Atomic Self-Consistency for Better Long Form Generations

Raghuveer Thirukovalluru, Yukun Huang and **Bhuwan Dhingra**
Duke University
{raghuveer.thirukovalluru,yukun.huang}@duke.edu, bdhingra@cs.duke.edu

Abstract

Recent work has aimed to improve LLM generations by filtering out hallucinations, thereby improving the precision of the information in responses. Correctness of a long-form response, however, also depends on the recall of multiple pieces of information relevant to the question. In this paper, we introduce Atomic Self-Consistency (ASC), a technique for improving the recall of relevant information in an LLM response. ASC follows recent work, Universal Self-Consistency (USC) in using multiple stochastic samples from an LLM to improve the long-form response. Unlike USC which only focuses on selecting the best single generation, ASC picks authentic subparts from the samples and merges them into a superior composite answer. ASC demonstrates significant gains over USC on multiple factoids and open-ended QA datasets with ChatGPT and Llama2.

1 Introduction

Long-form question answering (LFQA) is an important benchmark task whose performance reflects the reliability of these AI systems at providing comprehensive and accurate responses to user queries. In LFQA, each response comprises multiple pieces of information, described in the literature as atomic facts (Min et al., 2023), that collectively contribute to the overall correctness of the answer. Despite various improvements, LLMs are still very prone to producing hallucinatory content such as incorrect atomic facts, especially when the responses are longer (Ren et al., 2023). Recent works on mitigating hallucinations have primarily involved the removal of inaccurate atomic facts from the generated content. While these methods produce responses with high precision over atomic facts, the correctness of the response also depends on the inclusion of all information relevant to the question, i.e., recall of atomic facts relevant to the question.

On the other hand, in QA with closed-form answers (such as a math problem with a numeric answer), remarkable improvements were made by stochastically sampling multiple model responses and then using *consistency* criteria to select one as the final answer (Wang et al., 2022). Recently, similar efforts were extended to long-form generation. Universal Self Consistency (USC) (Chen et al., 2023), is one example which uses LLMs to determine consistency between model responses. Output of USC is the single most consistent generation among multiple samples from the model.

However, picking a single final answer among the candidate generations might miss out on relevant atomic facts from other candidates and not optimize the recall of information. Further, it is still prone to some atomic hallucinations within the final selected candidate. To overcome these challenges, we propose a simple approach called Atomic Self-Consistency (ASC), which combines authentic atomic facts from multiple candidate responses to generate a superior composite response.

2 Methodology (ASC)

Given a question q, our task is to use an LLM $\mathcal{L}$ to produce an answer which answers the questions both accurately (with high precision) and comprehensively (with high recall). Let $a_1, a_2, .., a_m$ be m independent samples directly generated by $\mathcal{L}$ when query q is fed to it in a prompt.

2.1 Atomic Facts

Each generation to a question might comprise multiple sentences and multiple atomic facts within each sentence. Min et al. (2023) used an InstructGPT to break down longform generation a_i into its atomic facts. In our case, this would be extremely expensive as this needs to be performed for m different generations per question. In this work, we confined to the use of simple sentence tokenization models (Bird et al., 2009). In case of list style answers, directly use individual entires of the list as its atomic facts.

		ASQA					*ELI5*			
	Method	**#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
	USC		64.52	40.19	39.05	30.88		97.36	24.09	17.4
	ASC (Ours)	15.7	101.17	47.01	**44.1**	**32.22**	16.68	163.58	21.29	**21.43**

Table 1: ASQA, ELI5 results. ASC does the best on QA-F1 and demonstrates strong Str_EM. ASC also demonstrates strong Mauve. ASC achieve best Claims_Nli score on ELI5. Results justify that merging of samples is better.

		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
	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	7.09	22.98	20.5	33.04	**19.46**	**26.21**	8.44	12.47	10.41	19.15	**9.75**	**14.09**

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.

2.2 Clustering and Self Consistency

USC used an LLM to pick the most consistent of m responses. Such methods however cannot work with higher m values due to context length limitations of LLMs. Further, atomic facts are much higher in count ($>> m$). Hence, we perform clustering over all atomic facts. We use agglomerative clustering with sentence embeddings if the atomic facts are sentences and edit distance based clustering if the atomic facts are entities. Finally, we use strength of individual clusters to pick the most consistent of them. Specifically, all clusters having count above a fixed threshold Φ (tuned on a validation set) are left $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, ...\}$.

2.3 Final Answer

Given the clusters $\{\mathcal{C}_1, ...\}$, we select the longest sentence in a cluster as its representative atomic fact. In case when atomic facts are entities, we randomly select a representative from the cluster. Representative atomic facts from all clusters are finally combined using an LLM to generate an answer a. This only requires a single call to the LLM.

3 Experiments

We compare ASC with USC and Direct generation on four datasets - ASQA (Stelmakh et al., 2022), ELI5 (Fan et al., 2019), QAMPARI (Rubin et al., 2022) and QUEST (Malaviya et al., 2023). QAMPARI and QUEST are both long form answer datasets where answer is a list of entities. ASQA is a long form answer dataset where an answer is expected to contain all disambiguations of an ambiguous question. ELI5 contains how/why/what questions from Reddit. We use $m = 50$ for generations. Sentence embeddings by SimCSE (Gao et al., 2021), agglomerative clustering ($d = 0.2$) is used to perform clustering in ASQA. ChatGPT is the LLM used. Table 1 and 2 show results. Trends are similar with Llama-70b.

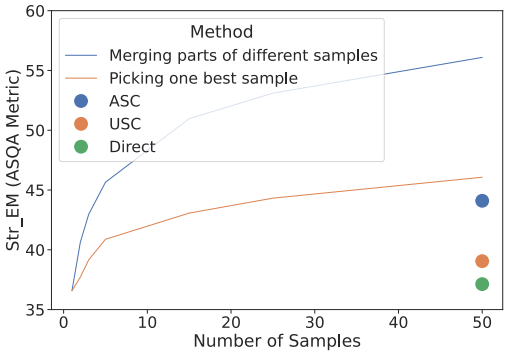

Figure 1: Best possible recall (oracle performance) with increasing number of samples on ASQA(ChatGPT). Merging subparts from multiple samples has a much higher ceiling. Significant potential still left over ASC as evident from the gap with oracle curve.

3.1 Ablation: Oracle Recall

To further show benefits of this approach, Fig. 1 shows the oracle performance (best possible performance) of picking one single sample vs merging multiple samples on ASQA dataset (using golden answers to pick samples). Merging answers from multiple samples have significant performance potential over picking a single answer. It also has significant potential remaining on top of ASC.

4 Conclusion

ASC combines atomic facts from multiple generations to produce a final answer with much higher degree of correctness. Experiments show that ASC has substantial gains over Direct, USC baselines. ASC can further be combined with Retrieval/self evaluation to further improve recall/correctness.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2023. Quest: A retrieval dataset of entity-seeking queries with implicit set operations. *arXiv preprint arXiv:2305.11694*.

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. *arXiv preprint arXiv:2305.14251*.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. *arXiv preprint arXiv:2312.09300*.

Samuel Joseph Amouyal Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. 2022. Qampari:: An open-domain question answering benchmark for questions with many answers from multiple paragraphs. *arXiv preprint arXiv:2205.12665*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. *arXiv preprint arXiv:2204.06092*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.