QUASI-RANDOM MULTI-SAMPLE INFERENCE FOR LARGE LANGUAGE MODELS

Avinash Amballa*, **Aditya Parashar***, **Aditya Vikram Singh***, **Jinlin Lai** , **Benjamin Rozonoyer**
College of Information & Computer Sciences, University of Massachusetts Amherst
{aamballa, aparashar, avsingh, jinlinlai, brozonoyer}@umass.edu

ABSTRACT

Large language models (LLMs) are often used with decoding strategies that require sampling multiple outputs. Vilnis et al. (2023) show that an LLM implicitly defines an arithmetic code book, facilitating efficient and embarrassingly parallelizable **arithmetic sampling** to produce multiple samples using quasi-random codes. Traditional text generation methods, such as beam search and sampling-based techniques, have notable limitations: they lack parallelizability or diversity of sampled sequences. This study explores the potential of arithmetic sampling, contrasting it with ancestral sampling across two decoding tasks that employ multi-sample inference: chain-of-thought reasoning with self-consistency and machine translation with minimum Bayes risk decoding. Our results demonstrate that arithmetic sampling produces more diverse samples, significantly improving reasoning and translation performance as the sample size increases. We observe a $\mathbf{3}$-$\mathbf{5}\%$ point increase in accuracy on the GSM8K dataset and a $\mathbf{0.45}$-$\mathbf{0.89}\%$ point increment in COMET score for WMT19 tasks using arithmetic sampling without any significant computational overhead.

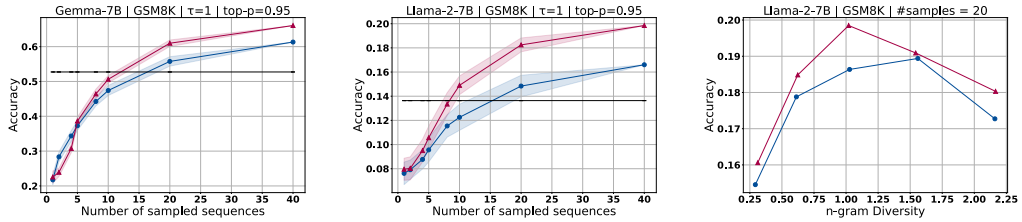
1 INTRODUCTION

There have been enormous efforts in improving the performance and efficiency of inference with large language models (Ippolito et al., 2019; Su & Collier, 2023; Grubisic et al., 2024; Zhou et al., 2024; Ding et al., 2024) based on system, data, and model level enhancements. In this paper, we consider that any decoding routine can be broadly assessed by its sample diversity (A.1) and parallelizability. Search-based techniques like beam search can approximate *maximum a posteriori* (MAP) decoding, mitigating duplicate samples at the expense of not being embarrassingly parallel. Sampling-based methods grounded in ancestral sampling techniques are parallel but don't explicitly guarantee diverse sequences. The recently proposed arithmetic sampling Vilnis et al. (2023) technique enables parallel inference with diverse samples – by interpreting the inference as sampling from code points from a unit interval, given code points generating sequences becomes embarrassingly parallel and the sample diversity is guaranteed by construction.

Decoding from pre-trained LLMs requires varying strategies for different downstream tasks. For complex reasoning and question-answering tasks, chain-of-thought (CoT) prompting Wei et al. (2022) is established for improving inference by instructing the model to generate intermediate reasoning paths. Wang et al. (2023) propose self-consistency as an additional improvement over chain-of-thought reasoning with multi-sample inference, attributable to diverse reasoning paths enhancing the confidence of the majority answer. For machine translation, minimum Bayes risk (MBR) decoding Kumar & Byrne (2004) is a classical approach for selecting the optimal translation from candidate translations generated by an LLM, requiring diversity to ensure performance.

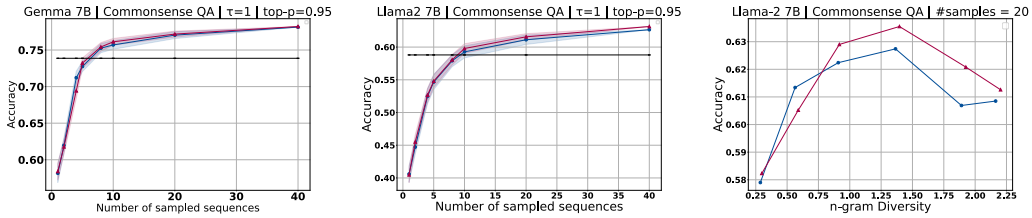
Thus, the inherent diversity of sequences generated via arithmetic sampling offers significant potential for enhancing decoding strategies that rely on multi-sample inference. Recognizing the importance of exploring this approach, we apply arithmetic sampling to both reasoning and translation tasks. For CoT reasoning with self-consistency and machine translation with MBR decoding, we observe accuracy improvements on the GSM8K and Commonsense QA datasets, along with substantial COMET score gains as the number of sampled sequences increases.

*indicates equal contribution



(a) Accuracy vs. # sampled sequences: Gemma-7B (b) Accuracy vs. # sampled sequences: Llama-2-7B (c) Accuracy vs. n-gram diversity: Llama-2-7B

Figure 1: 8-shot evaluation on GSM8K with Gemma-7B and Llama-2-7B



(a) Accuracy vs. # sampled sequences: Gemma-7B (b) Accuracy vs. # sampled sequences: Llama-2-7B (c) Accuracy vs. n-gram diversity: Llama-2-7B

Ancestral sampling | Arithmetic sampling | Greedy

Figure 2: 6-shot evaluation on Commonsense QA with Gemma-7B and Llama-2-7B

2 BACKGROUND

2.1 ARITHMETIC SAMPLING

Arithmetic sampling Vilnis et al. (2023) reinterprets the standard ancestral sampling process as lazily constructing an arithmetic codebook in the unit $[0, 1)$ interval where each code (point) is uniformly distributed and corresponds to a sequence from the output distribution.

This process ensures that the generated samples are diverse, as codes far apart in the code book usually correspond to different token prefixes. Moreover, it is embarrassingly parallel across the N samples, since the ith sample can be generated independently given its code c_i. Arithmetic sampling can also be applied orthogonally to other sampling-based techniques that directly manipulate the next token distribution, such as top-k, top-p (nucleus) Holtzman et al. (2020), temperature sampling and epsilon sampling Hewitt et al. (2022).

2.2 SELF-CONSISTENCY

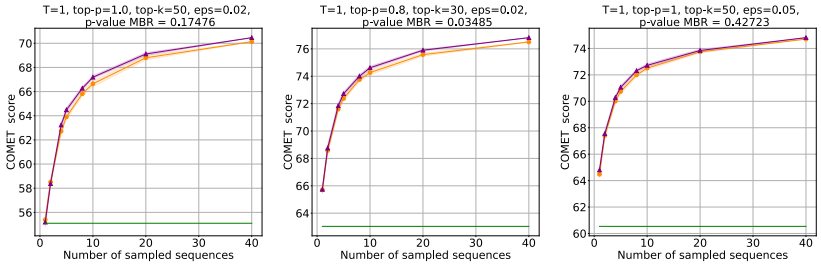
Self-consistency Wang et al. (2023) is a method designed to improve the performance of chain-of-thought (CoT) reasoning by generating multiple reasoning paths with answers for a given prompt and then selecting the most consistent answer based on, generally, majority voting. This approach leverages the diversity of the generated candidate reasoning paths to identify the most frequent outcome, thereby enhancing the accuracy of the final answer.

2.3 MBR DECODING

Minimum Bayes risk (MBR) Eikema & Aziz (2020) is based on the principle of maximizing the expected utility of a given hypothesis. When making predictions, we lack information about the ideal (target) translations and must make decisions under uncertainty. MBR allows the model to probabilistically estimate ideal decisions as it searches for the candidate that maximizes expected utility. We used COMET Rei et al. (2020) as the utility metric.

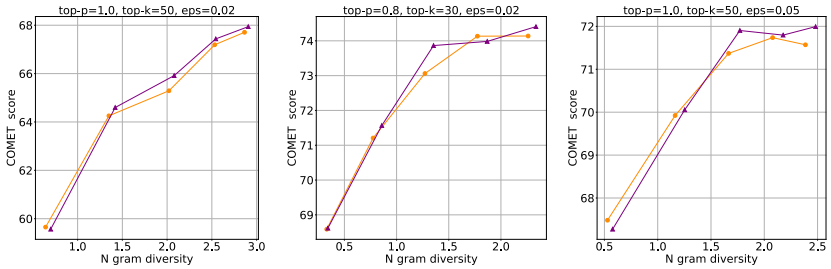
We use the sampling-based approximation to MBR decoding as posited in Eikema & Aziz (2022), using the Monte Carlo estimate and formulating the candidate space from the generated N (pseudo-reference) samples:

$$y^{\text{N-by-N}} = \underset{h \in \{y^{(1)}, y^{(2)}, \ldots y^{(N)}\}}{\arg\max} \frac{1}{N} \sum_{n=1}^{N} u(y^{(n)}, h)$$



(a) De-En task: Flan-T5 (b) De-En task: MT0 (c) Ru-En task: MT0

Figure 3: COMET vs. #sampled sequences on Flan T5, MT0



(a) De-En task: Flan-T5 (b) De-En task: MT0 (c) Ru-En: MT0

Ancestral sampling (MBR) Arithmetic sampling (MBR) Greedy

Figure 4: COMET vs. n-gram diversity on Flan-T5, MT0 varying temperature T

3 Experiments

3.1 Datasets

We conduct experiments with only validation splits of datasets aimed at chain-of-thought reasoning and neural machine translation. We use **GSM8K** Cobbe et al. (2021) and **CommonsenseQA** Talmor et al. (2019) for chain-of-thought self-consistency; and **WMT19 De – En (German-English)** and **Ru-En (Russian-English)** Barrault et al. (2019) for MBR decoding machine translation experiments.

3.2 Models and Baselines

In our experiments, we employ Gemma-7B Team et al. (2024), and Llama-2-7B Touvron et al. (2023) models for CoT self-consistency; and MT0 Muennighoff et al. (2023) for De-En and Ru-En tasks, Flan-T5 Chung et al. (2022) for De-En task with MBR decoding.

In addition to greedy decoding, we adopt ancestral sampling as our baseline method to compare with arithmetic sampling for sampling reasoning paths in self-consistency and pseudo-reference translations in MBR decoding. We include additional sampling strategies such as temperature, top-k, nucleus, and epsilon sampling which are orthogonal to ancestral and arithmetic sampling.

3.3 Results

Subsampling We perform subsampling for all the models and datasets to estimate performance and variance for sampled sequences less than a given N which is 40 for most of our results. For both ancestral and arithmetic sampling, we averaged over multiple runs, sampling sequence sets of lengths $\{d \in \mathcal{N} \text{ such that } d \,|\, N\}$ from a pool generated by runs with n samples. For arithmetic sampling, we randomly select an offset for each dataset instance and pick indices in fixed intervals to simulate selecting quasi-random samples from a codebook. For ancestral sampling, we randomly sample from

all elements in d with replacement. This allowed us to run an experiment for 40 sampled sequences and obtain results for $d \in \{1, 2, 4, 5, 8, 10, 20\}$ sequences.

3.3.1 SELF-CONSISTENCY

We primarily evaluate the performance of the aforementioned sampling techniques with parameters controlled as temperature $= 1$, top-$k = 40$, and top-$p = 0.95$ across 40 sampled reasoning paths per question-answer pair, using the subsampling results to plot the standard deviation as the shaded region in Figures 1 and 2. We select the majority-voting accuracy for self-consistency as our performance evaluation metric. We also analyze accuracy performance across varying n-gram diversity (controlled by setting temperatures from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$).

Arithmetic reasoning For GSM8K, which consists of open-ended integral answers for grade-school math problems, we performed an 8-shot evaluation as presented in the literature by Kojima et al. (2023). From Figure 1, we observe significant performance gain from arithmetic sampling over ancestral sampling at 40 sampled sequences: $+\mathbf{4.75}\%$ for Gemma-7B in Figure 1 (a) and $+\mathbf{3.25}\%$ for Llama-2-7B in Figure 1 (b). The performance v/s diversity analysis from Figure 1 (c), shows that as diversity increases, the accuracy increases initially and declines after reaching the maxima at some intermediate diversity level, where arithmetic sampling yields better accuracy than ancestral sampling for all diversity levels.

Commonsense reasoning Commonsense QA, a multiple-choice question-answering dataset that necessitates various types of commonsense knowledge to predict correct answers, was utilized for experiments on commonsense reasoning. Similar to arithmetic reasoning, we observe performance gains with arithmetic sampling, as shown in Figure 2. With Gemma-7B, arithmetic sampling initially trails behind ancestral sampling; however, as the number of sampled sequences surpasses 5, it demonstrates better performance, with a $\mathbf{0.1}\%$ improvement at 40 sampled sequences in Figure 2(a). In Figure 2 (b), for Llama-2-7B, arithmetic sampling consistently outperforms ancestral sampling across the number of sampled sequences, showing a $\mathbf{1.05}\%$ increase over ancestral at 40 sampled sequences. Conducting performance v/s diversity analysis on Llama-2-7B for 20 sampled reasoning paths with top-p set to 1, as seen in Figure 2 (c), we find that accuracy improves for both sampling methods, with arithmetic reaching a higher global maximum than ancestral and subsequently dominating in performance.

3.3.2 MACHINE TRANSLATION WITH MBR

We ran all the experiments on 1000 machine translation examples with different parameters of temperature $T = \{0.1, 1\}$, top-$p = \{0.8, 1\}$, top-$k = \{30, 50\}$, epsilon cutoff $eps = \{0.02, 0.05\}$. We selected COMET Rei et al. (2020); Freitag et al. (2022) as the evaluation metric. We use subsampling to report the mean COMET score across all the data instances versus the number of sampled sequences = $\{1, 2, 4, 5, 8, 10, 20, 40\}$, plotting the standard deviation as the shaded region in Figure 3. All the experiments are performed in a zero-shot setting with a simple prompt "`Translate the following German/Russian sentence to an English sentence.`"

Figures 3 (a) - (c) show the COMET score of Flan-T5 De-En, MT0 De-En and MT0 Ru-En tasks respectively. Across all COMET score graphs presented (refer Appendix Figure 5 for more plots), it is evident that arithmetic sampling consistently outperforms ancestral sampling as the number of sampled sequences increases. Figures 3 (a), 3 (b), 3 (c) show that arithmetic has **0.89**%, **0.48**%, **0.45**% increase in the mean COMET scores over ancestral at 5 sampled sequences (see Appendix Figure 6) without any significant computational overhead Vilnis et al. (2023).

We conduct the COMET score vs. n-gram diversity analysis on above tasks by varying temperature $T = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and plotting the COMET score for 20 sampled sequences (4). As the temperature increases, the diversity of both the sampling techniques increases, thus we chose a higher temperature value of $T = 1$ for reporting the results in Figure 3. The COMET score values for arithmetic outperform those of ancestral, as arithmetic demonstrates greater diversity (4).

We also observe that arithmetic sampling has a lower spread than ancestral (Appendix Figure 6). In addition, we conduct the paired t-test at 40 sampled sequences to compare the arithmetic and ancestral

COMET scores across 1000 datapoints (Figure 3). The low p-values indicate that the difference in the mean COMET scores of arithmetic and ancestral sampling is statistically significant.

3.4 Discussion

We observe that GSM8K demonstrates greater performance gains compared to Commonsense QA and Machine Translation tasks. This is attributed to nature of the tasks. GSM8K involves open-ended reasoning, where there are no predefined options, and diverse reasoning paths can significantly improve performance by exploring a broader solution space. Arithmetic sampling encourages this diversity, enabling the model to better capture complex reasoning patterns.

In contrast, tasks like Commonsense QA or Machine Translation are more constrained. Commonsense QA has only a limited set of options (e.g., 5 choices), reducing the likelihood of errors and the potential impact of diversity. Similarly, in Machine Translation, translations are typically close to the target sentence, leaving little room for diversity to improve outcomes significantly. This explains why arithmetic sampling shows limited gains in these tasks compared to GSM8K.

4 Conclusions

This work investigated the application of arithmetic sampling, a diverse parallelizable sampling technique for multi-sample inference from large language models (LLMs), in chain-of-thought reasoning with self-consistency and machine translation with minimum Bayes risk (MBR) decoding. Experiments on arithmetic reasoning and commonsense reasoning demonstrated that arithmetic sampling improves performance as the number of sampled sequences increases, yielding 3-5% and 1.05% point increases in accuracy, respectively, over ancestral sampling at 40 sampled sequences. In machine translation, arithmetic sampling consistently outperformed ancestral sampling, with improvements ranging from 0.45% to 0.89% in mean COMET scores for 5 sampled sequences across different language pairs and parameter settings. Arithmetic sampling also exhibited a lower standard deviation compared to ancestral sampling. The superior performance of arithmetic sampling with multi-sample inference based decoding methods like self-consistency and MBR decoding is due to its ability to generate more diverse samples, enhancing LLM performance in tasks like chain-of-thought reasoning and machine translation without any computational overhead.

5 Limitations

1. **Limited to random vocabulary ordering:** As the unit interval associated with arithmetic sampling corresponds to an ordering of the vocabulary tokens (from which sequences are decoded via the code points), we limited ourselves to randomizing the vocabulary ordering and then performing arithmetic sampling. At the expense of some constant overhead, creating an ordering of the vocabulary that follows some measure of semantic similarity may show greater diversity (and potentially, better performance) as the uniform lattice of code points would now be corresponding to semantically dissimilar vocabulary tokens more often – rendering this idea an interesting future direction.
2. **Limited to one dimension:** In arithmetic sampling, tokens are limited in one dimension space with softmax sampling. It may be beneficial to consider higher dimensional embeddings of the tokens with box embeddings Vilnis et al. (2018). Diverse sampling in high dimensional embedding spaces may involve advanced sampling techniques like Quasi-Monte Carlo Caflisch (1998).
3. **Difference between sampling and maximizing:** Arithmetic sampling was formulated to generate parallel diverse sequences, which differs from strategies that find the most probable answer like beam search. Our application of arithmetic sampling is limited to multi-sample decoding strategies that require diversity. It may be an interesting future work to extend arithmetic sampling for MAP decoding.

REFERENCES

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Russel E Caflisch. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 7:1–49, 1998.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. The efficiency spectrum of large language models: An algorithmic survey, 2024.

Bryan Eikema and Wilker Aziz. Is map decoding all you need? the inadequacy of the mode in neural machine translation, 2020.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.754. URL https://aclanthology.org/2022.emnlp-main.754.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.2.

Dejan Grubisic, Chris Cummins, Volker Seeker, and Hugh Leather. Priority sampling of large language models for compilers, 2024.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models, 2019.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

Shankar Kumar and Bill Byrne. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, 2004.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2023.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation, 2020. URL https://arxiv.org/abs/2009.09025.

Yixuan Su and Nigel Collier. Contrastive search is what you need for neural text generation, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 263–272. Association for Computational Linguistics, 2018.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Passos, and Sumit Sanghai. Arithmetic sampling: Parallel diverse decoding for large language models, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. A survey on efficient inference for large language models, 2024.

A APPENDIX

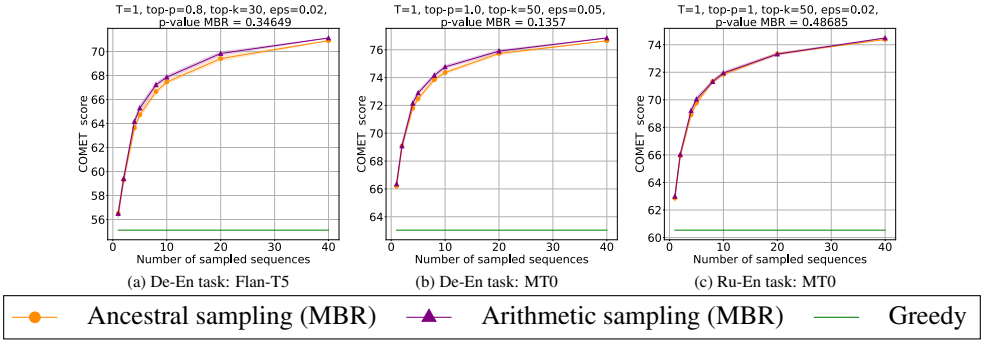
A.1 DEFINITION OF SAMPLE DIVERSITY

The n-gram diversity score is defined as $d = \sum_{n=1}^{4} d_n$ where

$d_n = \frac{\text{\# of unique } n\text{-grams in } N \text{ translations}}{\text{total \# of } n\text{-grams in } N \text{ translations}}$.

A.2 MACHINE TRANSLATION WITH MBR

In this section, we present more plots on De-En, Ru-En task for Flan-t5 and MT0 models across different parameter setting as shown in the Figure 5.



(a) De-En task: Flan-T5 (b) De-En task: MT0 (c) Ru-En task: MT0

Figure 5: COMET vs. #sampled sequences on Flan-T5, MT0

In addition, we also report the mean and standard deviation of performance (COMET) for arithmetic and ancestral sampling for the best performing parameter setting in Figure 6.

#Sampled Seq	Arithmetic	Ancestral
1	55.19 ± 0.35	55.39 ± 0.36
2	58.38 ± 0.21	58.51 ± 0.29
4	63.24 ± 0.19	62.72 ± 0.33
5	64.48 ± 0.18	63.91 ± 0.21
8	66.27 ± 0.14	65.82 ± 0.21
10	67.19 ± 0.12	66.64 ± 0.24
20	69.11 ± 0.12	68.80 ± 0.18
40	70.46 ± 0.00	70.13 ± 0.00

(a) De-En task: Flan-T5 (T = 1, top-p = 1.0, top-k = 50, eps = 0.02)

#Sampled Seq	Arithmetic	Ancestral
1	65.74 ± 0.28	65.71 ± 0.18
2	68.75 ± 0.25	68.58 ± 0.22
4	71.85 ± 0.22	71.59 ± 0.19
5	72.72 ± 0.13	72.37 ± 0.15
8	74.00 ± 0.13	73.75 ± 0.13
10	74.62 ± 0.11	74.27 ± 0.16
20	75.90 ± 0.07	75.57 ± 0.12
40	76.82 ± 0.00	76.50 ± 0.00

(b) De-En task: MT0 (T = 1, top-p = 0.8, top-k = 30, eps = 0.02)

#Sampled Seq	Arithmetic	Ancestral
1	64.80 ± 0.23	64.48 ± 0.23
2	67.55 ± 0.19	67.41 ± 0.27
4	70.29 ± 0.14	70.01 ± 0.13
5	71.06 ± 0.12	70.74 ± 0.18
8	72.32 ± 0.10	71.99 ± 0.14
10	72.72 ± 0.11	72.51 ± 0.12
20	73.85 ± 0.09	73.74 ± 0.08
40	74.82 ± 0.00	74.71 ± 0.00

(c) Ru-En task: MT0 (T = 1, top-p = 1.0, top-k = 50, eps = 0.05)

Figure 6: Tables showing mean$\pm$std of mean COMET for 20 runs using arithmetic and ancestral sampling, based on the number of sampled sequences