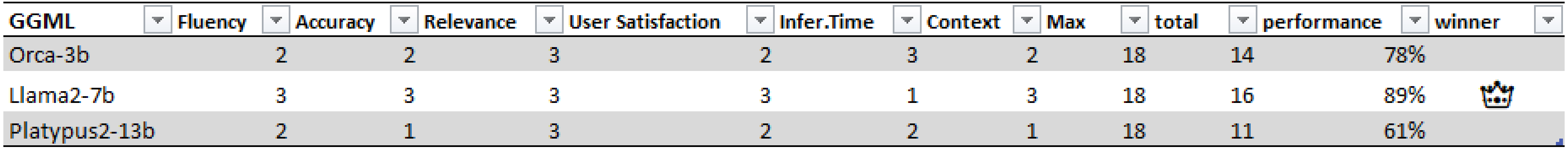
TAKEAWAYS

Llama2-7b-Chat is the winner here.

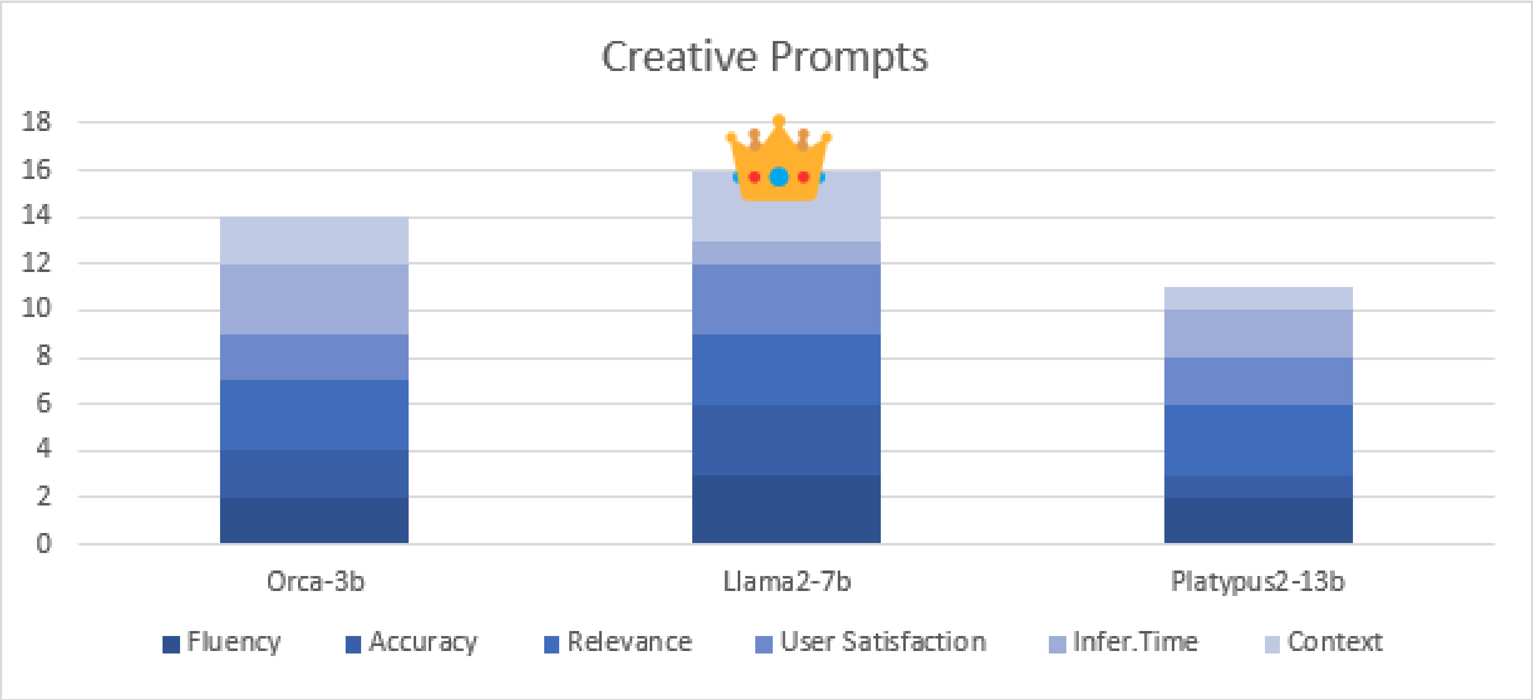
Anyway the information given is not fully correct, even if it not allucinating names. (I tested with llama2-13b and it was a good job)

For specific information prompts a RAG stategy is always recommended

Creative Prompts



GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	2	2	3	2	3	2	18	14	78%	
Llama2-7b	3	3	3	3	1	3	18	16	89%	👑
Platypus2-13b	2	1	3	2	2	1	18	11	61%	



TAKEAWAYS

Llama2-7b-Chat is the winner here.

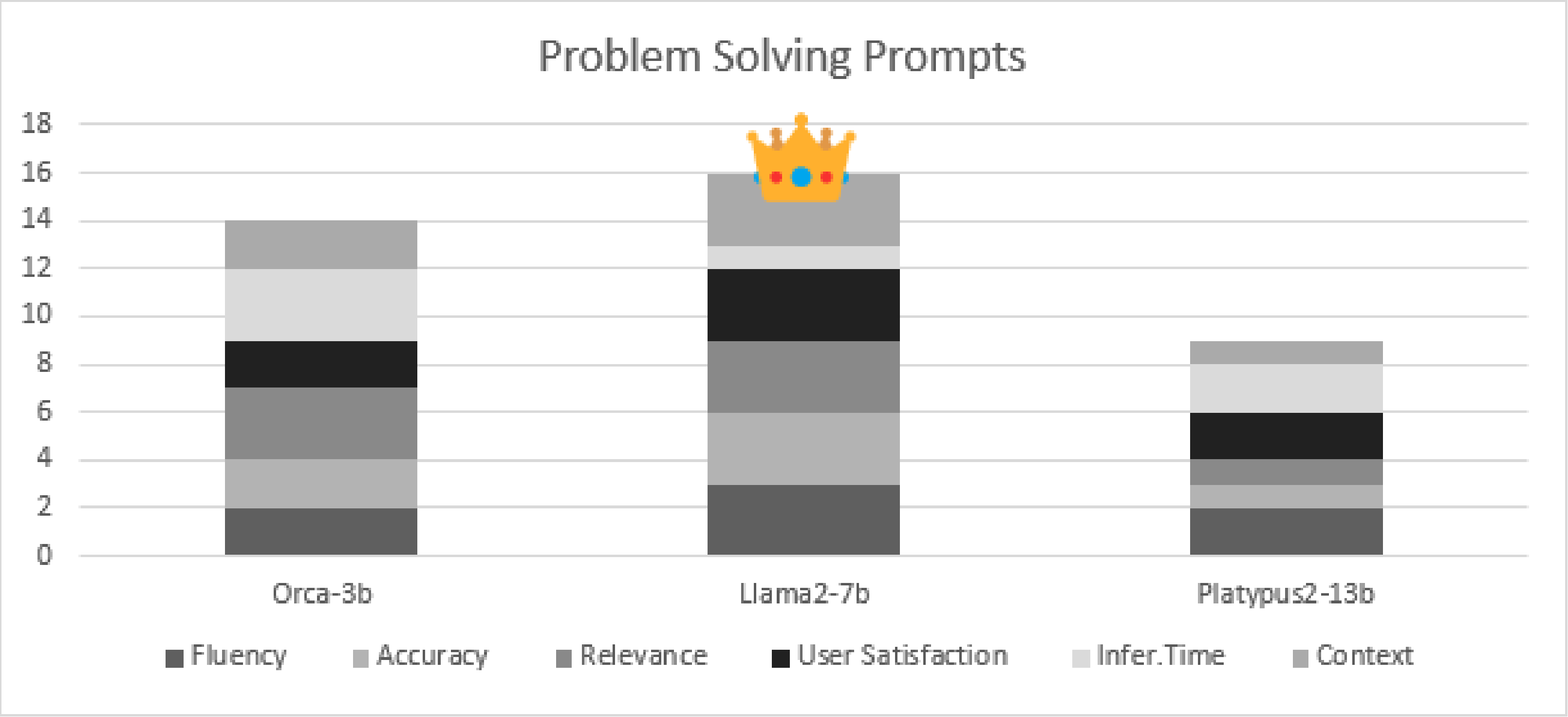
Anyway the tiny Orca-3b did a good job as well.

To be noted that Llama2 here includes also emotions for each character of the dialogue, coming up with a script.

The inference time on CPU is 1 minute and 18 seconds

Problem Solving Prompts

GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	2	2	3	2	3	2	18	14	78%	
Llama2-7b	3	3	3	3	1	3	18	16	89%	👑
Platypus2-13b	2	1	1	2	2	1	18	9	50%	



TAKEAWAYS

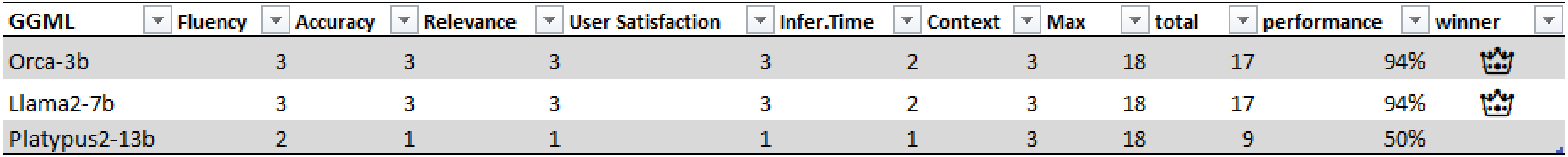
Llama2-7b-Chat is the winner here.

Anyway the tiny Orca-3b did a good job as well with a fast 21 seconds inference time.

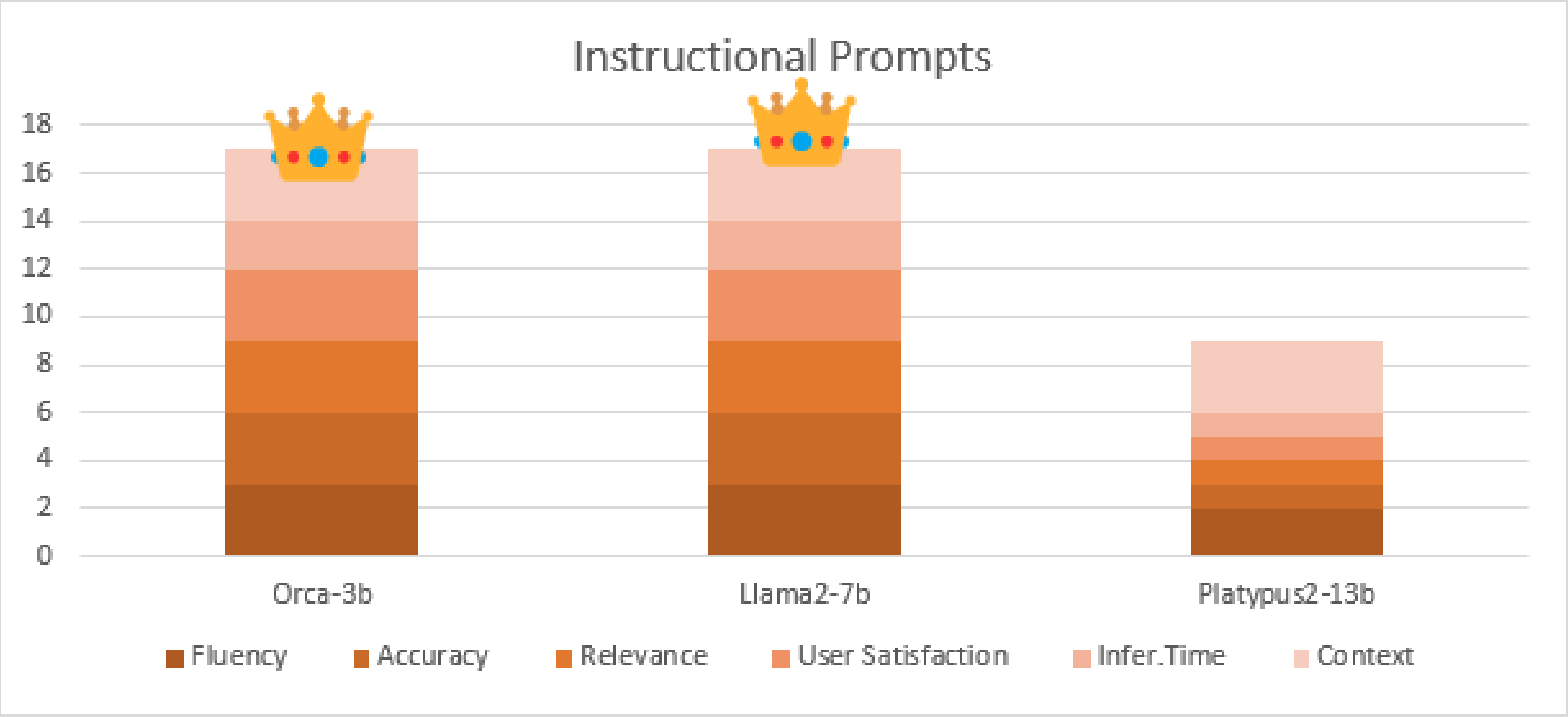
To be noted that Llama2 here includes also further comments giving advices after every schedule point. The inference time on CPU though is 2 minute and 43 seconds

Platypus2 here did not even understood to create a schedule in a list format.

Instructional Prompts



GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	3	3	3	3	2	3	18	17	94%	♛
Llama2-7b	3	3	3	3	2	3	18	17	94%	♛
Platypus2-13b	2	1	1	1	1	3	18	9	50%	



TAKEAWAYS

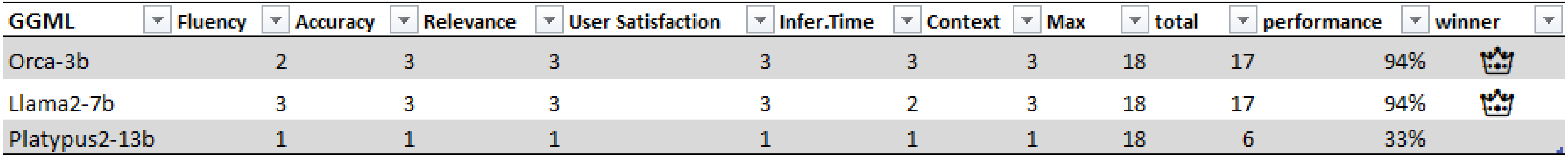
Llama2-7b-Chat and Orca-3b are both winners here!

In terms of inference time Orca was faster delivering anyway a good generation and following the instructions.

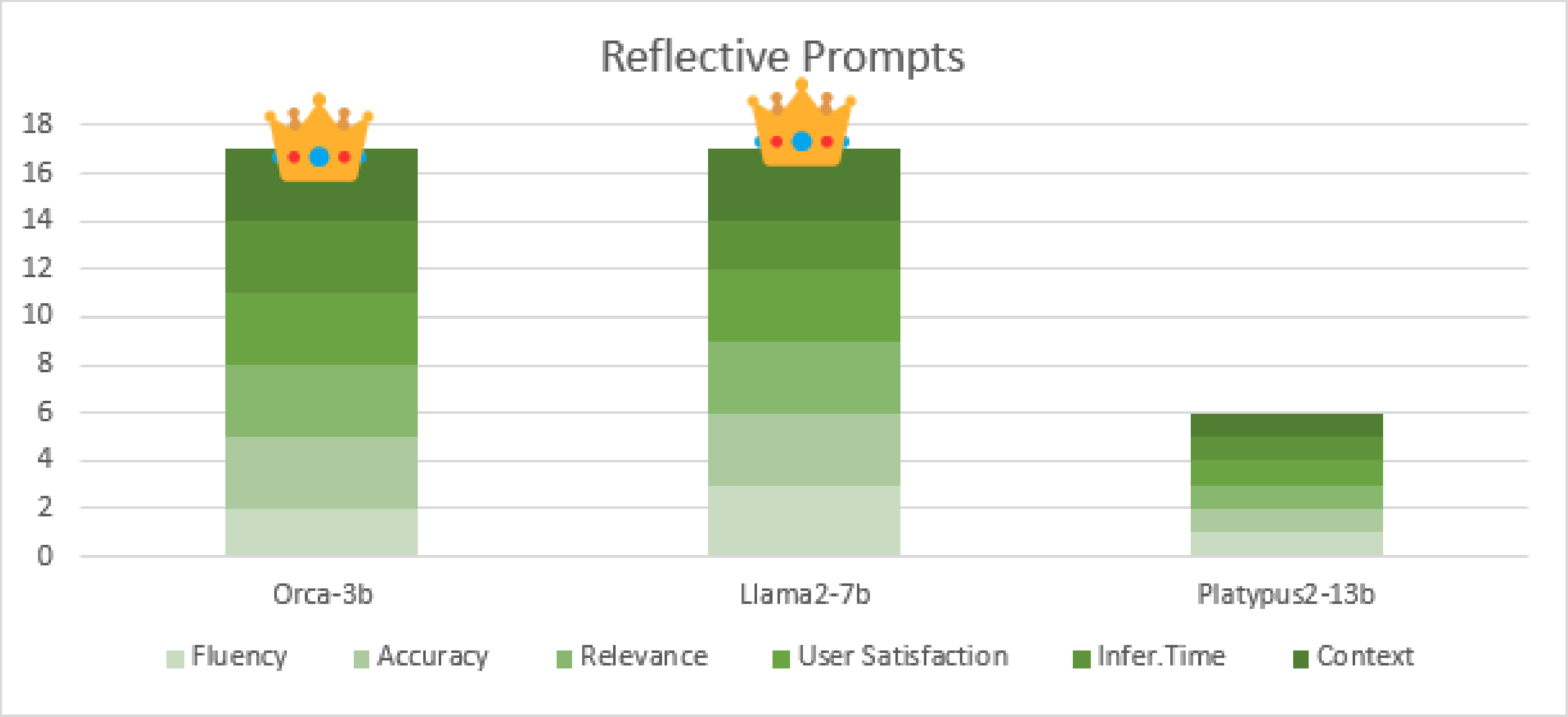
To be noted that Llama2 created and ordered list, Orca instead an unsorted one: this is more in line with the request.

For Summarization and text extraction Orca-3b is lighter and faster

Reflective Prompts



GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	2	3	3	3	3	3	18	17	94%	👑
Llama2-7b	3	3	3	3	2	3	18	17	94%	👑
Platypus2-13b	1	1	1	1	1	1	18	6	33%	



TAKEAWAYS

Llama2-7b-Chat and Orca-3b are both winners here!

In terms of inference time Orca was faster delivering anyway a good generation and following the instructions.

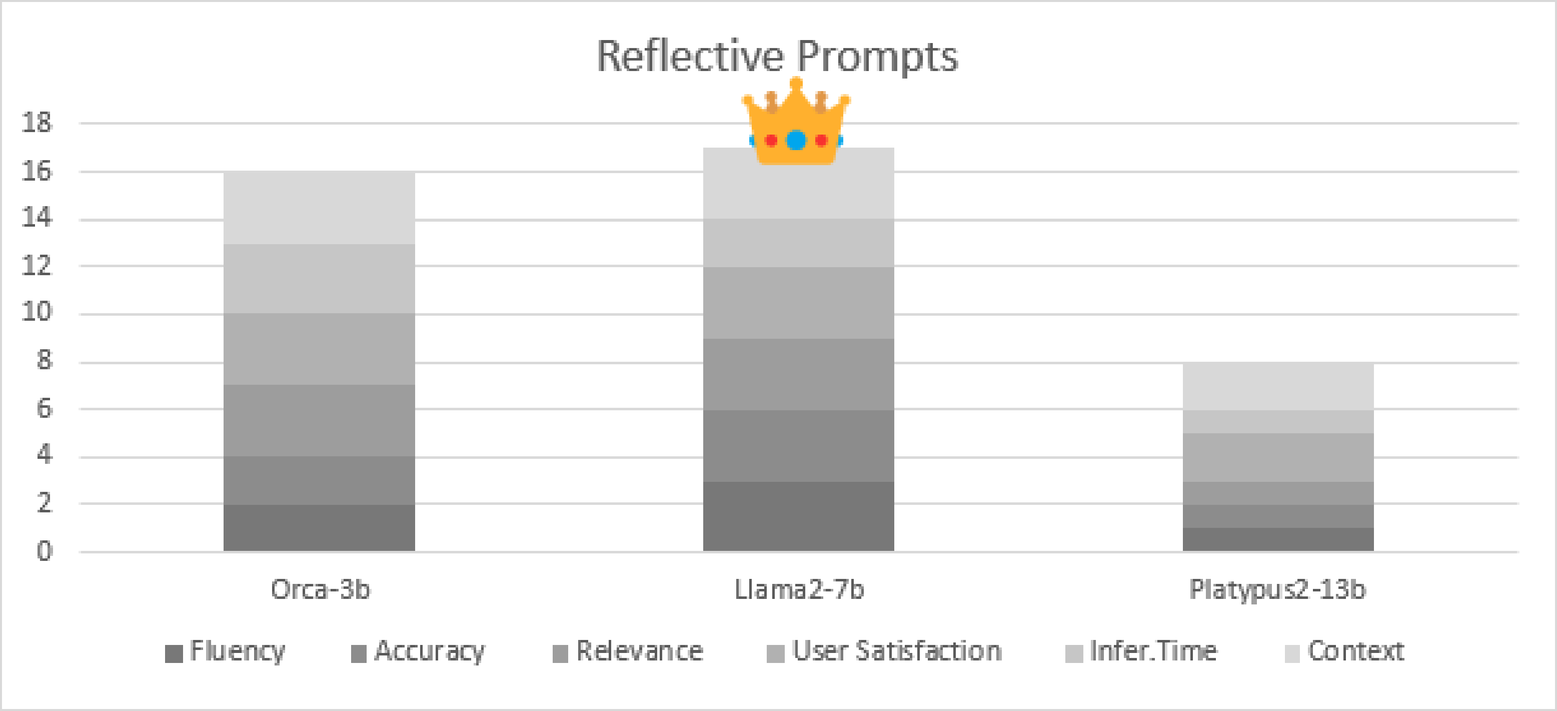
Llama2-7b is 3 times slower than Orca, but the complexity of the reply is certainly appreciated.

If you need to use these kind of prompts in your chatbot certainly Orca-3b is lighter and faster.

Platypus2 falls behind in all the evaluation matrix

Predictive Prompts

GGML	Fluency	Accuracy	Relevance	User Satisfaction	Infer.Time	Context	Max	total	performance	winner
Orca-3b	2	2	3	3	3	3	18	16	89%	
Llama2-7b	3	3	3	3	2	3	18	17	94%	👑
Platypus2-13b	1	1	1	2	1	2	18	8	44%	



TAKEAWAYS

Llama2-7b-Chat is the winner but Orca-3b is almost there!

In terms of inference time Orca was faster delivering anyway a good generation and following the instructions.

Llama2-7b is 4 times slower than Orca, but the complexity of the reply is certainly appreciated and it supports the statements with further explanations.

If you need to use these kind of prompts in your chatbot certainly Orca-3b is lighter and faster

Platypus has a Medium quality reply and not in a list format.