Question 3:

In this report, we evaluate three machine translation models: NLLB (Distilled model), IndicTrans, and ChatGPT. We assess their performance on various translation tasks including English to Hindi, Hindi to English, Hindi to Marathi, and Marathi to Hindi. Evaluation metrics include BLEU scores and ROUGE scores (different variants).

NLLB

1. English to Hindi

   Bleu Score : 0.6250193265147185
   rouge-1 : {'r': 0.613366995297033, 'p': 0.56946024378588, 'f': 0.5854537774344653}
   rouge-2 : {'r': 0.36684542276580157, 'p': 0.338663627863593, 'f': 0.3487086921766035}
   rouge-l : {'r': 0.5668975357075521, 'p': 0.5267644681034256, 'f': 0.5413654922464272}

2. Hindi to English

   Bleu Score : 0.41071331466827654
   rouge-1 : {'r': 0.3575507191043718, 'p': 0.3396192165534538, 'f': 0.3426810871348884}
   rouge-2 : {'r': 0.17104409841665175, 'p': 0.16424276886555525, 'f': 0.16602515305388552}
   rouge-l : {'r': 0.33763822684270406, 'p': 0.32079676950920816, 'f': 0.3236639801712981}

3. Hindi to Marathi

   Bleu Score : 0.5288606654478695
   rouge-1 : {'r': 0.4408873535305414, 'p': 0.4000567126617757, 'f': 0.41431805110763037}
   rouge-2 : {'r': 0.19317377764134774, 'p': 0.17509588297254036, 'f': 0.18122718934246504}
   rouge-l : {'r': 0.4108021617142739, 'p': 0.3736810561004369, 'f': 0.38661957131894154}

4. Marathi to Hindi

   Bleu Score : 0.5661748186832466
   rouge-1 : {'r': 0.5526599664651987, 'p': 0.5018477953310488, 'f': 0.5203456055339494}
   rouge-2 : {'r': 0.2963443998371247, 'p': 0.26915091303651834, 'f': 0.27879480476596846}
   rouge-l : {'r': 0.513027193867327, 'p': 0.46592536830715736, 'f': 0.48319167068788826}

IndicTrans

1. English to Hindi

   Bleu Score : 0.7066665712972036
   rouge-1 : {'r': 0.6334738647139899, 'p': 0.6309396061149714, 'f': 0.6288335453247389}
   rouge-2 : {'r': 0.40193678215286766, 'p': 0.4001237680798042, 'f': 0.39880689659490826}
   rouge-l : {'r': 0.5963587954788969, 'p': 0.5934031989040611, 'f': 0.5917419199684094}

2. Hindi to English

Bleu Score : 0.4479995439238356
rouge-1 : {'r': 0.37192230117286434, 'p': 0.3746205928655386, 'f': 0.3674154342312152}
rouge-2 : {'r': 0.18995708385720272, 'p': 0.19326541320016094, 'f': 0.19041753461205355}
rouge-l : {'r': 0.352345940654747, 'p': 0.3550702411316087, 'f': 0.34813489840498846}

3. Hindi to Marathi

Bleu Score : 0.604148489191587
rouge-1 : {'r': 0.4539714757017372, 'p': 0.4538173721847189, 'f': 0.45031956962538044}
rouge-2 : {'r': 0.2088336782352174, 'p': 0.20935763703562818, 'f': 0.20739598353402164}
rouge-l : {'r': 0.4281218711429381, 'p': 0.42828062411443457, 'f': 0.42482330944493624}

4. Marathi to Hindi

Bleu Score : 0.6182068355709924
rouge-1 : {'r': 0.5398530377948986, 'p': 0.5308555708190866, 'f': 0.5307159539999723}
rouge-2 : {'r': 0.2943369385786013, 'p': 0.2893510128190569, 'f': 0.2891481651027939}
rouge-l : {'r': 0.5018032389295699, 'p': 0.4927852802546739, 'f': 0.4930793403840655}

ChatGPT

1. English to Hindi

Bleu Score : 0.6166150885037212
rouge-1 : {'r': 0.5902300304856638, 'p': 0.5696823367283433, 'f': 0.5780784982462601}
rouge-2 : {'r': 0.3359028474679372, 'p': 0.32437603093811, 'f': 0.32907551627614046}
rouge-l : {'r': 0.5422047065761293, 'p': 0.5217512585472651, 'f': 0.5302962952480269}

2. Hindi to English

Bleu Score : 0.6924387569942543
rouge-1 : {'r': 0.6101105603115969, 'p': 0.6120678425977186, 'f': 0.6078969426790547}
rouge-2 : {'r': 0.37684577550017484, 'p': 0.3757120607497727, 'f': 0.37371356450207055}
rouge-l : {'r': 0.5889275074593623, 'p': 0.5907629632665234, 'f': 0.5867006520139373}

3. Hindi to Marathi

Bleu Score : 0.45228646094340247
rouge-1 : {'r': 0.27939858916528054, 'p': 0.2825290975457889, 'f': 0.2788451747579006}
rouge-2 : {'r': 0.08202905135799872, 'p': 0.08596615802498157, 'f': 0.083175773575077}
rouge-l : {'r': 0.2701323553990468, 'p': 0.2732892365296121, 'f': 0.26968610059219605}

4. Marathi to Hindi

Bleu Score : 0.4945372323286291
rouge-1 : {'r': 0.4054991131368117, 'p': 0.3856036760255194, 'f': 0.38927085115862603}
rouge-2 : {'r': 0.16763423122322701, 'p': 0.15829215710966804, 'f': 0.16067218411262232}
rouge-l : {'r': 0.3394996574105549, 'p': 0.3220572322651, 'f': 0.32563843892387717}

Conclusion

From this evaluation, several insights can be drawn:

- ChatGPT demonstrates competitive performance across different translation tasks, especially in English to Hindi and Hindi to English translations.
- IndicBART performs consistently well across all evaluated tasks, showcasing its effectiveness in translation between Indian languages.
- NLLB shows decent performance but lags slightly behind ChatGPT and IndicBART in most scenarios.
- BLEU and ROUGE scores provide valuable insights into the quality of translations, but they should be considered alongside other factors such as model efficiency and computational resources required.

Overall, the choice of machine translation model should be tailored to specific language pairs and desired translation quality, considering factors such as BLEU and ROUGE scores along with practical considerations like model size and inference speed.

Learning

- **Performance Variability:** Different translation models perform differently across tasks and language pairs.

- **IndicTrans Effectiveness:** IndicTrans consistently performs well, particularly for Indian language translations.

- **NLLB Performance:** NLLB shows decent performance but is slightly behind other models in most scenarios.

- **Metric Importance:** BLEU and ROUGE scores offer valuable metrics for assessing translation quality.

- **Consideration Factors:** Model efficiency, computational resources, and task requirements should influence model choice.

- **Tailored Selection:** Choose the translation model based on specific language pairs and desired quality.