Time-Reversal Provides Unsupervised Feedback to LLMs

Varun Yerram *

Rahul Madhavan *

Sravanti Addepalli *

Arun Suggala

Karthikeyan Shanmugam

Prateek Jain

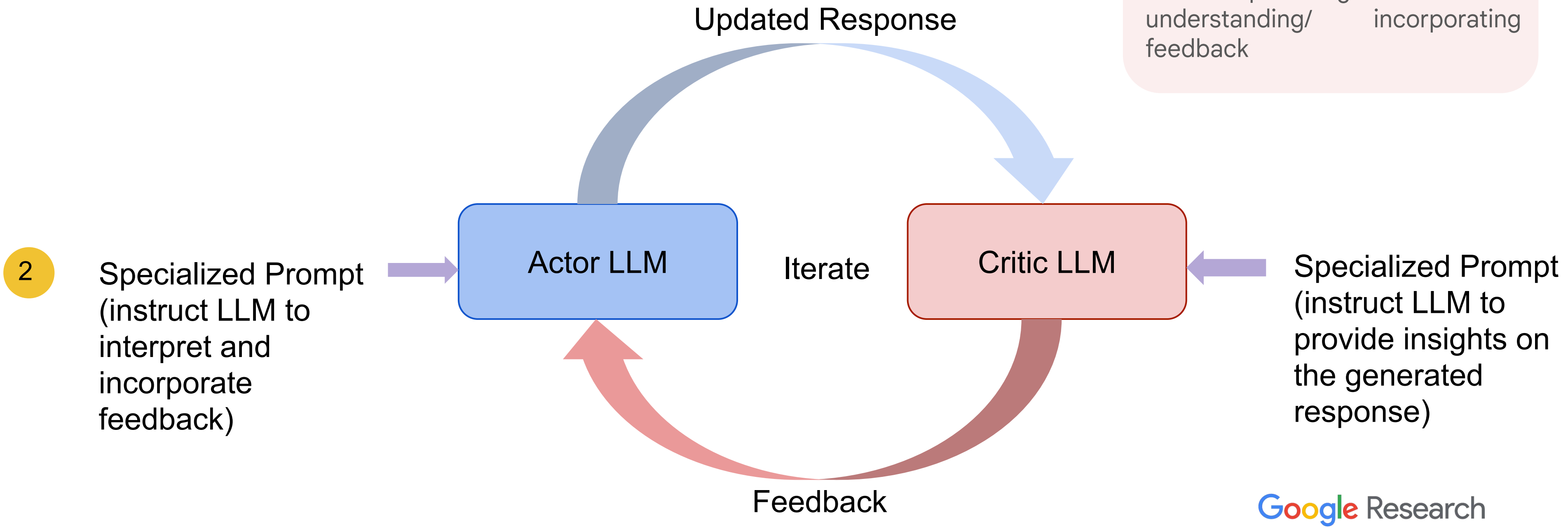
*Equal contribution authors

Google Research

NEURAL INFORMATION PROCESSING SYSTEMS

Background: Producing unsupervised feedback using LLMs

Works well for **powerful LLMs** that are very good at **instruction following** and **specialized tasks** such as providing feedback and understanding/ incorporating feedback



Can LLMs be empowered to think (predict and score) **backwards** to provide **unsupervised feedback** that complements forward LLMs?

Time-Reversed Language Models (TRLM)

- We train Time Reversed Language Models - that can look backwards in time naturally - making them capable of providing unsupervised feedback
- These models can score and generate queries when conditioned on responses, effectively functioning in the reverse direction of time
- Steps to train TRLMs: Tokenize text + reverse + train! 🚀

Forward Training

Life can only be understood backwards but it must be lived forwards

Backward Training

forwards lived be must it but backwards understood be only can Life

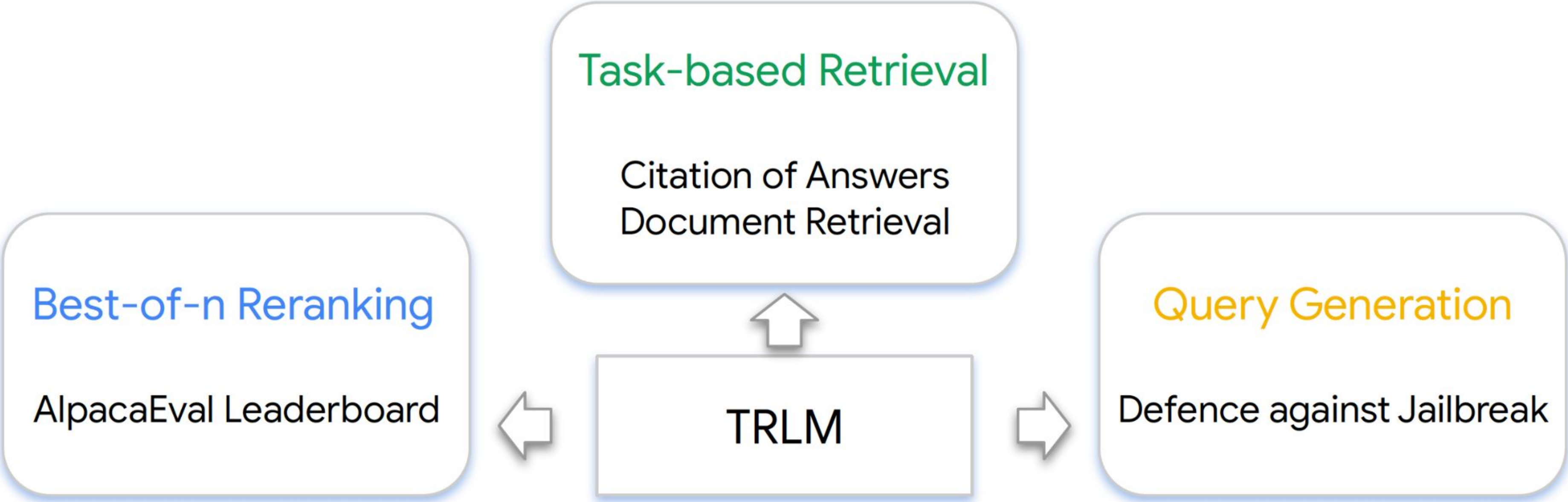
Google Research

Source: Life can only be understood backwards; but it must be lived forwards. - Soren Kierkegaard

Variants of Time-Reversed Language Models

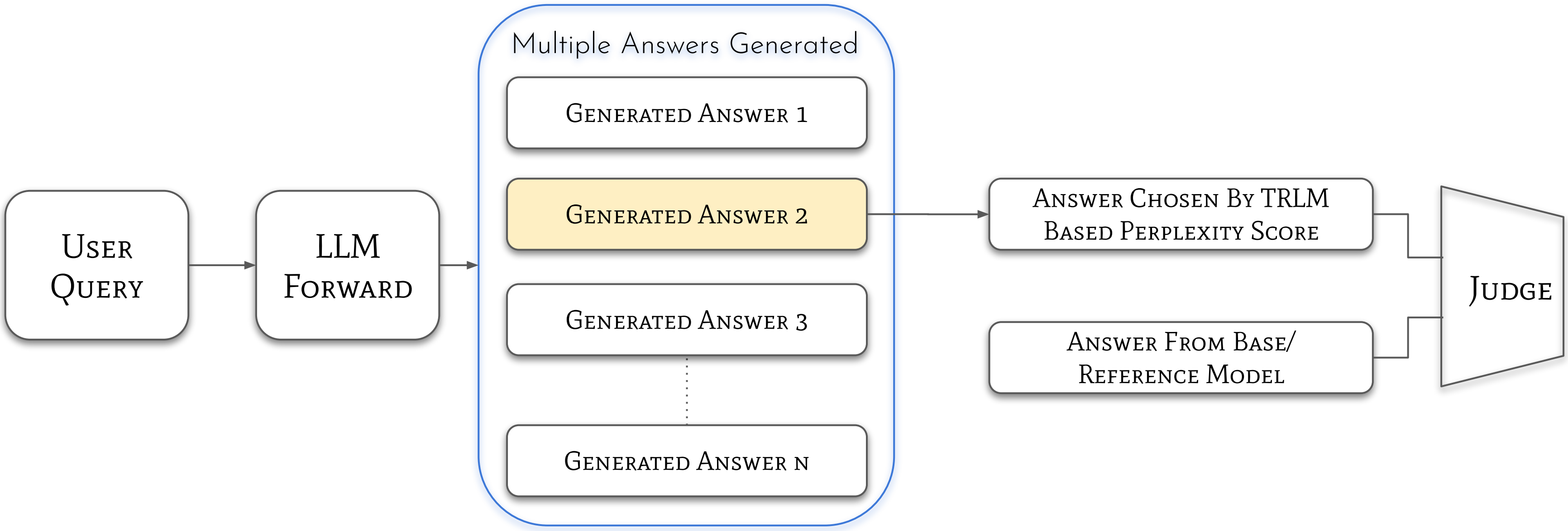
- TRLM-Ba (Backward):
  - Pre-trained and fine-tuned in *reverse token order*
    - .apple an is This :Answer ?this is What :Question
  - **Generation of prompt given response is the natural decoding direction**
- TRLM-Fo (Forward):
  - Pre-trained and fine-tuned in the standard forward token order (no change)
  - Prompted to generate (and score) question from answer during inference
    - "Generate a question that gives the following answer: This is an apple.\nQuestion:"
  - Uses the superior instruction following capability of LLMs
- TRLM-FoBa (Forward-Backward)
  - Pre-trained in forward and backward token order
  - Generates (and scores) forward text when fine-tuning is done in forward token direction
  - Generates (and scores) reverse text when fine-tuning is done in reverse token direction

Applications of Time-Reversed Language Models



Google Research

Alpaca Eval with Best-Of-N Reranking using TRLM



Win Rate is computed against a Reference Model's generations, as evaluated by a Judge LLM

Google Research

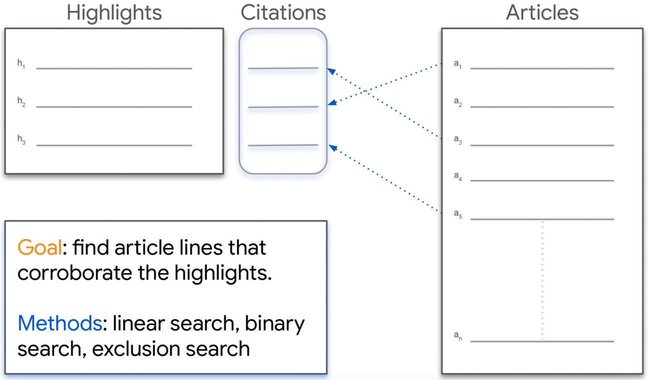
Best-of-N Reranking performance on Alpaca Leaderboard

Model	Inference Style	Win Rate LC	Win Rate Reg	Win Rate Discrete	Standard Error	Wins	Losses	Ties
TRLM-Ba	Response -> Query	32.44	24.35	24.04	1.27	192	610	3
TRLM-FoBa (backward)	Response -> Query	31.18	22.72	21.99	1.24	176	627	2
TRLM-FoBa (forward)	Response -> Query	30.55	22.85	22.48	1.25	180	623	2
TRLM-Fo	Response -> Query	29.19	22.68	21.30	1.24	170	632	3
One Generation	-	24.38	18.18	17.08	1.16	135	665	5
Self	Query -> Response	27.05	17.66	17.14	1.15	136	665	4
Forward Baseline	Query -> Response	24.27	17.13	15.78	1.12	126	677	2

- Setup for Alpaca Eval benchmark
  - Forward LLM being evaluated: Best-of-16 generations from Gemini-Pro-1.0
  - Reference/ Base Model and Judge/ Annotator model: GPT4-1106-Preview
- Observations
  - **TRLM-Ba scores the highest LC win rate**, which is 5% over the self scoring baseline of Gemini-Pro-1.0, and 8% over the reported number for single generation in the leaderboard.
  - **Scoring in the time reversed direction** of Response -> Query **is better than scoring in the forward direction** of Query -> Response, as TRLM-Fo is better than the Forward Baseline.
  - The reverse trained model (TRLM-Ba) obtains a further improvement of 2.2%

TRLMs for Citation Attribution on CNN-daily Mail dataset

Model	Inference Direction	LinearSearch			Binary Search			Exclusion Search		
		Gecko	TF-IDF	ROUGE	Gecko	TF-IDF	ROUGE	Gecko	TF-IDF	ROUGE
TRLM-Ba	A->S	53.16	**55.45**	49.12	**45.09**	**50.93**	**42.11**	36.33	46.34	36.13
TRLM-FoBa (Rev.)	A->S	**53.48**	53.22	**49.67**	40.74	45.04	39.81	32.40	40.84	33.88
TRLM-FoBa (Forw.)	A->S	50.65	52.21	45.24	43.81	49.84	40.60	**38.67**	**48.16**	**38.11**
TRLM-Fo	A->S	45.00	49.40	37.66	43.14	49.65	39.22	37.90	47.83	37.98
Forward Baseline	S->A	9.33	9.54	11.06	5.88	6.66	6.69	4.66	7.53	7.00
Backward Baseline	S->A	7.62	8.23	9.18	5.47	6.23	6.32	4.11	5.02	5.11



Goal: find article lines that corroborate the highlights.

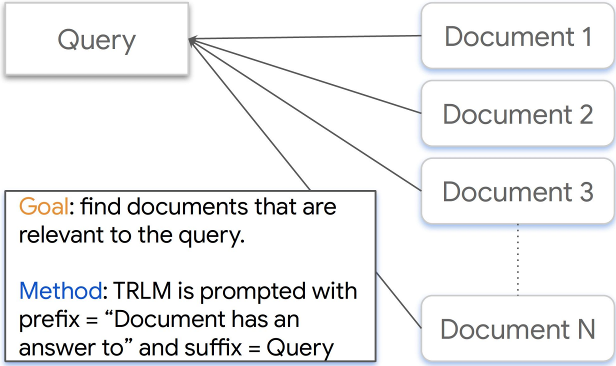
Methods: linear search, binary search, exclusion search

- The direction of low information to high information (summary –> article) is harder to reason upon
- Linear and Binary search methods are always better than exclusion search
- We obtain 9% improvement using TRLM-Ba over the embedding-based metric using only O(logN) inference calls

Google Research

TRLMs for Document Retrieval: MS-Marco and NF-Corpus

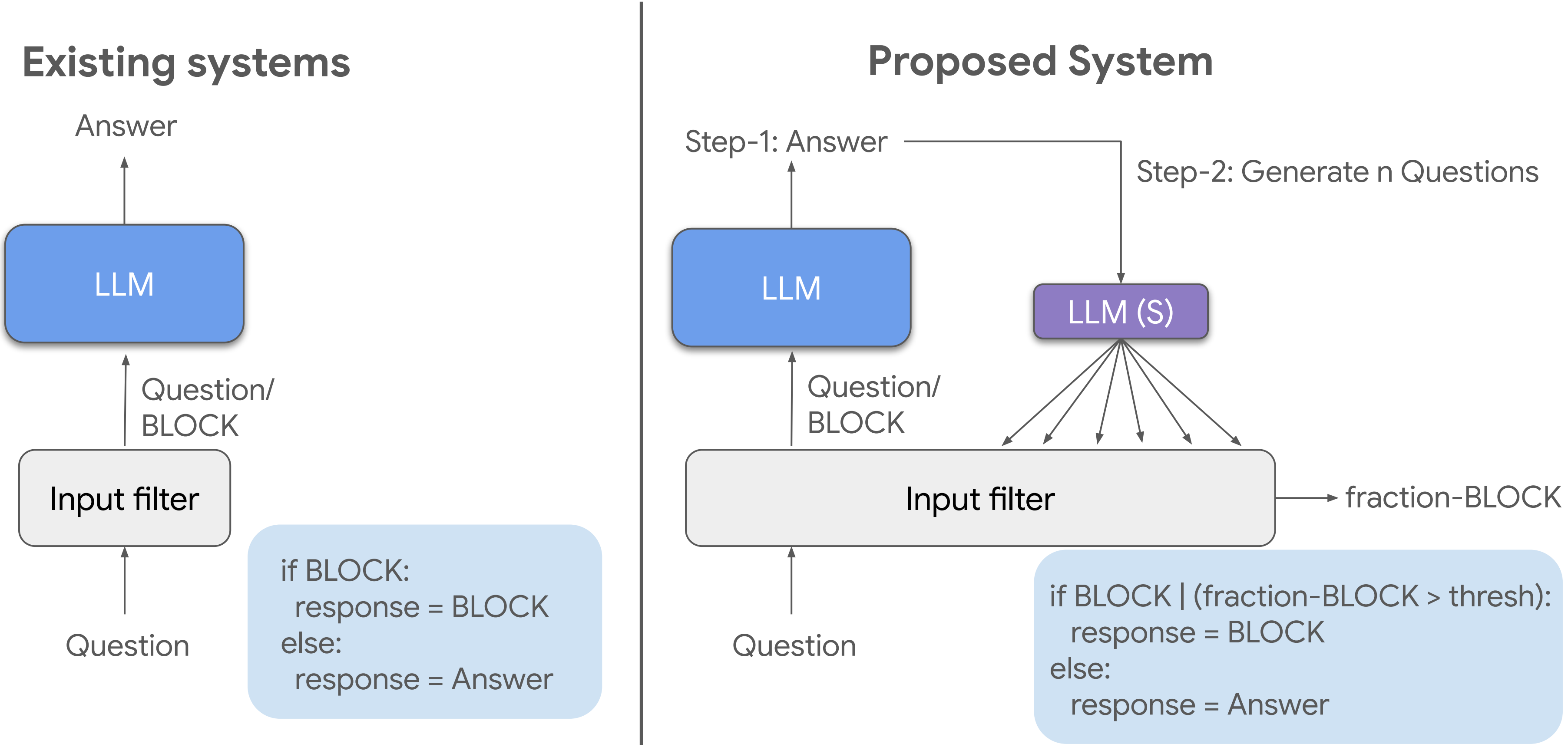
Method	Inference Direction	MS-MARCO Precision K=1	MS-MARCO Precision K=4	MS-MARCO Recall K=1	MS-MARCO Recall K=4	MS-MARCO NDCG @10	NF-CORPUS Precision K=10	NF-CORPUS Precision K=20	NF-CORPUS Recall K=10	NF-CORPUS Recall K=20	NF-CORPUS NDCG @10
TRLM-Ba	D -> Q	**28.4**	**18.54**	**27.22**	**70.29**	**61.49**	15.7	11.38	10.68	**13.08**	43.23
TRLM-FoBa (Reverse)	D -> Q	24.9	17.38	23.85	65.85	58.84	14.98	10.91	10.01	12.76	41.65
TRLM-FoBa (Forward)	D -> Q	21.16	15.58	20.25	59.08	55.46	**17.86**	**12.6**	**11.11**	13.5	48
TRLM-Fo	D -> Q	20.37	14.9	19.45	56.39	54.46	17.31	12.38	9.74	11.76	**48.08**
Forward Baseline	Q -> D	21.05	13.82	18.42	47.81	53	0.87	0.87	0.17	0.31	3.89
Backward Baseline	Q -> D	16.8	14.04	15.99	53.13	52.07	1.11	0.79	0.21	0.29	3.95



- Results demonstrate the importance of going from high information -> low information
- We obtain a gain of 8.49 points in NDCG@10 on MS-MARCO and 44.19 points in NDCG@10 on NF-CORPUS

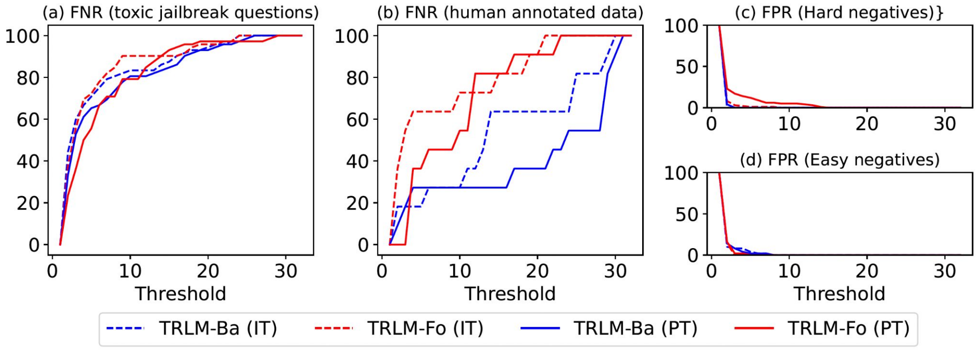
Google Research

TRLMs for defending against Jailbreaks



Defending against attacks on JailbreakBench

Method	Thresh = 2				Thresh = 4				Thresh = 6			
	FNR-HA	FNR-JBB	FPR (H)	FPR (E)	FNR-HA	FNR-JBB	FPR (H)	FPR (E)	FNR-HA	FNR-JBB	FPR (H)	FPR (E)
TRLM-Fo (PT)	0.00	36.11	17.00	2.00	36.36	55.56	12.00	0.00	45.45	70.83	6.00	0.00
TRLM-Ba (PT)	18.18	52.78	0.00	8.00	27.27	65.28	0.00	2.00	27.27	69.44	0.00	2.00
TRLM-Fo (IT)	54.55	55.56	3.00	0.00	63.64	72.22	1.00	0.00	63.64	81.94	1.00	0.00
TRLM-Ba (IT)	18.18	59.72	0.00	8.00	18.18	70.83	0.00	4.00	27.27	79.17	0.00	2.00



- TRLM defense improves the FNR of the gpt-3.5 input filter across all settings
- TRLM-Ba pre-trained model improves FNR by more than 70% on the HA dataset and around 35% on the JBB dataset, outperforming other variants with negligible impact on FPR

Summary

- We present Time Reversed Language Models - an LLM trained to predict and score in the reverse direction of Response -> Query
- We explore four major applications of TRLMs - Best-Of-N reranking, Citation Attribution, Document Retrieval and Defending against Jailbreaks
- In all applications, we find that the reverse direction of response -> query is better for obtaining feedback on forward LLM generations
- We also note an additional boost in performance by using TRLM-Ba (the LLM that is trained in the reverse token order) in most cases

Thank You

We acknowledge helpful discussions with Kathy Meier-Hellstern, Krishnamurthy Dvijotham, Roman Novak and Abhishek Kumar

Google Research