University of Pisa

Human Language Technologies Project: Zero Shot Learning with Llama 2

Pasquale Esposito[1] and **Sergio Latrofa**[2]

[1]*M.Sc. Computer Science, Free Curriculum* - 649153 - p.esposito8@studenti.unipi.it
[2]*M.Sc. Computer Science, Artificial Intelligence* - 640584 - s.latrofa1@studenti.unipi.it

July, 2023

1 Model

Llama 2, released in July 2023 by Meta.ai, is an updated version of Llama 1, trained on a new mix of publicly available data, with an increased size of the pretraining corpus by 40%, doubling the context length of the model, and adopting grouped-query attention , an enhanced kind of attention recently released as well (Ainslie et al., 2023). Variants of Llama 2 were released with 7B, 13B, and 70B parameters respectively. Llama 2-Chat, a fine-tuned version of Llama 2 that is optimized for dialogue use cases, variants of this model with 7B, 13B, and 70B parameters were released as well. In this work, due to computational and time limitations we only evaluated the 7B version of Llama 2 chat, thanks to the pipeline implementation available on `huggingface` after being authorized by META.

The training process begins with the initial pretraining, using publicly available online sources. The outcoming initial of the model would then undergo several stages of supervised fine-tuning. Subsequently, the model is iteratively refined using Reinforcement Learning with Human Feedback (RLHF) methodologies, specifically through rejection sampling and Proximal Policy Optimization (PPO). Throughout the RLHF stage, the accumulation of iterative reward modeling data in parallel with model enhancements is crucial to ensure the reward models remain within distribution.

2 Results

The evaluated version was the one provided with the zero shot classification pipeline for `huggingface`. Results were not so brilliant as one could expect, not reaching at all the nice zero-shot adaptation capabilities of the smallest version of FLAN previously analyzed (80M parameters vs 7B parameters). Such poorness of performance is an additional point in favour of FLAN training strategy. As authors claimed, Llama is probably more suitable for few-shot adaptation tasks, which is something that we did not manage to neither try because of the unexpected and "bizarre" responses of the generative pipeline.

Here follow a recap of prompt selection and evaluation performance.

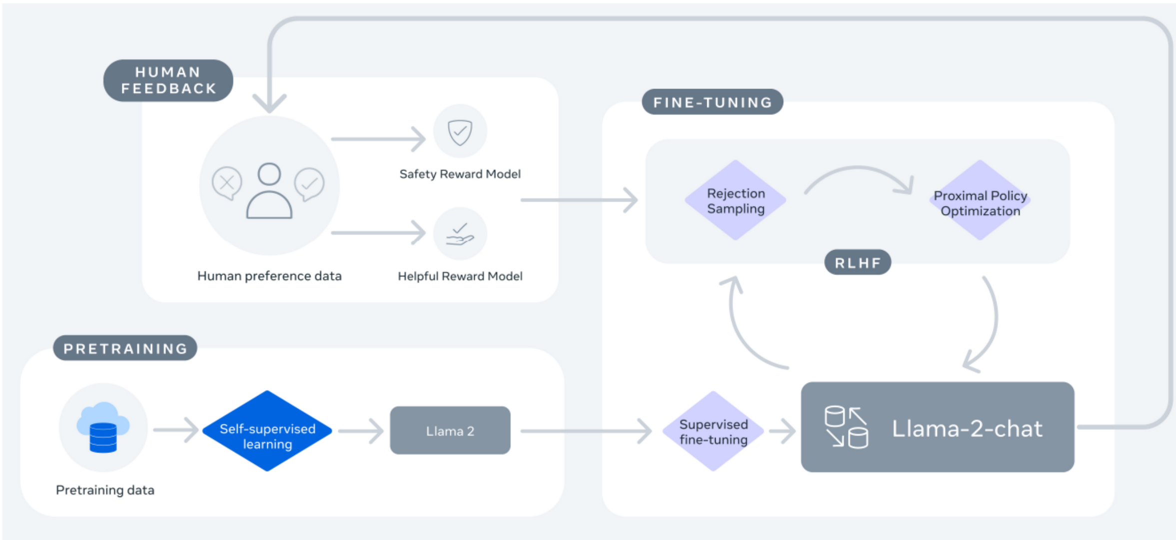

Figure 1: Schema of Llama2 training (taken from the official preprint.)

2.1 NLI

2.1.1 Prompt selection

Prompt	1	2	3	4	5	6	7	8	9	10	*Mean*
Accuracy	**0.563**	0.531	0.563	0.5	0.5	0.563	0.438	**0.656**	**0.625**	0.563	0.55

Table 1: Binary NLI Prompt Selection Results on *scitail* (32 samples).

Prompt	anli	glue-mnli	sick	superglue-cb	*Mean*	*St.dev.*
Prompt 1	**0.354**	**0.375**	0.25	0.406	0.346	0.068
Prompt 2	0.333	**0.354**	0.333	0.469	0.372	0.065
Prompt 3	**0.396**	0.313	0.333	0.188	0.308	0.087
Prompt 4	0.313	**0.417**	0.333	0.313	0.344	0.05
Prompt 5	0.271	**0.354**	0.354	0.344	0.331	0.04
Prompt 6	0.313	0.333	0.333	**0.5**	0.37	0.087
Prompt 7	0.333	**0.458**	0.333	**0.531**	0.414	0.098
Prompt 8	0.333	0.354	0.313	0.438	0.36	0.055
Prompt 9	0.313	0.292	0.208	0.406	0.305	0.081
Prompt 10	0.333	0.292	**0.417**	**0.5**	0.386	0.092
Prompt 11	0.333	0.271	0.354	0.344	0.326	0.037
Prompt 12	0.333	0.333	0.354	**0.5**	0.38	0.081
Prompt 13	**0.354**	0.292	**0.458**	0.406	0.378	0.071
Prompt 14	0.313	0.188	0.292	0.469	0.316	0.116
Prompt 15	**0.354**	0.354	0.188	0.313	0.302	0.079
Prompt 16	0.313	0.354	0.292	0.344	0.326	0.028
Prompt 17	0.333	0.333	0.333	**0.5**	0.375	0.084
Prompt 18	0.313	0.333	0.375	**0.5**	0.38	0.084
Prompt 19	0.333	0.313	0.354	0.469	0.367	0.07
Prompt 20	**0.354**	0.313	**0.417**	0.469	0.388	0.069
Prompt 21	0.333	0.313	0.333	0.25	0.307	0.039
Prompt 22	0.313	0.313	**0.417**	0.406	0.362	0.057
Prompt 23	0.313	0.313	0.354	0.313	0.323	0.021
Mean	0.33	0.329	0.336	0.408	0.351	-
St.dev.	0.024	0.052	0.063	0.092	-	0.07

Table 2: NLI Prompt Selection Results.

2.1.2 Test results

Dataset	$(acc$ - $1^{st}BestPrompt)$	$(acc - 2^{nd}BestPrompt$	$(acc - 3^{rd}BestPrompt)$
scitail	0.317 - *(8)*	**0.362** - *(15)*	**0.362** - *(17)*
anli	0.332 - *(1)*	0.331 - *(3)*	**0.337** - *(8)*
$glue - mnli$	**0.357** - *(1)*	0.324 - *(4)*	0.341 - *(2)*
sick	**0.422** - *(8)*	0.403 - *(15)*	0.42 - *(17)*
$superglue - cb$	**0.446** - *(1)*	0.339 - *(2)*	0.393 - *(5)*

Table 3: NLI Test Results.

2.1.3 Query time

Dataset	*Time*
scitail	*7.632*
anli	**22.98**
glue-mnli	13.25
sick	11.058
superglue - cb	23.2
Mean	15.624
STD	7.104

Table 4: Classification time of 32 NLI samples (in seconds).

2.2 QA

2.2.1 Prompt selection

In the column names have been used some abbreviations for datasets' names:

- "QuaRTz-no" stands for "QuaRTz-no_knowledge";
- "QuaRTz-with" stands for "QuaRTz-with_knowledge";
- "RACE" stands for "RACE-middle";
- "Social" stands for "Social IQA";
- "COPA" stands for "SUPERGLUE-COPA";
- "Wino" stands for "Wino Grande".

Prompt	QuaRel	QuaRTz-no	QuaRTz-with	RACE	SciQ	Social	COPA	Wino	*Mean*	*St.dev.*
Prompt 1	0.438	0.531	**0.5**	0.281	0.094	0.25	0.469	**0.5**	0.383	0.156
Prompt 2	0.344	**0.563**	**0.531**	0.219	0.094	0.219	0.5	0.406	0.36	0.17
Prompt 3	0.281	0.469	0.406	0.281	**0.219**	0.25	0.5	0.406	0.352	0.107
Prompt 4	0.438	0.438	0.375	**0.313**	0.094	0.125	0.5	0.438	0.34	0.153
Prompt 5	0.406	0.5	0.469	**0.313**	0.156	0.219	0.5	0.438	0.375	0.132
Prompt 6	0.438	0.5	0.438	0.25	0.063	0.344	0.375	0.438	0.356	0.141
Prompt 7	0.406	0.5	**0.5**	**0.313**	0.094	0.219	0.5	0.438	0.371	0.151
Prompt 8	0.438	**0.688**	**0.5**	**0.313**	0.188	0.281	0.5	0.469	0.422	0.157
Prompt 9	0.438	**0.563**	0.438	0.281	0.125	0.313	0.5	**0.5**	0.395	0.145
Prompt 10	**0.469**	0.406	0.375	0.281	**0.219**	0.344	0.438	0.469	0.375	0.09
Prompt 11	**0.469**	0.469	0.375	0.281	**0.219**	**0.406**	**0.531**	0.469	0.402	0.106
Prompt 12	0.438	0.438	0.375	**0.344**	0.156	**0.406**	0.5	**0.5**	0.395	0.111
Prompt 13	**0.469**	0.406	0.406	**0.313**	0.156	**0.375**	**0.531**	0.469	0.391	0.116
Prompt 14	0.438	0.438	0.438	0.281	**0.344**	0.313	**0.563**	**0.5**	0.414	0.096
Prompt 15	0.438	**0.563**	**0.5**	0.281	**0.25**	0.313	0.5	**0.5**	0.418	0.119
Mean	0.423	0.498	0.442	0.29	0.165	0.292	0.494	0.463	0.383	-
St.dev.	0.05	0.075	0.055	0.03	0.076	0.079	0.043	0.034	-	0.126

Table 5: QA Prompt Selection Results (32 samples).

2.2.2 Test results

Dataset	(acc - $1^{st}BestPrompt$)	($acc - 2^{nd}BestPrompt$)	($acc - 3^{rd}BestPrompt$)
$QuaRel$	0.522 - *(10)*	**0.54** - *(11)*	0.525 - *(13)*
$QuaRTz - no$	**0.484** - *(2)*	0.474 - *(8)*	0.337 - *(9)*
$QuaRTz - with$	**0.479** - *(1)*	0.471 - *(2)*	0.448 - *(7)*
$RACE$	0.228 - *(7)*	0.216 - *(8)*	**0.237** - *(12)*
$SciQ$	0.139 - *(3)*	0.171 - *(14)*	**0.188** - *(15)*
$Social$	0.316 - *(11)*	0.31 - *(12)*	**0.335** - *(13)*
$COPA$	0.41 - *(11)*	**0.45** - *(13)*	0.43 - *(14)*
$Wino$	**0.5** - *(1)*	0.48 - *(14)*	0.484 - *(15)*

Table 6: QA Test Results.

2.2.3 Query time

Dataset	*Time*
QuaRel	7.28
QuaRTz-no	7.073
QuaRTz-with	7.233
RACE	**46.883**
SciQ	26.088
Social	9.419
COPA	*5.723*
Wino	6.147
Mean	14.481
STD	14.715

Table 7: Classification time of 32 QA samples (in seconds).