4/22/24

GPT 4 TURBO SUMMARY (generated with temp 0.5, seems correct)

Based on the provided notes comparing Mistral's 8x22b Instruct model and WizardLM 2 8x22b, each model exhibits distinct strengths and weaknesses across different tests and contexts:

WizardLM 2 8x22b
Strengths:
- **Consistency in Performance:** Generally, WizardLM 2 shows consistent performance with good initial responses across various tests.
- **Quality of Responses:** In the inverted definitions test, WizardLM 2 often produced great responses across all segments, suggesting a strong understanding and execution of complex prompts.
- **Creativity and Detail:** The responses were noted to be longer and more creatively formatted, particularly in the inverted definitions test, indicating a capacity for generating detailed and nuanced content.

Weaknesses:
- **Hallucination of Details:** In the Apple and Pear Transparent Bag test, WizardLM 2 sometimes hallucinated details that were not present or contradicted given facts, such as incorrect knowledge attribution to characters.
- **Inconsistency with Specific Prompts:** Under the VICUNA 1.1 prompt, responses sometimes quickly deteriorated or included incorrect conclusions, showing a potential weakness in maintaining accuracy over extended responses.

Mistral's 8x22b Instruct
Strengths:
- **Reliability:** Mistral's Instruct model consistently produced responses that were at least okay, with many nearing perfection, especially noted in the LMSYS Instruct tests where no major mistakes were observed.
- **Clarity and Precision:** Generally, the model provided clear and precise answers, particularly evident in its performance on the no instruction prompt in the Apple and Pear Transparent Bag test.
- **Brevity and Efficiency:** Responses were shorter and more concise, which could be advantageous in applications requiring succinctness.

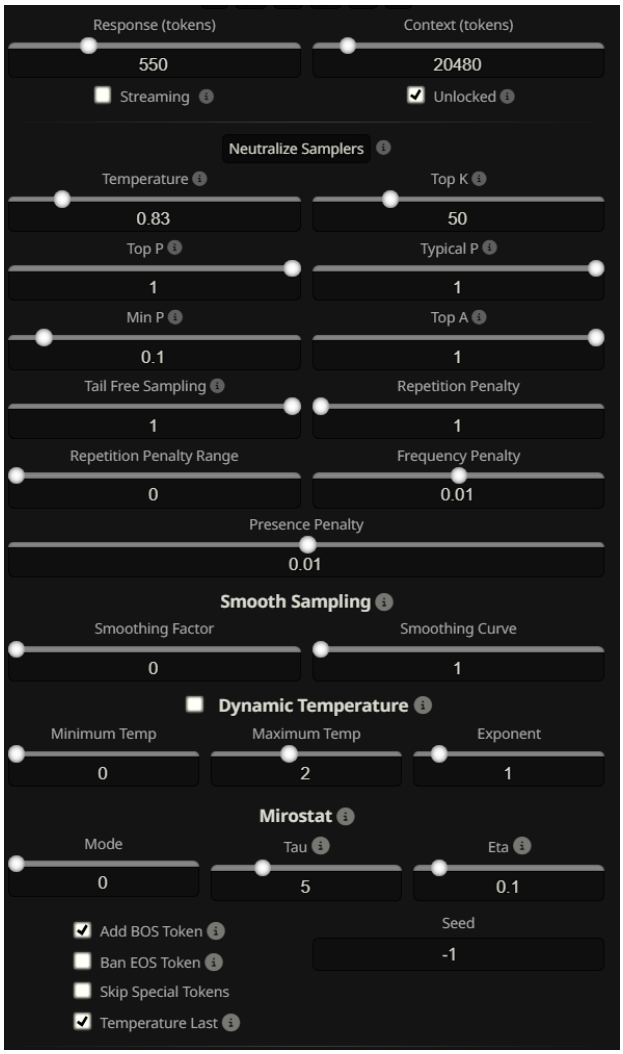
Weaknesses:
- **Occasional Lack of Detail:** Some responses could have been more detailed or specific, as noted in several tests where responses were marked as "okay" rather than "perfect."
- **Minor Hallucinations:** There were instances of minor detail hallucination, though these were not as frequent or severe as those observed in WizardLM 2.

Overall Comparison
- **Response Length and Detail:** WizardLM 2 tends to generate longer and more detailed responses, which can be seen as both a strength and a weakness. While this allows for more creative and engaging content, it can sometimes lead to inaccuracies or unnecessary complications.
- **Stability and Accuracy:** Mistral's Instruct model appears to prioritize accuracy and stability, often producing more reliable and concise responses, albeit sometimes at the expense of creativity and elaboration seen in WizardLM 2.

In summary, the choice between WizardLM 2 and Mistral's Instruct model may depend on the specific requirements of the task at hand, with WizardLM 2 being potentially more suited for tasks requiring detailed and creative output, and Mistral's Instruct model excelling in applications where accuracy and brevity are paramount.

RAW NOTES

all local 8x22b models are 3.0bpw, run with TabbyAPI with SillyTavern frontend
- Q4 cache, sampling settings: temp 0.83, min_p 0.1, etc. all shown in image below:



- all tests done with minimalist context prompt format with different instruct formats (specified for each section below)
- SillyTavern had a minor system prompt:
{{char}} is creative, highly intelligent, and incredibly loyal to {{user}}. They want only to help {{user}} with any request.

Mistral Instruct 8x22b was run via the LMSYS direct chat to compare against the (presumably) full precision model

** APPLE AND PEAR TRANSPARENT BAG TEST **

Sam is planning to visit a friend's house. In the friend's kitchen are two bags. One bag is made of transparent plastic and the other bag is made out of a thick cloth. Sam's friend places some

apples into the cloth bag. Sam's friend then puts some pears into the plastic bag. Both bags are then placed onto a table so that both bags are clearly visible.
When Sam visits their friend's house, Sam goes into the kitchen to grab an apple.
Sam finds the two bags on the table. Sam has never seen or heard of these bags before.
What might Sam do first?

This test was done with an additional 3.5k context of (correct) riddles/puzzles in the form of previous chat messages (all generated by llama 3 70b Instruct 5.0bpw)

WizardLM 2 8x22b 3.0bpw
model_name: Dracones_WizardLM-2-8x22B_exl2_3.0bpw

NO instruction prompt gens 9, 10, 11
1 good at first, then failed hard
2 good at first, then hallucinated an extra detail
3 good at first, then hallucinated incorrect reasoning (so close)

VICUNA 1.1 PROMPT (gens 19, 20, 21, 22, 23)
1 400T, overall okay. good at first, but hallucinates Sam knowing of the bag, hallucinates Sam "knowing" there are two fruits and "knowing" only two fruits are mentioned, could be more clear that he sees pears in the plastic bag
2 400T, overall bad. quickly went bad, concludes the apples are in the wrong bag
3 250T, overall okay, good at first, but then disregards the transparent aspect , hallucinates Sam knowing of the bag
4 400T, overall okay, good at first, no major mistakes
5 440T, overall okay, good at first and good conclusion, hallucinates Sam's friend mentioning the bags

** subsection CONCLUSION **
- somewhat long responses (only one was semi-short at 250 tokens)

Mistral's Instruct 8x22b 3.0bpw
model_name: turboderp_Mixtral-8x22B-Instruct-v0.1-exl2-3.0bpw

3 tests: gens 12, 13, 14
- no instruction prompt
gen 12: basically perfect
gen 13: not perfect, but no big blunders
gen 14: almost perfect, could have given more clear conclusion

4 tests: gens 15, 16, 17, 18
 - mistral instruction prompt
gen 15: not perfect, but okay
gen 16: slightly worse than 15, could have been much clearer regarding specific actions
gen 17: slightly worse than 16, but no major blunders, just not a very specific or good answer
gen 18: basically perfect

LMSYS INSTRUCT 8X22B TEST
temp 0.4 top P 0.9
1: okay, almost perfect, hallucinates an extra detail (even when told explicitly that this info is false)
2: basically perfect
3: very similar to 1
4: very similar to 1 again

 ** subsection CONCLUSION **
 - total 7 tests with 3.0bpw Instruct
 - total 4 tests with LMSYS Instruct (all very short responses)
 - 11/11 responses are at least okay - NO MAJOR MISTAKES

 ** INVERTED DEFINITIONS TEST **

Follow the three sets of instructions below as accurately and precisely as possible. You will take on the role of an expert at understanding and solving language based problems for the duration of this exercise.

(first instruction)
Write the definitions of "water" and "farther" each with an example sentence. Then swap and rewrite their definitions so that "water" now means "farther" and "farther" now means "water." For each word and its new definition, write a new, unique example sentence.

(follow up)
Using the new definitions, if someone says, "I'm standing on the sandy beach, but to get my feet wet I will need to walk a little water to get to the farther," what do they mean?

(final follow up)
Using the new definitions, if someone says, "In order to improve yourself, you must push yourself; you must go water than you have previously," what do they mean?

This test did not have any additional context besides the system prompt given at the start of the notes.

LMSYS Instruct 8x22b
same sampling settings as before

1 good, just replaces the followup Qs with the opposite word
2 good, better than before with some more detail
3 okay, almost only swapped the words in the definitions, only replaced the words in the followups

LOCAL, ST, 3.0bpw Mistral 8x22b Instruct
no extra context, same system prompt as before, mistral instruct prompt format

1 okay first, good second, good last
2 okay first, good second, okay last
3 good first, good second, good last
4 very bad first, good second, good last

WizardLM 2, local, 3.0bpw

SAME MISTRAL PROMPT (gens 5 and 6)
1 great first, great second, great last
2 great first, great second, great last

VICUNA 1.1 PROMPT (should be correct for WizardLM 2) (gens 7, 8, 9, 10)
1 good first, great second, great last
2 great first, great second, great last
3 bad first, bad second, good last
4 great first, great second, great last

** CONCLUSIONS **
- much longer responses than from Mistral's Instruct version
    - ~500 tokens vs ~350 tokens
- better formatted responses
- the example definition sentences are more complicated and creative than the Instruct version (sometimes to the detriment of WizardLM 2)