Advanced Machine Learning

April 2, 2025

Spring 2025

CS 726: Programming assignment **Submitted by:** Pinak Mahapatra, Danish Siddiqui and Aansh Samyani

Contents

1 Problem Statement

In this assignment, we will be working with an introduction to LLM decoding techniques, word-constrained decoding, and a concept called "Staring into Medusa's Heads."

2 Introduction to LLM Decoding Techniques

This section was designed to get us familiar with the decoding process in Large Language Models (LLMs) and how different sampling techniques impact its text generation. Our task was to implement and analyze the following decoding strategies on Llama-2 when evaluated on Hindi to English translation task with IN22-Gen dataset using relevant metrics.

2.1 Greedy LLM Sampling Results

At every step, you simply pick the token with the highest probability from the LLM's output distribution. Formally, at the tth step, you obtain the next token as follows:

$$y_t = \arg\max_w P(w \mid y_{1:t-1}, x)$$

where $y_{1:t-1}$ denotes previously generated tokens and x is the input prompt. This process is repeated iteratively until the end-of-sentence (EOS) token is generated.

2.1.1 Evaluation Mterics

- BLEU Score: 0.3099861303744798
- ROUGE-1: 0.35427226007345536
- ROUGE-2: 0.12963403848897576
- ROUGE-LCS: 0.27101764007551854

2.1.2 Translation Examples

Example 1

Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

मोटे तौर पर, आय मान्यता की नीति व्यक्तिनिष्ठ कारणों पर आधारित होने के बजाय वस्तुनिष्ठ और वसूली के विवरण पर आधारित होनी चाहिए।

Figure 1: Example 1

Completion:

- **Reference:** in macroeconomics the subject is typically a nation and how all markets interact to generate big phenomena that economists call aggregate variables
- **Ground Truth:** in economics a country is a distinct entity in which the market is affected by one another in a certain way which is called the economic cycle

Example 2 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

मोटे तौर पर, आय मान्यता की नीति व्यक्तिनिष्ठ कारणों पर आधारित होने के बजाय वस्तुनिष्ठ और वसूली के विवरण पर आधारित होनी चाहिए।

Figure 2: Example 2

Completion:

- **Reference:** liquidity refers to the extent to which financial assets can be sold at close to full market value at short notice
- **Ground Truth:** chal nidhi ka arth hai kiis had tak alpkal me pure bazar bhav se bahut kam antar par vittiya sanpatti becchi ja sakati hai

2.2 Random Sampling with Temperature Scaling

Instead of always selecting the most probable token, here we randomly sample from the probability distribution while adjusting its sharpness using a temperature parameter τ. That is, first, we modify the probabilities as follows:

$$P'(w \mid y_{1:t-1}, x) = \frac{P(w \mid y_{1:t-1}, x)^{\frac{1}{\tau}}}{\sum_{w' \in V} P(w' \mid y_{1:t-1}, x)^{\frac{1}{\tau}}}$$

A token is then randomly sampled from P'. Like before, keep repeating this process until the EOS token is generated. Here, you must experiment with $\tau \in \{0.5, 0.9\}$ and report your findings.

2.2.1 Evaluation Metrics at $\tau = 0.5$

- BLEU: 0.31428571428571433
- ROUGE-1: 0.33972420543105697
- ROUGE-2: 0.13453839042358834
- ROUGE-LCS: 0.27320559361364677

Example 1 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

सेवा संबंधी लोगों के लिए भेष कई गुणों का संयोजन है, जैसे कि उनके जूते, कपड़े, टाई, आभूषण, केश शैली, मेक-अप, घड़ी, कॉस्मेटिक, इत्र, आदि।

Figure 3: Example 1

Translation Results:

- **Reference:** an appearance is a bunch of attributes related to the service person like their shoes clothes tie jewellery hairstyle makeup watch cosmetics perfume etc
- **Ground Truth:** service is the coordination of many qualities for people such as their shoes clothes ties accessories makeup hairstyle cosmetics etc
- **Comment:** your sentence is not valid

Example 2 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

महाराष्ट्र के औरंगाबाद जिले में स्थित अजंता में उन्तीस चैत्य और विहार गुफाएँ हैं जो पहली शताब्दी ई.पू. से ले कर पाँचवीं शताब्दी ईस्वी तक की मूर्तियों तथा चित्रकारियों से सुसज्जित हैं।

Figure 4: Example 4

Translation Results:

- **Reference:** ajanta located in the aurangabad district of maharashtra has twentynine caitya and vihara caves decorated with sculptures and paintings from the first century bce
- **Ground Truth:** maharashtras aurangabad district in the state of uttar pradesh has 35 caves and 23 viharas that date back to the first century ad the caves are known

2.2.2 Evaluation Metrics $\tau = 0.9$

- BLEU: 0.19441944194419442
- ROUGE-1: 0.15688305019988685
- ROUGE-2: 0.049388172391750744
- ROUGE-LCS: 0.12356232704785278

Example 1 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

अशोक ने व्यापक रूप से मूर्तियों और शानदार स्मारकों को बनाने के लिए पत्थर का प्रयोग करना शुरू किया, जबकि उससे पहले पारंपरिक रूप से लकड़ी और मिट्टी का प्रयोग किया जाता है।

Figure 5: Example 1

Translation Results:

- **Reference:** ashoka started making extensive use of stone for sculptures and great monuments whereas the previous tradition consisted of working with wood and clay
- **Ground Truth:** the ashook performed the carving of the stone and the beautiful sculptures for the first time although previously traditional wood and clay was used for the purpose

2.3 Variation of BLEU Score with τ

The BLEU score is an essential evaluation metric for translation quality. In this section, we analyze how the BLEU score changes with different values of τ, which represents the level of randomness in token selection.

2.3.1 Observations

- As τ increases, the BLEU score decreases.
- This suggests that introducing more randomness in the decoding process reduces the accuracy of the generated translations.
- At $\tau = 0.5$, we observe the highest BLEU score of 0.3143, indicating relatively high translation accuracy.
- As τ increases to 0.9, the BLEU score drops significantly to 0.1944, suggesting a decline in translation precision.
- This trend highlights the trade-off between deterministic decoding (low τ) and diverse generation (high τ).

2.3.2 BLEU Scores at Different τ Values

τ	BLEU Score
0.5	0.3143
0.6	0.3083
0.7	0.2732
0.8	0.2656
0.9	0.1944

Table 1: BLEU Score Variation with τ

2.3.3 Plot of BLEU Score vs τ

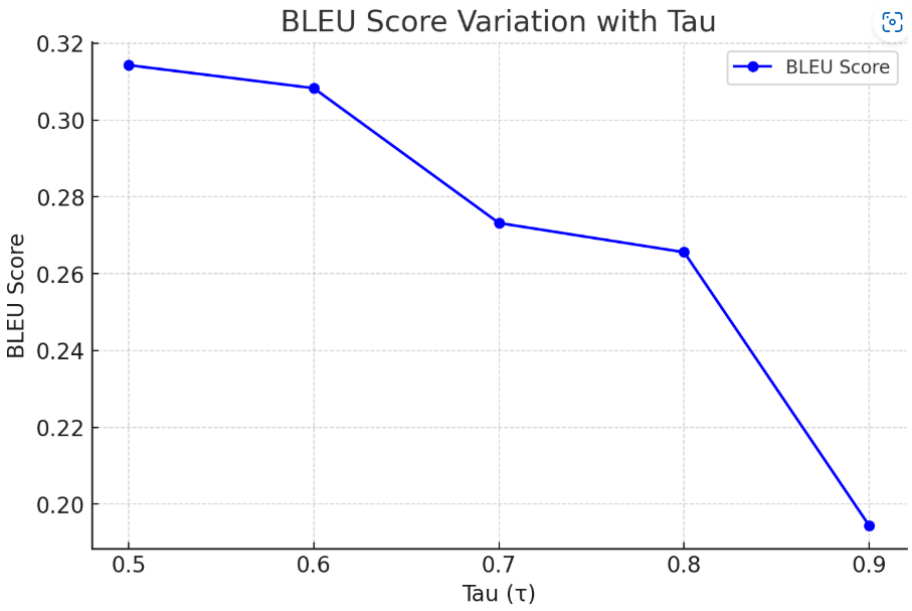

Figure 6: Plot of Blue scores

The plot in Figure **??** visually confirms the trend of decreasing BLEU scores as τ increases. This suggests that while increasing τ allows for more diverse outputs, it comes at the cost of reduced translation fidelity.

2.4 Top-K Sampling

Rather than sampling from the entire vocabulary, in Top-k sampling, we restrict our choices to the k most probable tokens. To do this, first, we sort the vocabulary by probability and keep only the top k tokens:

$$V_k = \{w_1, w_2, \ldots, w_k\}, \quad \text{where} \quad P(w_i) \geq P(w_{i+1}) \quad \text{for} \quad i < k$$

The probabilities within V_k are then normalized as follows:

$$P'(w) = \begin{cases} \frac{P(w)}{\sum_{w' \in V_k} P(w')} & \text{if } w \in V_k \\ 0 & \text{otherwise} \end{cases}$$

A token is then randomly sampled from P'. As before, repeat the process until the EOS token is generated. Here, experiment with $k \in \{5, 10\}$ and report your findings.

2.4.1 Evaluation Metrics at $k = 5$

- BLEU: 0.30437903804737976
- ROUGE-1: 0.3441461065458107
- ROUGE-2: 0.12217469017065748
- ROUGE-LCS: 0.26448142335633174

Example 1 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

महाराष्ट्र के औरंगाबाद जिले में स्थित अजंता में उन्तीस चैत्य और विहार गुफाएँ हैं जो पहली शताब्दी ई.पू. से ले कर पाँचवीं शताब्दी ईस्वी तक की मूर्तियों तथा चित्रकारियों से सुसज्जित हैं।

Figure 7: Example 1

Translation Results:

- **Reference:** ajanta located in the aurangabad district of maharashtra has twentynine caitya and vihara caves decorated with sculptures and paintings from the first century bce
- **Ground Truth:** maharashtras aurangabad district has 30 caves and viharas in the ajanta caves that are the first from the 2nd century bc to the fifth century ad which are

2.4.2 Evaluation Metrics at $k = 10$

- BLEU: 0.292021688613478
- ROUGE-1: 0.304920781920225
- ROUGE-2: 0.09839440147559848
- ROUGE-LCS: 0.24102308933558741

Example 1 Input Prompt:

<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:

महाराष्ट्र के इस स्वादिष्ट और प्रसिद्ध व्यंजन में आलुओं में मसाले को मिलाकर, बेसन के घोल की परत लगाकर, उसे अच्छी तरह से तल कर बनाया जाता है।

Figure 8: Example 1

Translation Results:

- **Reference:** potatoes mixed in masalas coated in besan batter and deep fried to perfection form this delicious and famous dish of maharashtra
- **Ground Truth:** in maharashtra the famous and delicious curry is prepared by mixing the masala in the onion the leaves of the basil plant and the onion is prepared in a good way

2.4.3 Inference from Top-K Sampling

- As k increases, the BLEU score does not follow a strictly increasing or decreasing trend but fluctuates.
- The highest BLEU score (0.3103) is observed at $k = 9$, suggesting an optimal balance between controlled selection and diversity.
- The BLEU score is lowest at $k = 8$ (0.2806), indicating that excessive vocabulary restriction can harm translation quality.
- ROUGE scores also exhibit similar fluctuations, with the best overall performance at $k = 9$.
- Overall, a moderate k (not too low or too high) provides the best balance between diversity and accuracy.

2.4.4 BLEU Scores for Different k Values

k	BLEU Score
5	0.3044
6	0.2959
7	0.2950
8	0.2806
9	0.3103
10	0.2920

Table 2: BLEU Score Variation with k

2.4.5 Plot of BLEU Score vs k

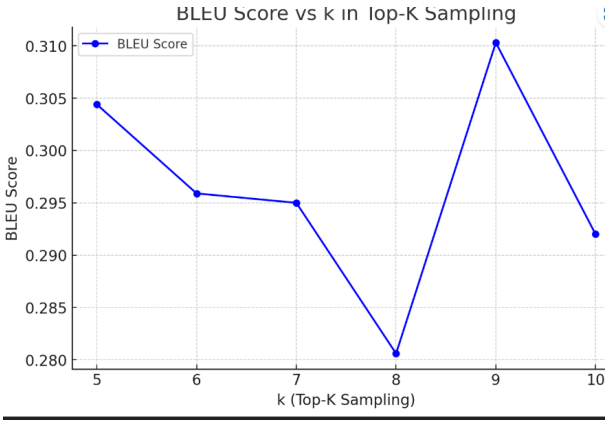

Figure 9: Plot of Variation of Blue score

The plot in Figure **??** shows the variation of BLEU scores with different values of k. The trend suggests that an optimal k (around 9) yields the best translation quality, while very low or very high values may lead to suboptimal performance.

2.5 Nucleus Sampling

Instead of picking a fixed number of tokens as in Top-k Sampling, here we dynamically choose the smallest set of tokens whose cumulative probability exceeds a threshold p:

$$V_p = \{w_1, w_2, ..., w_m\}, \quad \text{such that} \quad \sum_{i=1}^{m} P(w_i) \geq p$$

We then normalize probabilities over this set and sampling occurs as follows:

$$P'(w) = \begin{cases} \frac{P(w)}{\sum_{w' \in V_p} P(w')} & \text{if } w \in V_p \\ 0 & \text{otherwise} \end{cases}$$

Similar to before, we repeat the process until the EOS token is generated. Here, experiment with $p \in \{0.5, 0.9\}$ and report your findings.

2.5.1 P = 0.5

- BLEU: 0.3167520117044623

- ROUGE-1: 0.3465546104497868
- ROUGE-2: 0.12836815346391814
- ROUGE-LCS: 0.27175564185690826

Example 1

- **Input Prompt:**

सेवा संबंधी लोगों के लिए भेष कई गुणों का संयोजन है, जैसे कि उनके जूते, कपड़े, टाई, आभूषण, केश शैली, मेक-अप, घड़ी, कॉस्मेटिक, इत्र, आदि।

Figure 10: Example 1

- **Completion:**

  Reference: an appearance is a bunch of attributes related to the service person like their shoes clothes tie jewellery hairstyle makeup watch cosmetics perfume etc.
  Ground Truth: service is the combination of many qualities for people such as their clothes shoes ties accessories makeup hairstyle cosmetics etc.

2.5.2 P = 0.9

- BLEU: 0.31149927219796214
- ROUGE-1: 0.33882255547285145
- ROUGE-2: 0.12336814636044396
- ROUGE-LCS: 0.2569024897849652

Example: 1

- **Input Prompt:**

महाराष्ट्र के इस स्वादिष्ट और प्रसिद्ध व्यंजन में आलुओं में मसाले को मिलाकर, बेसन के घोल की परत लगाकर, उसे अच्छी तरह से तल कर बनाया जाता है।

Figure 11: Example 5

- **Completion:**

  Reference: Potatoes mixed in masalas, coated in besan batter, and deep fried to perfection form this delicious and famous dish of Maharashtra.
  Ground Truth: The famous and delicious dish of Maharashtra, in which the spices are mixed with onions, is cooked by putting the leaves of basil on it, and it is cooked well.

2.6 Inference on BLEU Score Variation with Different Values of P

From the BLEU score plot, we observe the following key trends:

1. **Optimal BLEU Score Around $P = 0.6$:** The highest BLEU score of **0.3224** is achieved at $P = 0.6$, indicating that this level of randomness in nucleus sampling produces the most accurate translations. This suggests that a moderate level of probability mass selection balances fluency and accuracy effectively.

2. **Performance Decreases at Higher P:** As P increases beyond 0.6 (towards 0.7 and 0.9), BLEU scores show a decline. This suggests that increasing randomness leads to more diverse but less precise translations, making the outputs deviate from the reference translations. Higher values of P allow for more unpredictable choices, sometimes leading to less faithful translations.

3. **Lower P Also Limits Performance:** At $P = 0.5$, the BLEU score is slightly lower than at $P = 0.6$, indicating that excessive restriction on token selection may cause the model to generate overly deterministic outputs. This can reduce diversity and prevent the model from capturing variations that might still be correct translations.

Conclusion: The results suggest that $P = 0.6$ provides the best balance between diversity and accuracy in translation tasks. Lower values restrict the model too much, while higher values introduce excessive randomness, leading to a decline in performance.

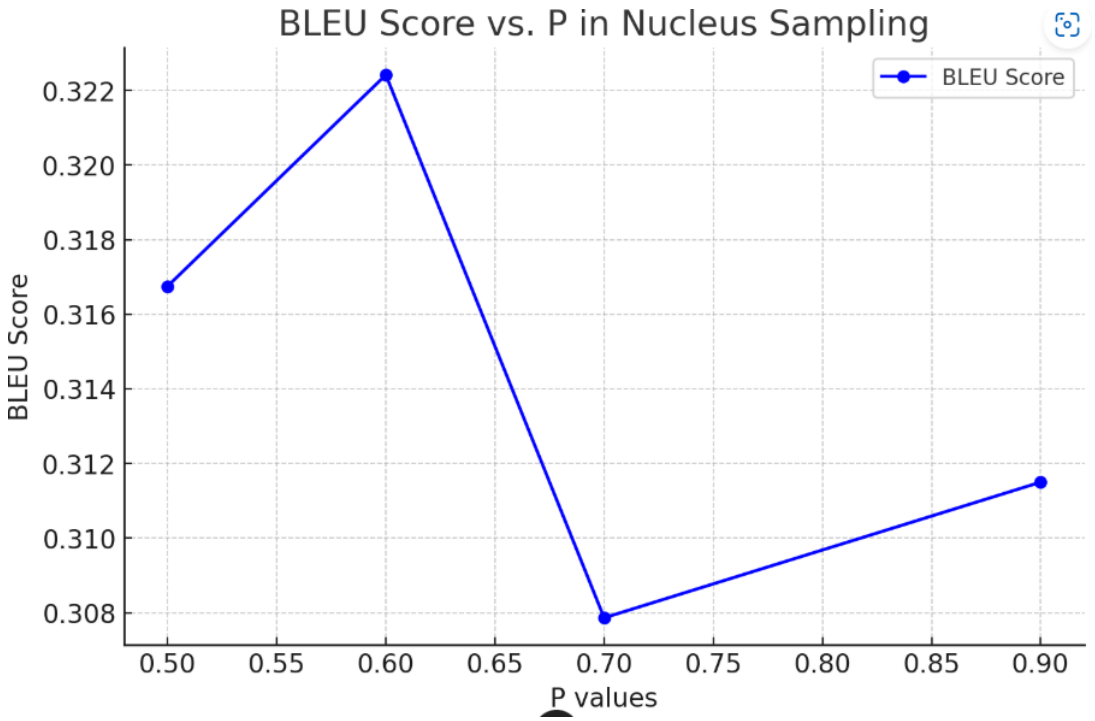

Figure 12: Plot of blue score variation

3 Word-Constrained Decoding

In this section, we will implement a variant of Grammar-Constrained Decoding called Word-Constrained Decoding. Let us assume that there is an oracle that has magically provided you, for every test example, a bag of all the words that appear in its output. Our task is to design a greedy decoding technique that takes advantage of this additional word list and improve the LLM performance.

Algorithm 1 Trie-based Word-Constrained Decoding

```
function ConstructTrie(word_list, tokenizer)
    Initialize root node
    for each word in word_list do
        Convert word to tokenized form
        Insert token sequence into Trie
    end for
    return Trie
end function

function ApplyConstraints(input_ids, scores, Trie)
    Extract last generated token
    if token sequence exists in Trie then
        Allow only valid next tokens
        Mask invalid tokens in scores
    else
        Boost probability of valid first tokens
    end if
    return updated scores
end function

function GenerateText(model, tokenizer, input_ids, word_list, eos_id, max_output_len)
    Construct Trie from word_list
    Initialize LogitsProcessor with Trie constraints
    Generate output using model.generate()
    Extract generated tokens (excluding input)
    return generated tokens
end function
```

For word-constrained decoding, we employed a **token-based Trie** to enforce the model to generate words exclusively from the given list. Initially, we experimented with a **character-based Trie**, but synchronizing it with the token generation process of the model proved challenging.

In the token-based Trie approach, each word from the list was stored in its tokenized form. Our primary idea was to allow the model to predict the **first token freely**, after which we would mask all tokens except those belonging to valid words in the Trie. Additionally, we reduced the scores of other tokens to constrain the generation.

However, the model struggled to predict the first token of words from the list. To address this, we introduced a boosting mechanism, increasing the probability of valid first tokens. The boost factor was a tunable hyperparameter, and after extensive experimentation, we determined that 5.6 was the optimal value, yielding excellent results.

3.1 Example:1

Example Details

Word List: ['policy', 'subjective', 'should', 'rather', 'recovery,', 'be', 'record', 'Broadly,', 'objective', 'based', 'any', 'considerations.', 'and', 'on', 'of', 'income', 'than', 'the', 'recognition']

Input Prompt: `<s> You are an AI assistant whose purpose is to perform translation. Given the following sentence in Hindi, translate it to English:`

मोटे तौर पर, आय मान्यता की नीति व्यक्तिनिष्ठ कारणों पर आधारित होने के बजाय वस्तुनिष्ठ और वसूली के विवरण पर आधारित होनी चाहिए।

Figure 13: Example 1

Completion:
Reference: `broadly the policy of income recognition should be objective and based on the record of recovery rather than on any subjective considerations`
Ground Truth: `on the surface it may seem that the reason for the individuals policy is based on their own subjective and personal considerations rather than being based on any objective or rational considerations however on closer inspection it`

Evaluation Scores:

- **BLEU:** 0.37535211267605634
- **ROUGE-1:** 0.4375376168085542
- **ROUGE-2:** 0.1680307664272397
- **ROUGE-LCS:** 0.34197911469295644

3.2 Comparison of Decoding Techniques Based on Evaluation Scores

Metric	Greedy Decoding	Word-Constrained Decoding	Improvement
BLEU	0.3099	0.3754	**+21.2%**
ROUGE-1	0.3543	0.4375	**+23.5%**
ROUGE-2	0.1296	0.1680	**+29.6%**
ROUGE-LCS	0.2710	0.3419	**+26.2%**

Table 3: Performance comparison between Greedy Decoding and Word-Constrained Decoding

Key Observations:

- **Word-Constrained Decoding Outperforms Greedy Decoding Across All Metrics**
- BLEU score improves from **0.3099** to **0.3754** (+21.2%), indicating better alignment with the reference text.

- ROUGE-1 and ROUGE-2 scores increase, signifying better word and phrase overlap.
- ROUGE-LCS increases by **26.2%**, suggesting improved structural similarity.

Effect of Additional Word Constraints on Decoding:

- Word-Constrained Decoding utilizes a word list, ensuring key words appear in output.
- This reduces randomness and helps maintain topical consistency.

Greedy Decoding is Simpler but Less Effective:

- Greedy Decoding selects the highest probability token at each step without awareness.
- The lack of constraints can lead to divergence from the reference text.

Inference: Word-Constrained Decoding significantly improves text generation performance compared to Greedy Decoding. The structured constraints help the model generate more accurate and coherent outputs, improving its alignment with reference translations.

4 Medusa: A Speculative Decoding Framework

In this section, we will explore a speculative decoding framework called Medusa [**?**]. Figure **??** provides an illustration of Medusa's architecture. The core idea behind Medusa is straightforward: in addition to the standard Language Modeling (LM) head, you also train multiple Medusa heads (a.k.a linear layers) that operate on the final hidden states of the LLM and are responsible for predicting some token in the future.

Specifically, if $y_{1:t-1}$ is the input sequence to the LLM, then:

- The LM head predicts token y_t.
- The first Medusa head predicts token y_{t+1}.
- The second Medusa head predicts token y_{t+2}.
- This continues until the Kth decoding head predicts y_{t+K}, i.e., the $(K+1)$th token in the future.

4.1 Single Head Decoding

In this approach, we use only the LM head to perform inference. That is, at each step, we greedily pick the most probable token from the LM head's output distribution. This predicted token is then fed back as input to the LLM, and the cycle repeats.

Metric	**Score**
BLEU	0.2921
ROUGE-1	0.3963
ROUGE-2	0.1483
ROUGE-LCS	0.3177
RTF	0.0715

Table 4: Evaluation metrics for the decoding approach

4.2 MultiHead Decoding

Here, we will utilize Medusa's decoding heads along with the existing LM head to generate multiple future tokens simultaneously. Let us consider $y_{1:t-1}$ as the input to the LLM and K be the total number of Medusa heads available, out of which we want to use the first S heads. Then the decoding strategy works as follows:

4.3 Inference on BLEU Score Variation with Multi-Head Attention

From the BLEU score results across different values of width and head configurations, we observe the following key trends:

1. **Higher Width with Fewer Heads Improves BLEU Score:** The configurations **(width = 2, head = 2)**, **(width = 5, head = 2)**, and **(width = 10, head = 2)** achieve significantly higher BLEU scores compared to their respective counterparts with **head = 5**. This indicates that increasing the number of heads does not necessarily lead to better translation performance and may introduce excessive noise, reducing accuracy.

2. **More Heads Reduce BLEU Scores:** When the number of heads is increased from 2 to 5 for the same width, the BLEU score consistently drops. For example:

   - **Width = 5, Head = 2:** BLEU = **0.1368**
   - **Width = 5, Head = 5:** BLEU = **0.0554** (a sharp decline)

   This suggests that increasing the number of heads may lead to unnecessary redundancy in attention computation, which affects performance.

3. **ROUGE Scores Follow a Similar Pattern:** The ROUGE-1, ROUGE-2, and ROUGE-LCS scores also follow the same trend, reinforcing that a **ower number of heads** with a **wider model width** leads to better performance.

4. **RTF (Real-Time Factor) Increases with Width and Heads:** The RTF values increase as both **width and heads** increase. This suggests a higher computational cost with a greater number of attention heads, leading to slower inference times.

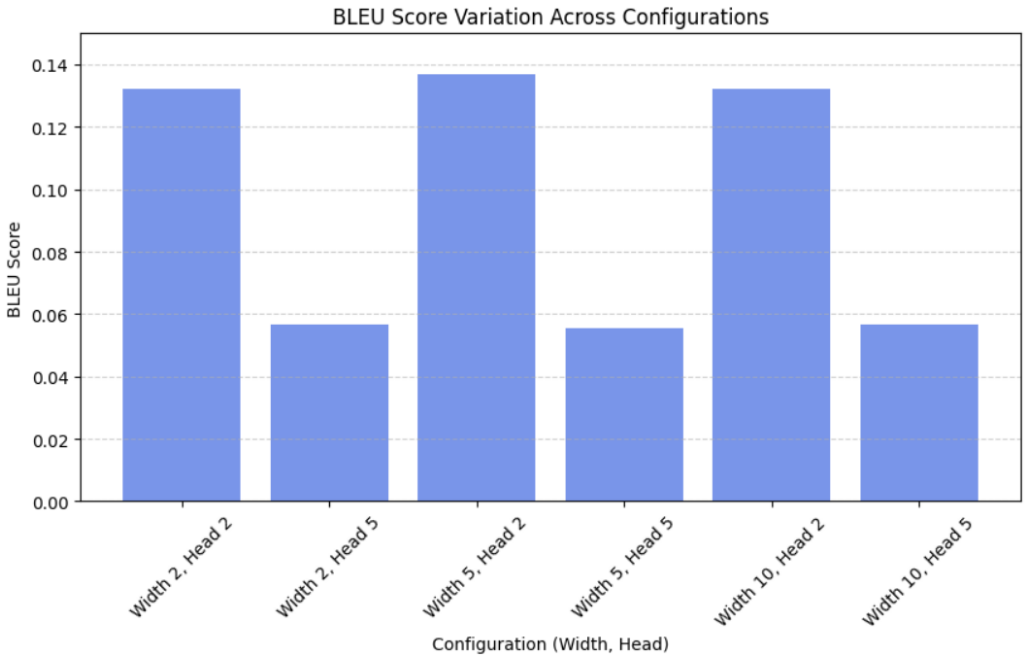

Figure 14: Plot Variation

References

1. Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. *Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads*, 2024.

2. Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, et al. *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for All 22 Scheduled Indian Languages*. Transactions on Machine Learning Research, 2023.

3. Hugo Touvron, Louis Martin, et al. *LLaMA 2: Open Foundation and Fine-Tuned Chat Models*, 2023.

4. ChatGPT and OpenAI resources

Contributions

Contributor	Task	Description
Pinak	Task 0	Implemented and analyzed different decoding strategies for LLMs.
Danish	Task 1	Developed a word-constrained decoding technique using a token-based trie.
Aansh	Task 2	Explored and implemented Medusa's multi-head speculative decoding.