1.

(g) Masks sets e_t of <pad> tokens to -inf. Hence the probability of these tokens will be zero in the softmax output and as a result, the attention weight of these tokens will be zero. By doing this, we are telling the model that <pad> tokens are never of importance and they should not affect the prediction in the decoder part. It is necessary to do so because as mentioned we do not want <pad> tokens to affect the decoder predictions.

(h) 12.13 > 10

```
[nltk_data]   Package punkt is already up-to-date!
load test source sentences from [./chr_en_data/test.chr]
load test target sentences from [./chr_en_data/test.en]
load model from model.bin
Decoding: 100%|██████████| 1000/1000 [01:06<00:00, 15.00it/s]
Corpus BLEU: 12.132865668925385
```

(i) i. advantage: It is computationally faster and more memory efficient. disadvantage: It does not learn how to balance the impact of encoder and decoder hidden states. It just simply calculates the dot product of them without any special modification but maybe in some cases the hidden state of the decoder should be considered more important or vice versa.
ii. advantage: It has more (and separated) learnable parameters and a non-linearity which makes it potentially more intelligent and flexible. disadvantage: It is computationally more costly.

2.

(a) From wikipedia: In linguistic typology, polysynthetic languages are highly synthetic languages, i.e. languages in which words are composed of many morphemes (word parts that have independent meaning but may or may not be able to stand alone). They are very highly inflected languages. Polysynthetic languages typically have long "sentence-words" such as the Yupik word tuntussuqatarniksaitengqiggtuq which means "He had not yet said again that he was going to hunt reindeer." The word consists of the morphemes tuntu-ssur-qatar-ni-ksaite-ngqiggte-uq with the meanings, reindeer-hunt-future-say-negation-again-third person-singular-indicative; and except for the morpheme tuntu "reindeer", none of the other morphemes can appear in isolation.

Having the above explanation, Cherokee as a polysynthetic language has words that are combinations of other words not separated by space. Hence the vocabulary size of this language will be huge. So it is computationally better to have separate embeddings for each sub-word. Additionally, in these languages the chance of having better representation for a compound(?) word is higher if we concat embedding of meaningful sub-words because each of these sub-words are more repeated than each combination of them.

(b) Because often there are some limited sub-words that are repeated over and over in the construction of every word. In the extreme case, we have 26 characters in English but over a million words.

(c) By using this technique we share learnable parameters across different languages. As there are many inherent similarities between languages, using this technique can improve low-resource languages by helping models to capture some fundamental and similar characteristics from high-resource languages and use them for low-resource ones.

(d) i. Cause: Maybe Cherokee word for hair and crown is similar. Fix: Use different embeddings for a subword in different contexts.

ii. Cause: Maybe pronouns are genderless or implicit or have a very subtle difference in Cherokee. Fix: Add a lot of instances of some sentences with just different pronouns. (For example I am happy and He is happy and She is happy)

iii. Cause: By the capitalization of Littlefish we find out that it is a special name. Maybe in Cherokee there is no explicit difference between special names and their literal equivalents. Or maybe it is the fault of the model that took the name translation literally. Fix: Add copying mechanisms to the model or add some special names to target vocab.

(e) I used **wdiff**.

i. At line 273: “the master of the house”. At line 515: “the face of the Lord”. At line 69: “know that the Son of man”. At line 370: “that the scripture might be”

These examples show that our NMT model learned the use of “the” and the structure of “A of B” pretty well.

ii. At line 865: (test) vs. (machine)
“But what saith the answer of God unto him? I have left for myself seven thousand men, who have not bowed the knee to Baal.” vs. “ But what saith the God of God? I pray thee, because of the seven thousand men.”

This example shows that sometimes the model wrongly increases the weights for previous decoder hidden states such that the encoder attentions has little effect on the output.

(f)

i. Source Sentence s: el amor todo lo puede
Reference Translation r 1 : love can always find a way
Reference Translation r 2 : love makes anything possible
NMT Translation c 1 : the love can always do
NMT Translation c 2 : love can make anything possible

c1:
p1 = 0 + 1 + 1 + 1 + 0 / 5 = 0.6
p2 = 0 + 1 + 1 + 0 / 4 = 0.5
len(c) = 5, len(r) = 4
BP = 1
BLEU = 1 * exp(0.5 * log(0.6) + 0.5 * log(0.5)) = 0.5477

c2:
p1 = 1 + 1 + 0 + 1 + 1 / 5 = 0.8
p2 = 1 + 0 + 0 + 1 / 4 = 0.5
len(c) = 5, len(r) = 5
BP = 1
BLEU = 1 * exp(0.5 * log(0.8) + 0.5 * log(0.5)) = 0.6325

c2 has a better BLEU score. Yes it is actually a better translation.

ii. c1:

p1 = 0 + 1 + 1 + 1 + 0 / 5 = 0.6
p2 = 0 + 1 + 1 + 0 / 4 = 0.5
len(c) = 5, len(r) = 4
BP = 1
BLEU = 1 * exp(0.5 * log(0.6) + 0.5 * log(0.5)) = 0.5477

c2:
p1 = 1 + 1 + 0 + 0 + 0 / 5 = 0.4
p2 = 1 + 0 + 0 + 0 / 4 = 0.25
len(c) = 5, len(r) = 4
BP = 1
BLEU = 1 * exp(0.5 * log(0.4) + 0.5 * log(0.25)) = 0.3162

c1 has a better BLEU score. No it is not a good translation.

iii. There is a risk of having a wrong translation to be the only reference. Sometimes the only translation is actually a correct one but it is not the only form possible and maybe our model predicts a correct translation which is not similar to the reference.

iv. adv: It is deterministic and the cost of computing it is much less than human evaluation (computationally and monetary).
disadv: Our model may sometimes predict a correct translation but due to difference in the choice of words and tenses with the reference sentence it gets a low score. But in such cases humans are more robust. Another disadvantage is that BLEU does not take into account the semantics and the grammar of the sentence whereas a human understands if a translation makes sense or not