Atomic Self-Consistency for Better Long Form Generations

Raghuveer Thirukovalluru, Yukun Huang, Bhuwan Dhingra

rt195@duke.edu

Duke University

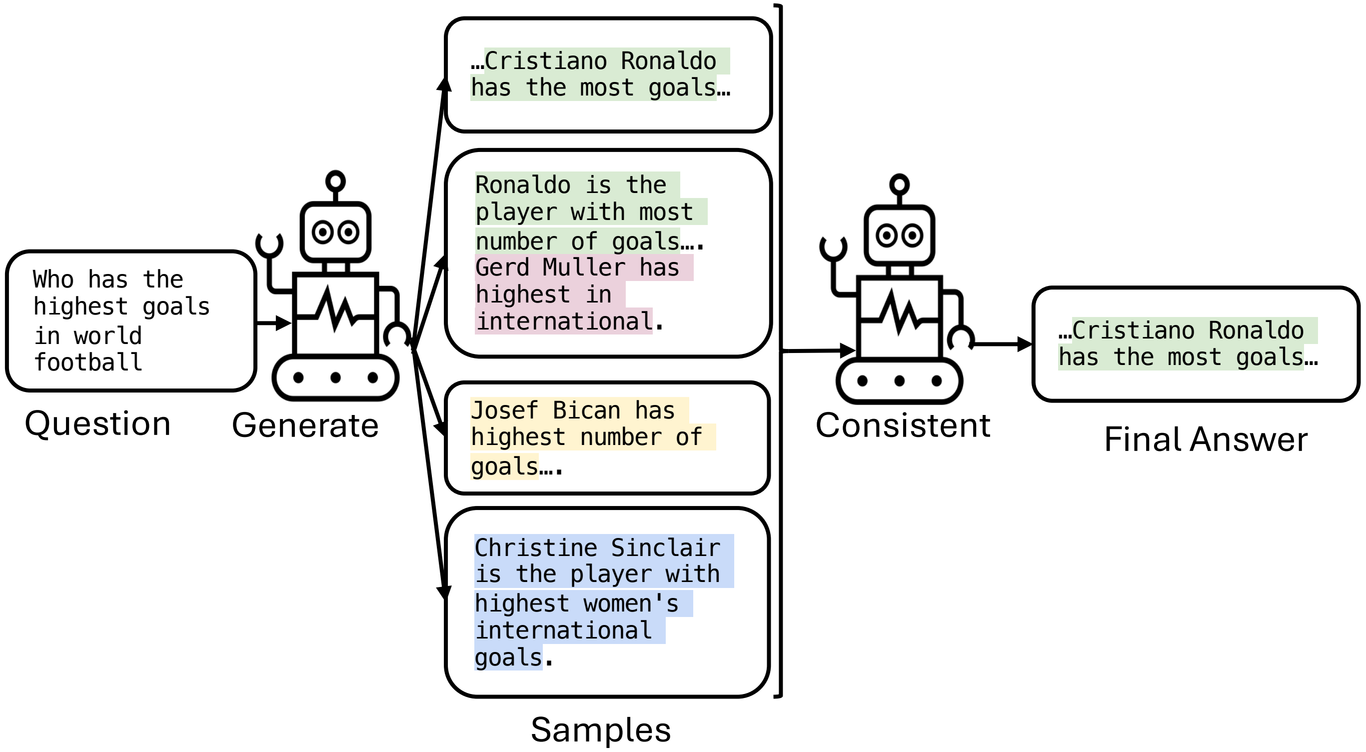
Introduction

- In Long Form Question Answering (LFQA), each response comprises multiple pieces of information (atomic facts) that collectively contribute to the overall correctness of the answer.

- Recent work has aimed to improve LLM generations by filtering out hallucinations, thereby improving the precision of the information in the response (Dhuliawalia et al. 2023; Min et al. 2023; Manukul et al., 2023).

- Higher response quality has also been achieved by stochastically sampling multiple model responses and then using consistency/other criteria to select one as the final answer (Chen et al., 2023; Ren et al., 2023).

Related Work: Universal Self Consistency (USC)

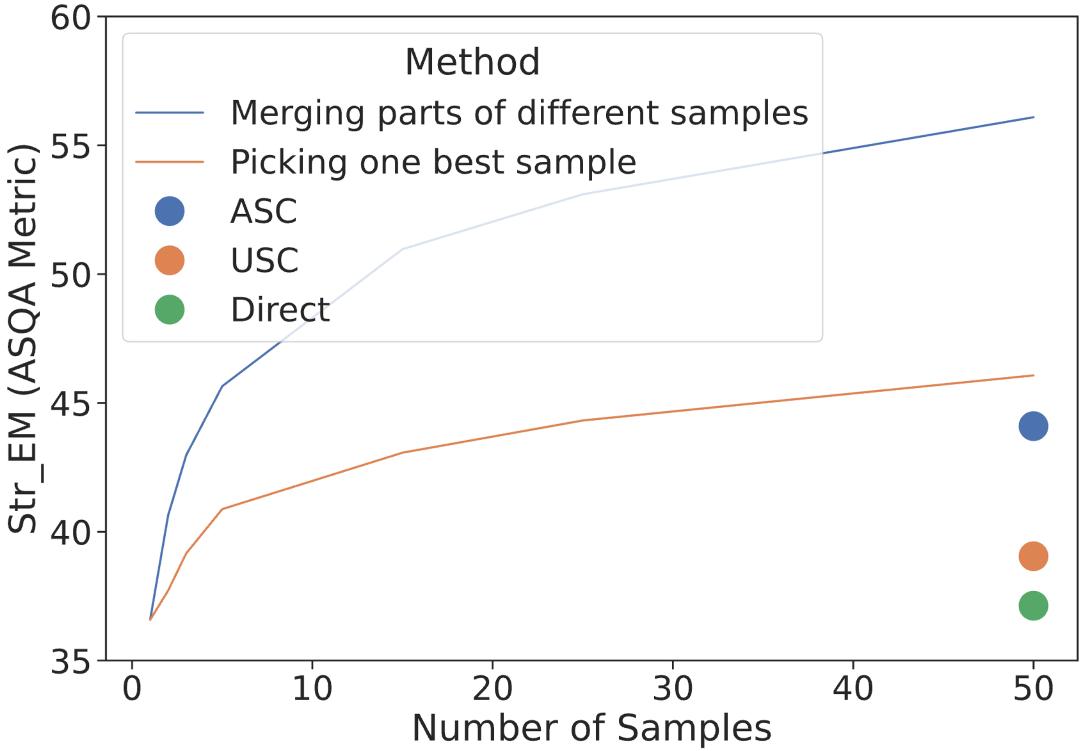


Limitations of Prior Work

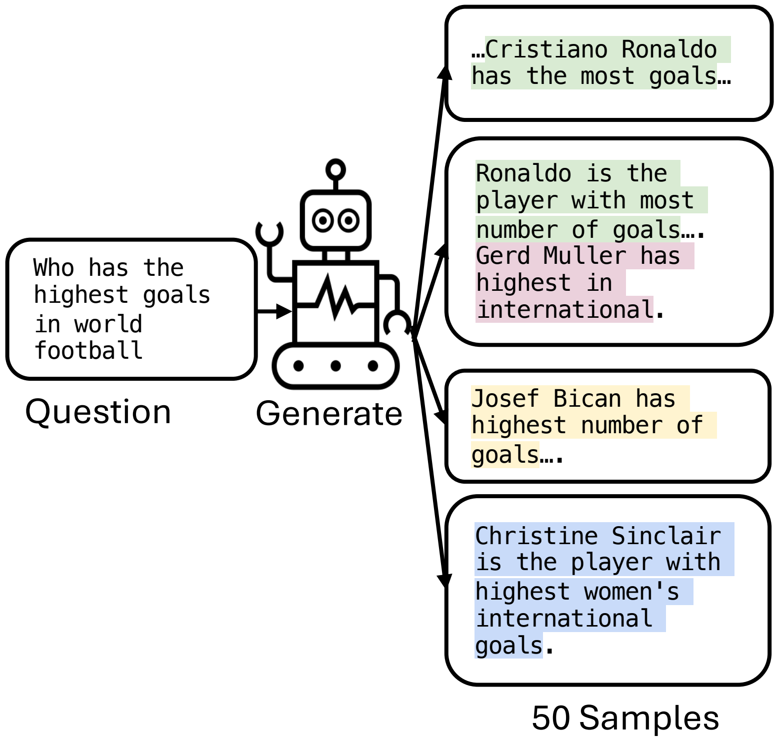
- Focused on <u>Precision of Atomic Facts</u>.
- Selects <u>one single sample</u> (among multiple samples) as the final answer. <u>Misses out on recall</u> of other samples. Also <u>allows for atomic hallucinations</u> within the sample selected.

Atomic Self Consistency (ASC)

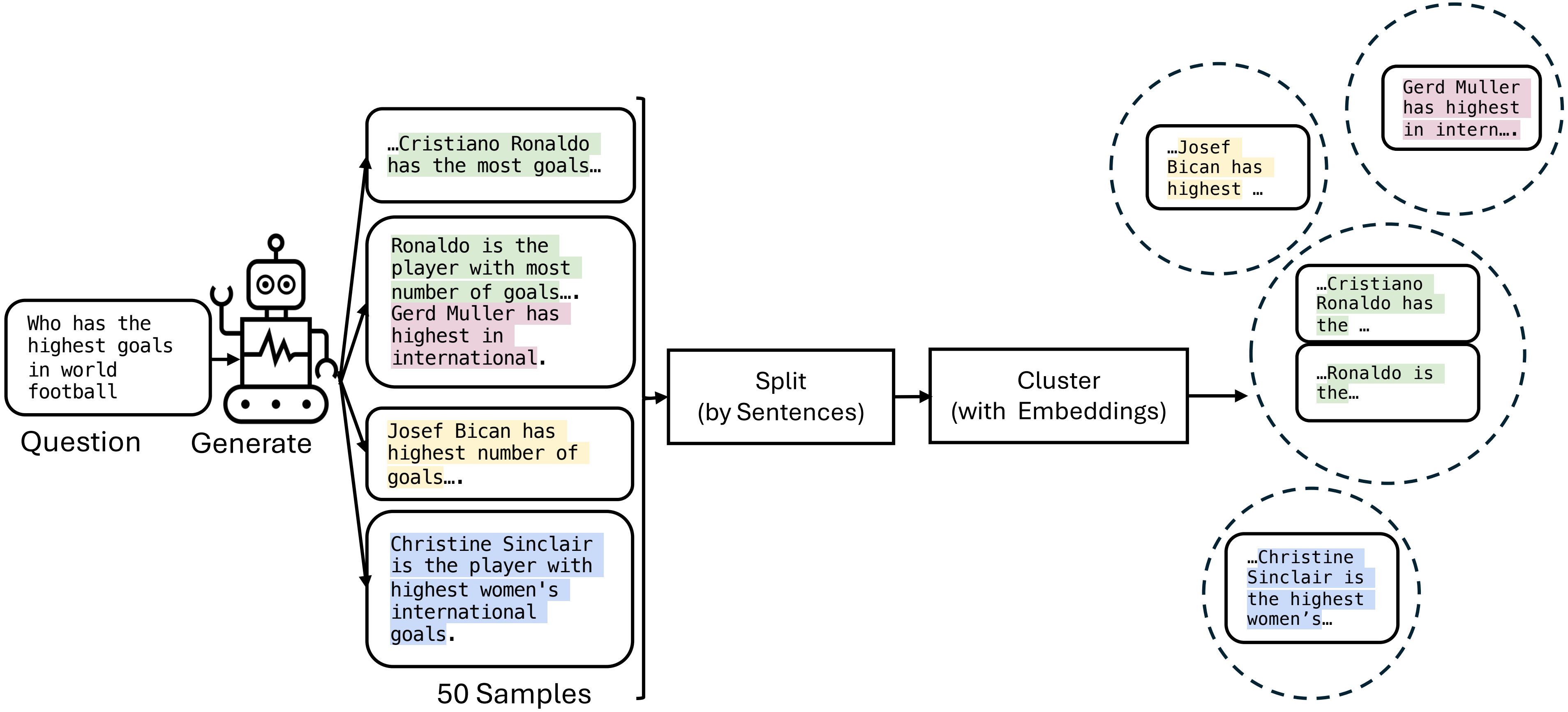
- Merges authentic subparts of multiple samples to generate a superior composite response.



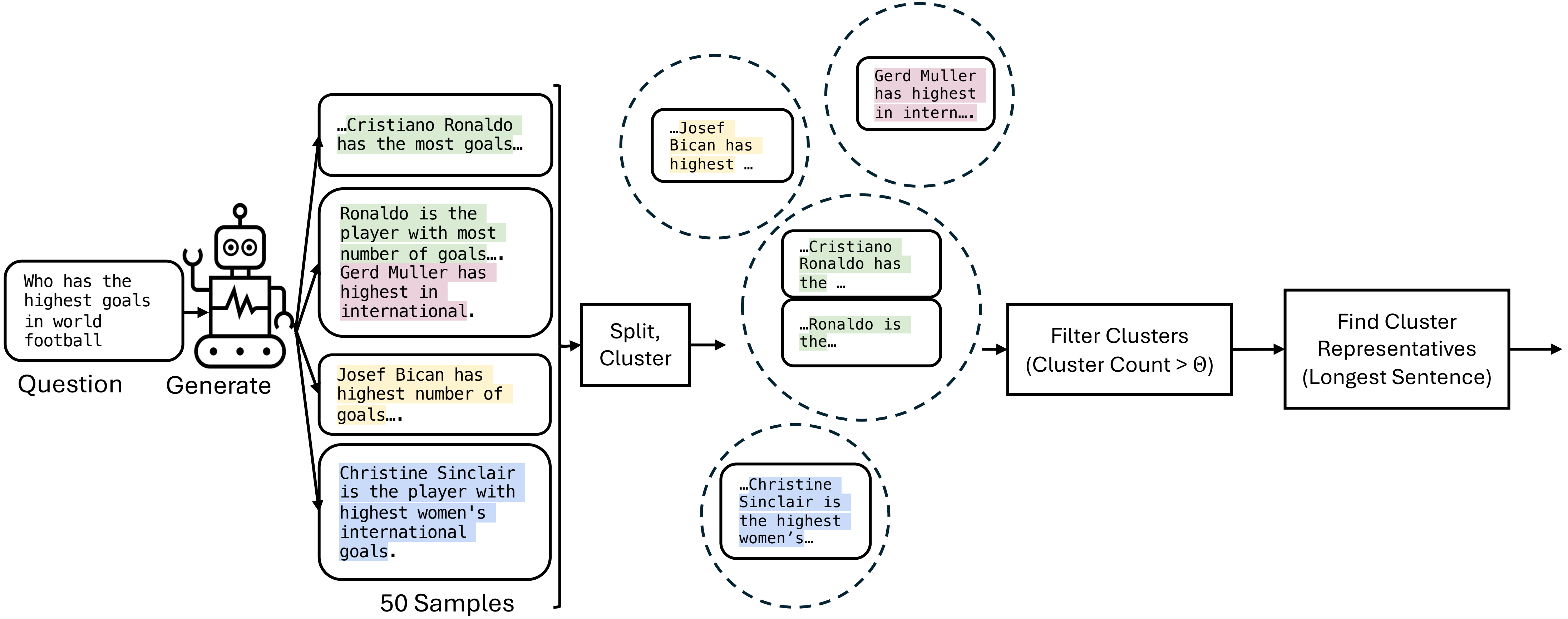
Atomic Self Consistency (ASC)



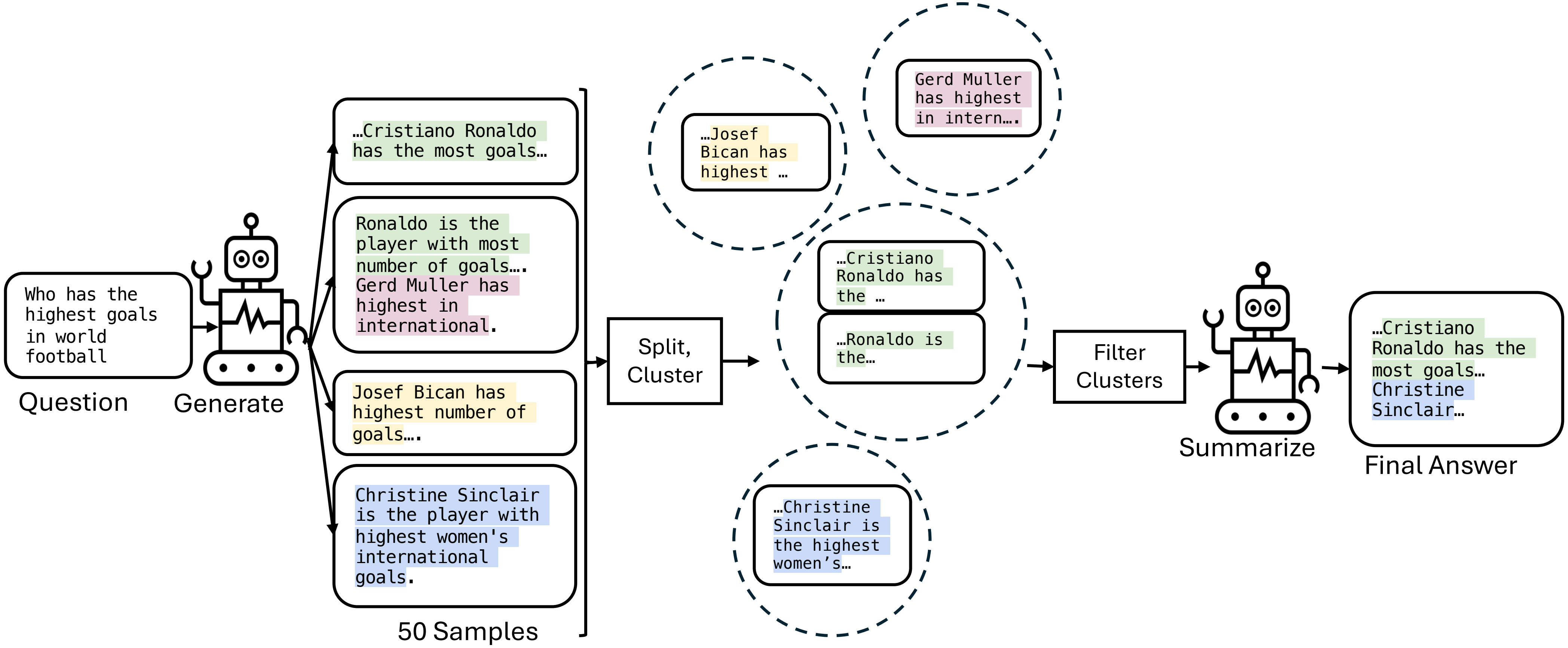
Atomic Self Consistency (ASC)



Atomic Self Consistency (ASC)



Atomic Self Consistency (ASC)



ASC Qualitative Analysis

- Question: When did the voortrekkers arrive in south africa?
- ASC Answer: ”..1835 onwards, with the main wave of migration taking place between 1836 and 1840. Led by pioneers such as Louis Trichardt, Piet Retief, and Andries Pretorius, they established several new independent republics, including the Orange Free State and the South African Republic (Transvaal)..... some groups arriving as early as 1834. ... KwaZulu-Natal in 1837, establishing the Republic of Natalia. The Great Trek, ..., began in earnest in 1836. ....”
- USC Answer: "The Voortrekkers, a group of Dutch-speaking settlers also known as Boers, arrived in South Africa in the 1830s. They embarked on a Great Trek, moving inland ....”
- Qualitative Analysis: ASC answer comprises a much detailed response with multiple years which can address most disambiguations of the question. USC gives a generalized response saying it was in the 1830s.

Experiments

		ASQA					ELI5			
		#Clusters	length	Mauve	Str_EM	QA-F1	#Clus.	length	Mauve	Claims_Nli
	Direct		56.29	44.64	37.13	29.33		104.35	24.57	18.66
	ACF		42.99	53.66	36.16	28.98		84.11	20.73	18.2
	FCF	-	45	52.68	36.84	29.64	-	94.75	27.97	18.7
ChatGPT	USC-LLM		56.72	44.88	37.91	29.71		104.13	21.11	18.76
	USC		64.52	40.19	39.05	30.88		97.36	24.09	17.4
	ASC-F (Ours)	30.74	106.7	41.25	**44.96**	31.91	56.83	172.66	22.68	**22.16**
	ASC (Ours)	15.7	101.17	47.01	44.1	**32.22**	16.68	163.58	21.29	21.43
	Direct		41.88	68	28.71	23.58		84.38	46.59	13.98
	ACF	-	25.78	63.79	28.48	24.73	-	58.20	38.22	13.70
Llama2	FCF		28.71	68.22	28.38	24.64		66.96	35.20	14.57
	USC		63.7	63.63	33.16	26.42		115.82	35.21	17.70
	ASC-F (Ours)	33.57	108.18	62.68	**39.26**	26.54	83.42	148.30	35.25	18.97
	ASC (Ours)	12.68	91.91	70.52	38.82	**27.16**	14.32	143.07	28.09	**19.40**

Table 1: ASQA, ELI5 results. ASC does the best on QA-F1 and demonstrates strong Str_EM. ASC-F picks a large number of clusters and does well on Str_EM. ASC also demonstrates strong Mauve. ASC, ASC-F achieve best Claims_Nli score on ELI5. Results justify that merging of samples is better than picking one sample.

Experiments

		ASQA					ELI5			
		#Clusters	**length**	**Mauve**	**Str_EM**	**QA-F1**	**#Clus.**	**length**	**Mauve**	**Claims_Nli**
ChatGPT	Direct	-	56.29	44.64	37.13	29.33	-	104.35	24.57	18.66
	ACF		42.99	53.66	36.16	28.98		84.11	20.73	18.2
	FCF		45	52.68	36.84	29.64		94.75	27.97	18.7
	USC-LLM		56.72	44.88	37.91	29.71		104.13	21.11	18.76
	USC		64.52	40.19	39.05	30.88		97.36	24.09	17.4
	ASC-F (Ours)	30.74	106.7	41.25	**44.96**	31.91	56.83	172.66	22.68	**22.16**
	ASC (Ours)	15.7	101.17	47.01	<u>44.1</u>	**32.22**	16.68	163.58	21.29	<u>21.43</u>
Llama2	Direct	-	41.88	68	28.71	23.58	-	84.38	46.59	13.98
	ACF		25.78	63.79	28.48	24.73		58.20	38.22	13.70
	FCF		28.71	68.22	28.38	24.64		66.96	35.20	14.57
	USC		63.7	63.63	<u>33.16</u>	<u>26.42</u>		115.82	35.21	17.70
	ASC-F (Ours)	33.57	108.18	62.68	**39.26**	<u>26.54</u>	83.42	148.30	35.25	<u>18.97</u>
	ASC (Ours)	12.68	91.91	70.52	<u>38.82</u>	**27.16**	14.32	143.07	28.09	**19.40**

Table 1: ASQA, ELI5 results. ASC does the best on QA-F1 and demonstrates strong Str_EM. ASC-F picks a large number of clusters and does well on Str_EM. ASC also demonstrates strong Mauve. ASC, ASC-F achieve best Claims_Nli score on ELI5. Results justify that merging of samples is better than picking one sample.

Experiments

	Method	QAMPARI						QUEST					
		#Pred	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
ChatGPT	`Direct`	5.2	21.35	13.82	23.47	15.35	21.83	5.56	12.05	6.76	12.91	7.45	11.6
	`ACF`	3.61	24.16	12.5	21.96	15.04	22.18	3.07	14.71	5.65	10.67	7.06	11.53
	`FCF`	4.41	22.59	13.29	23.16	15.33	22.16	3.61	13.55	5.91	11.03	7.01	11.27
	`USC-LLM`	4.95	20.88	13.39	22.91	14.94	21.33	5.10	11.86	6.18	11.92	7.08	11.16
	`USC`	8.97	20.7	19.21	31.28	18.07	24.2	7.83	11.98	8.43	15.19	8.23	12.21
	`ASC-F`	40.83	13.42	29.81	45.04	15.7	18.82	39.9	7.94	17.31	30.73	8.47	10.84
	`ASC`	7.09	22.98	20.5	33.04	**19.46**	**26.21**	8.44	12.47	10.41	19.15	**9.75**	**14.09**
Llama2	`Direct`	4.86	13.5	9.25	16.23	10.22	14.47	5.46	6.74	4.16	7.66	4.42	6.7
	`ACF`	3.17	14.94	7.96	13.84	9.69	13.85	3.48	7.9	3.47	6.34	4.14	6.54
	`FCF`	3.88	14.1	8.93	15.36	10.15	14.22	3.43	8.06	3.78	6.75	4.38	6.77
	`USC`	7.44	14.07	11.61	20.04	11.64	15.99	9.36	7.76	5.4	10.16	5.38	**7.96**
	`ASC-F`	27.35	10.74	18.44	29.88	11.52	14.4	28.07	5.63	10.64	19.08	**5.81**	7.67
	`ASC`	6.08	14.51	12.15	20.58	**12.15**	**16.44**	6.77	7.42	5.52	9.97	5.3	7.86

Table 2: `ASC` outperforms `Direct`, `USC` and `ASC-F`. `ASC-F` picks a large number of clusters and does worse on P, F1, F1-5. Results justify that consistency-based cluster selection does better than retrieval-based cluster selection.

Experiments

		QAMPARI						*QUEST*					
	Method	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
ChatGPT	`Direct`	5.2	21.35	13.82	23.47	15.35	21.83	5.56	12.05	6.76	12.91	7.45	11.6
	`ACF`	3.61	24.16	12.5	21.96	15.04	22.18	3.07	14.71	5.65	10.67	7.06	11.53
	`FCF`	4.41	22.59	13.29	23.16	15.33	22.16	3.61	13.55	5.91	11.03	7.01	11.27
	`USC-LLM`	4.95	20.88	13.39	22.91	14.94	21.33	5.10	11.86	6.18	11.92	7.08	11.16
	`USC`	8.97	20.7	19.21	31.28	18.07	24.2	7.83	11.98	8.43	15.19	8.23	12.21
	`ASC-F`	40.83	13.42	29.81	45.04	15.7	18.82	39.9	7.94	17.31	30.73	8.47	10.84
	`ASC`	7.09	22.98	20.5	33.04	**19.46**	**26.21**	8.44	12.47	10.41	19.15	**9.75**	**14.09**
Llama2	`Direct`	4.86	13.5	9.25	16.23	10.22	14.47	5.46	6.74	4.16	7.66	4.42	6.7
	`ACF`	3.17	14.94	7.96	13.84	9.69	13.85	3.48	7.9	3.47	6.34	4.14	6.54
	`FCF`	3.88	14.1	8.93	15.36	10.15	14.22	3.43	8.06	3.78	6.75	4.38	6.77
	`USC`	7.44	14.07	11.61	20.04	11.64	15.99	9.36	7.76	5.4	10.16	5.38	**7.96**
	`ASC-F`	27.35	10.74	18.44	29.88	11.52	14.4	28.07	5.63	10.64	19.08	**5.81**	7.67
	`ASC`	6.08	14.51	12.15	20.58	**12.15**	**16.44**	6.77	7.42	5.52	9.97	5.3	7.86

Table 2: `ASC` outperforms `Direct`, `USC` and `ASC-F`. `ASC-F` picks a large number of clusters and does worse on P, F1, F1-5. Results justify that consistency-based cluster selection does better than retrieval-based cluster selection.

Ablations (Random Selection)

		ASQA					QAMPARI					
Ablation	**Method**	**#Clusters**	**length**	**Mauve**	**Str_EM**	**QA-F1**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
1	ASC	15.7	101.17	47.01	**44.1**	**32.22**	7.09	22.98	20.5	33.04	**19.46**	**26.21**
	Random Clusters	15.7	85.31	49.97	<u>42.62</u>	<u>31.75</u>	7.09	11.86	10.08	18.62	9.77	14.05
	Random Sentences	15.7	99.45	42.08	41.5	29.36	7.09	22.19	13.8	24.42	15.39	22.1
2	USC	-	64.52	40.19	39.05	30.88	8.97	20.7	19.21	31.28	<u>18.07</u>	<u>24.2</u>
	High Token/#Pred		82.93	40.59	37.8	28.79	10.48	17.19	18.3	29.28	16.07	21.01

Table 3: Ablation 1: ASC performs better than randomly picking clusters and randomly picking sentences on ASQA, QAMPARI. Ablation 2: Larger length or higher #Predictions in response is not critical for better performance.

Ablations (Random Selection)

		ASQA					*QAMPARI*					
Ablation	**Method**	**#Clusters**	**length**	**Mauve**	**Str_EM**	**QA-F1**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
1	`ASC`	15.7	101.17	47.01	**44.1**	**32.22**	7.09	22.98	20.5	33.04	**19.46**	**26.21**
	`Random Clusters`	15.7	85.31	49.97	<u>42.62</u>	<u>31.75</u>	7.09	11.86	10.08	18.62	9.77	14.05
	`Random Sentences`	15.7	99.45	42.08	41.5	29.36	7.09	22.19	13.8	24.42	15.39	22.1
2	`USC`	-	64.52	40.19	39.05	30.88	8.97	20.7	19.21	31.28	<u>18.07</u>	<u>24.2</u>
	`High Token/#Pred`		82.93	40.59	37.8	28.79	10.48	17.19	18.3	29.28	16.07	21.01

Table 3: Ablation 1: ASC performs better than randomly picking clusters and randomly picking sentences on ASQA, QAMPARI. Ablation 2: Larger length or higher #Predictions in response is not critical for better performance.

Ablations (Random Selection)

		ASQA					QAMPARI					
Ablation	**Method**	**#Clusters**	**length**	**Mauve**	**Str_EM**	**QA-F1**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
1	ASC	15.7	101.17	47.01	**44.1**	**32.22**	7.09	22.98	20.5	33.04	**19.46**	**26.21**
	Random Clusters	15.7	85.31	49.97	42.62	31.75	7.09	11.86	10.08	18.62	9.77	14.05
	Random Sentences	15.7	99.45	42.08	41.5	29.36	7.09	22.19	13.8	24.42	15.39	22.1
2	USC	-	64.52	40.19	39.05	30.88	8.97	20.7	19.21	31.28	18.07	24.2
	High Token/#Pred		82.93	40.59	37.8	28.79	10.48	17.19	18.3	29.28	16.07	21.01

Table 3: Ablation 1: ASC performs better than randomly picking clusters and randomly picking sentences on ASQA, QAMPARI. Ablation 2: Larger length or higher #Predictions in response is not critical for better performance.

Ablations (Random Selection)

		ASQA					QAMPARI					
Ablation	**Method**	**#Clusters**	**length**	**Mauve**	**Str_EM**	**QA-F1**	**#Pred**	**Prec**	**Rec**	**Rec-5**	**F1**	**F1-5**
1	ASC	15.7	101.17	47.01	**44.1**	**32.22**	7.09	22.98	20.5	33.04	**19.46**	**26.21**
	Random Clusters	15.7	85.31	49.97	42.62	31.75	7.09	11.86	10.08	18.62	9.77	14.05
	Random Sentences	15.7	99.45	42.08	41.5	29.36	7.09	22.19	13.8	24.42	15.39	22.1
2	USC		64.52	40.19	39.05	30.88	8.97	20.7	19.21	31.28	18.07	24.2
	High Token/#Pred		82.93	40.59	37.8	28.79	10.48	17.19	18.3	29.28	16.07	21.01

Table 3: Ablation 1: ASC performs better than randomly picking clusters and randomly picking sentences on ASQA, QAMPARI. Ablation 2: Larger length or higher #Predictions in response is not critical for better performance.

Ablations (Varying Θ)

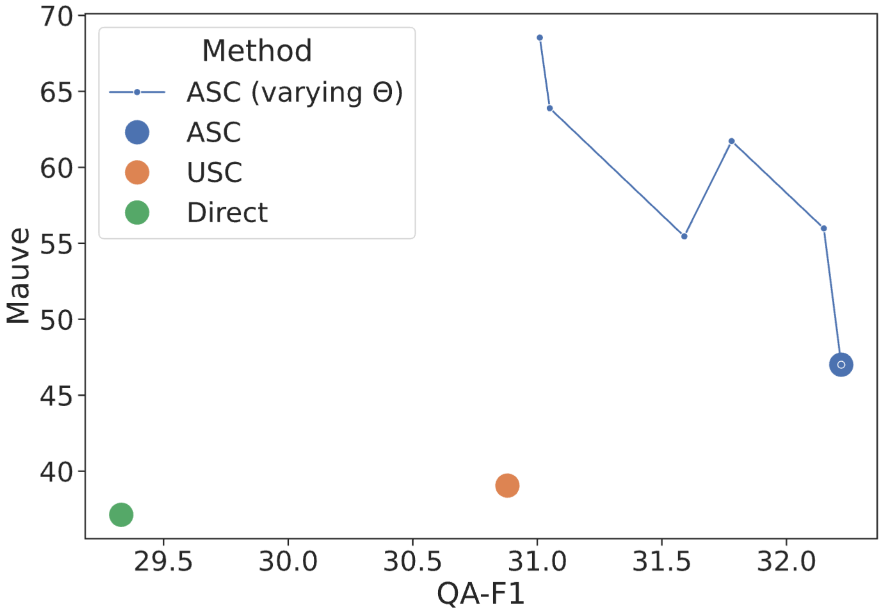


Figure 4: ASQA. Increasing Θ improves QA-F1, reduces Mauve. Adjusting Θ produces a preferred answer.

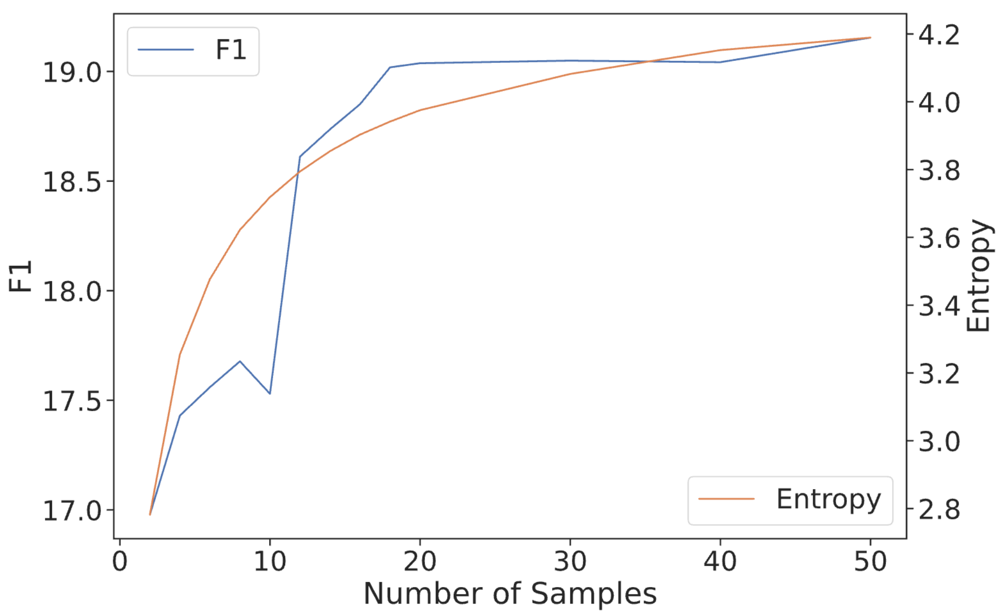


Figure 5: QAMPARI. Performance starts to stagnate when clusters' entropy stagnates.

Analysis (Room for Improvement)

		ASQA		*QAMPARI*	
Method	**#Gen**	**Str_EM**	**QA-F1**	**Rec**	**Rec-5**
Oracle	1	36.32	22.88	13.94	24.24
	2	40.64	28.05	18.15	30.46
	5	45.65	34.03	24.53	39.02
	15	50.97	39.28	32.29	48.78
	25	53.1	41.29	35.86	52.76
	50	56.09	45.2	40.06	56.90
ASC	50	44.1	32.22	20.50	33.04

Table 4: Oracle results reveal sizable scope for improvement using our approach of merging multiple responses.

Analysis (Room for Improvement)

		ASQA		*QAMPARI*	
Method	**#Gen**	**Str_EM**	**QA-F1**	**Rec**	**Rec-5**
Oracle	1	36.32	22.88	13.94	24.24
	2	40.64	28.05	18.15	30.46
	5	45.65	34.03	24.53	39.02
	15	50.97	39.28	32.29	48.78
	25	53.1	41.29	35.86	52.76
	50	56.09	45.2	40.06	56.90
ASC	50	44.1	32.22	20.50	33.04

Table 4: Oracle results reveal sizable scope for improvement using our approach of merging multiple responses.

References

- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation arXiv preprint arXiv:2311.17311
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.
- Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references?
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.
- Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Thank You

- Questions?