

Burmese Grammar Error Correction using Deep Learning Approach

Nay Win Hlaing, Khin Su Myat Moe

Content

Introduction

Model Selection

Results and Evaluation

Conclusions

Introduction

The Burmese language is an underrepresented language in the field of NLP. Its script presents unique challenges due to ambiguities in complex grammar rules, and the lack of standardized word segmentation. These challenges are especially evident in digital communication, where users often make usage and typing errors. Unlike widely studied languages like English, Burmese lacks reliable tools for automatic sentence correction.

This project aims to develop a deep learning-based model that can automatically detect and transform incorrect Burmese sentences into correct ones. The correction will focus on grammar structure, and word segmentation, with the scope limited to simple sentence patterns

ကြိယာတစ်ခုရှိသော ဝါကျဖွဲ့နည်းများ [ပြင်ဆင်ရန်]

၁။ ကတ္တားပုဒ်၊ ကြိယာပုဒ် တို့ဖြင့်ဖွဲ့သော ဝါကျ။

*ပုံစံ - တောင်(ကတ္တား) ပြီသည်။ (ကြိယာပုဒ်)

၂။ ကတ္တားပုဒ်၊ အဖြည့်ကတ္တား၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ဗာဝရီ ရသေ့ကား (ကတ္တားပုဒ်) အနာဂတ်(အဖြည့်ကတ္တား) ဖြစ်၏။ (ကြိယာပုဒ်)

၃။ ကတ္တားပုဒ်၊ ကံပုဒ် ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ယောကျားသည် (ကတ္တားပုဒ်) မိန်းမကို (ကံပုဒ်) ကြည့်၏။ (ကြိယာပုဒ်)

၄။ ကတ္တားပုဒ်၊ ကံပုဒ်၊ အဖြည့်ကံ၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ပြုမိဖုရားလည်း(ကတ္တားပုဒ်) မဟာသမ္မဝအား (ကံပုဒ်) မင်း(အဖြည့်ကံ) မြှောက်လေ၏။ (ကြိယာပုဒ်)

၅။ ကတ္တားပုဒ်၊ ကံပုဒ်၊ လက်ခံပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ဖေဒါရိကလည်း (ကတ္တားပုဒ်) အကြောင်းမျိုးကို (ကံပုဒ်) ခမည်းတော် ရှင်ရသေ့အား (လက်ခံပုဒ်) လျှောက်လေ၏။ (ကြိယာပုဒ်)

၆။ ကတ္တားပုဒ်၊ ထွက်ခွာရာ ပြသည့်ပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့်ဖွဲ့သော ဝါကျ။

*ပုံစံ - ပုဏ္ဏားတို့ (ကတ္တားပုဒ်) ကျွန်းအဖြစ်မှ (ထွက်ခွာရာပြသည့်ပုဒ်) လွတ်လေသတည်း။ (ကြိယာပုဒ်)

၇။ ကတ္တားပုဒ်၊ ရေးရှူရာပြသည့်ပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - တောင်ငူမင်းလည်း (ကတ္တားပုဒ်) တောင်ငူမြို့သို့(ရေးရှူရာ ပြသည့်ပုဒ်) ပြန်လေ၏။ (ကြိယာပုဒ်)

၈။ ကတ္တားပုဒ်၊ နေရာပြပုဒ်၊ ကြိယာပြပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - မိုးကြိုး(ကတ္တားပုဒ်) နှန်းမှာ (နေရာပြပုဒ်) ကျသည်။ (ကြိယာပုဒ်)

၉။ ကတ္တားပုဒ်၊ အချိန်ပြပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ဤအစာတို့သည် (ကတ္တားပုဒ်) အကုသိုလ်ကြောင့် (အကြောင်းပြပုဒ်) ကွယ်ခဲ့ကုန်၏။ (ကြိယာပုဒ်)

၁၀။ ကတ္တားပုဒ်၊ အကြောင်းပြပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - ထိအစာတို့သည် (ကတ္တားပုဒ်) အကုသိုလ်ကြောင့် (အကြောင်းပြပုဒ်) ကွယ်ခဲ့ကုန်၏။ (ကြိယာပုဒ်)

၁၁။ ကတ္တားပုဒ်၊ ကံပုဒ်၊ အသုံးခံပုဒ်၊ ကြိယာပုဒ်တို့ဖြင့် ဖွဲ့သော ဝါကျ။

*ပုံစံ - မြှို့မြို့ဖုရားသည် (ကတ္တားပုဒ်) တိုက်တံခါးရွက်ကို (ကံပုဒ်) လက်ဖြင့် (အသုံးခံပုဒ်) ခေါက်၏။ (ကြိယာပုဒ်)

- a) ကျောင်းမှ စောင့်နေပါ။ -> ကျောင်းက စောင့်နေပါ။
- b) သူတို့သည် ဒီမိုကရေစိရေးအတွက် တိုက်ကြသည်။ -> သူတို့သည် ဒီမိုကရေစိရှိရေးအတွက် တိုက်ကြသည်။
- c) သူအိမ်ကို ထွက်သွားတယ်။ -> သူအိမ်မှ ထွက်သွားတယ်။
- d) ကျောင်းသွား ပြင်ဆင်တယ်။ -> ကျောင်းသွားဖို့ ပြင်ဆင်တယ်။
- e) စာဖတ်မယ် ပြီးကျောင်းသွားမယ်။ -> စာဖတ်ပြီး ကျောင်းသွားမယ်။
- f) ငါ နံနက်ကို လာမယ်။ -> ငါ နံနက်မှာ လာမယ်။
- g) သူ အားနည်းခြင်းဖြင့် လဲကျေတယ်။ -> သူ အားနည်းခြင်းကြောင့် လဲကျေတယ်။
- h) သူလည်း မနက်စာလည်း မစားဘူး။ -> သူ မနက်စာလည်း မစားဘူး။
- i) သူ မ သွားမယ်။ -> သူ မသွားဘူး။
- j) မနေ့က သူ လာတယ်။ -> မနေ့က သူ လာခဲ့တယ်။

ဘောင်းဘီဖြင့် လူကို မှတ်မိ၏။
ဘောင်းဘီနှင့် လူကို မှတ်မိ၏။
အကြောင်းပြု။ တက္ကဖြစ်သော၊ တက္က ပါသောအရာတို့ တွင်(နှင့်)ကိုသာ သုံးရသည်။



သူသည် ပြည်သို့ မီးရထားနှင့် သွားသည်။
သူသည် ပြည်သို့ မီးရထားဖြင့် သွားသည်။
အကြောင်းပြု။ အားကိုး အမိုအဖြစ်၊ လက်ခွဲအသုံးချုပ္ပါည်းအဖြစ်၊ အထောက်အခဲ အကူအညီအဖြစ် သုံးစွဲသောပုဒ်များအတွက် (ဖြင့်) (အားဖြင့်)ကိုသာ သုံးရသည်။

အမှား။ ဦးဘ မှ ဖိတ်ကြားပါသည်။
အမှန်။ ဦးဘ က ဖိတ်ကြားပါသည်။

အမှား။ ကျောင်းသားသည် စာအုပ်အား ဆရာကိုပေး၏။
အမှန်။ ကျောင်းသားသည် စာအုပ်ကို ဆရာအားပေး၏။

အမှား။ ဘာက အန္ောင့်အယုက်ဖြစ်နေသလဲ။
အမှန်။ ဘာက အန္ောင့်အယုက်ပြုနေသလဲ။
အကြောင်းပြု။ (က)သည် ပြလုပ်ကြောင်းကြိယာနှင့်သာ တွဲချုပ်သုံးနိုင်သော ပုဒ်ဖြစ်သည်။ ဖြစ်ကြောင်းကြိယာနှင့် မသုံးနိုင်။ ထို့ကြောင့်-
ဘာ(သည်)အန္ောင့်အယုက် ဖြစ်နေသလဲ ဟူ၍သာ ရေးလိုက ရေးနိုင်သည်။

သို့ရာတွင် အားကိုးအကူအထောက်အဖြစ်ဖြင့် သုံးစွဲ စေကာမှာ၊ တစ်ချိန်တည်းတွင် (တပေါင်းတကွတည်း) ကဲသို့ သာောဝင်နေသည့်အခါ (နှင့်)ကိုလည်း သုံးကြပ်နှင့် သည်။ ထို့ကြောင့် (နှင့်)နှင့် (ဖြင့်)အသုံးများ ရောထွေးရ တတ်သည်။
လက်အိတ်ဖြင့် ကိုင်သည်၊ လက်အိတ်နှင့် ကိုင်သည်။
ရေဖြင့် နယ်သည်။ ရေနှင့် နယ်သည်။



Model Selection

1

ByT-5 Model

ByT5 is tokenizer-free version of the T5 model designed to work directly on raw UTF-8 bytes. It can process any language, more robust to noise like typos, and simpler to use.

2

Mt-5 base

Google mT5 is a multilingual variant of the T5 (Text-To-Text Transfer Transformer) model designed for natural language processing tasks. It supports over 100 languages. mT5 uses SentencePiece (Unigram) subword tokenization.

Model Selection

mbart-50 Model

mBART-50 is a multilingual version
of **BART** (a seq2seq Transformer)
trained by **Meta (Facebook)** as a
denoising autoencoder across 50
languages.

SentencePiece shared vocab;
requires language codes.

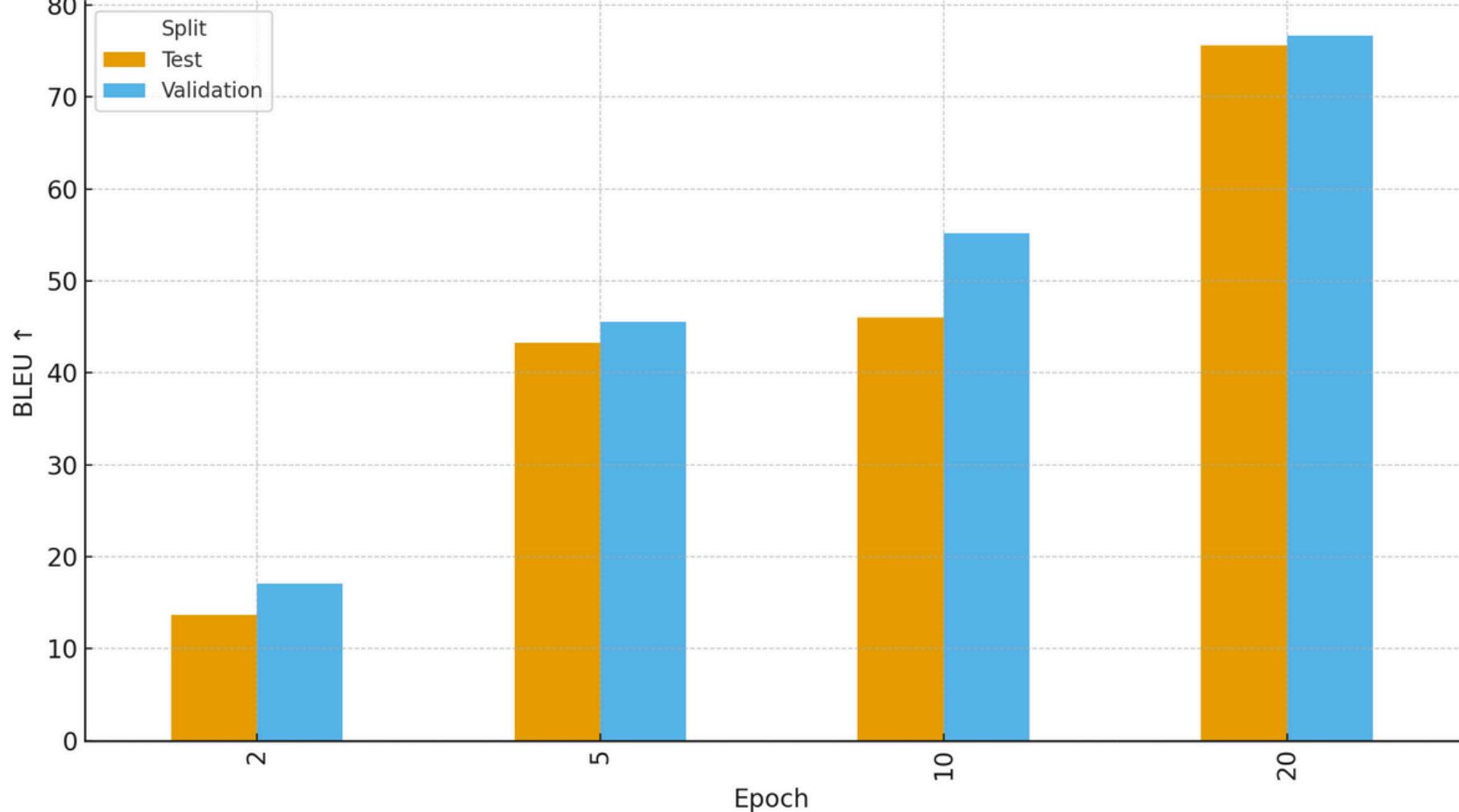
Evaluation Metrics

1

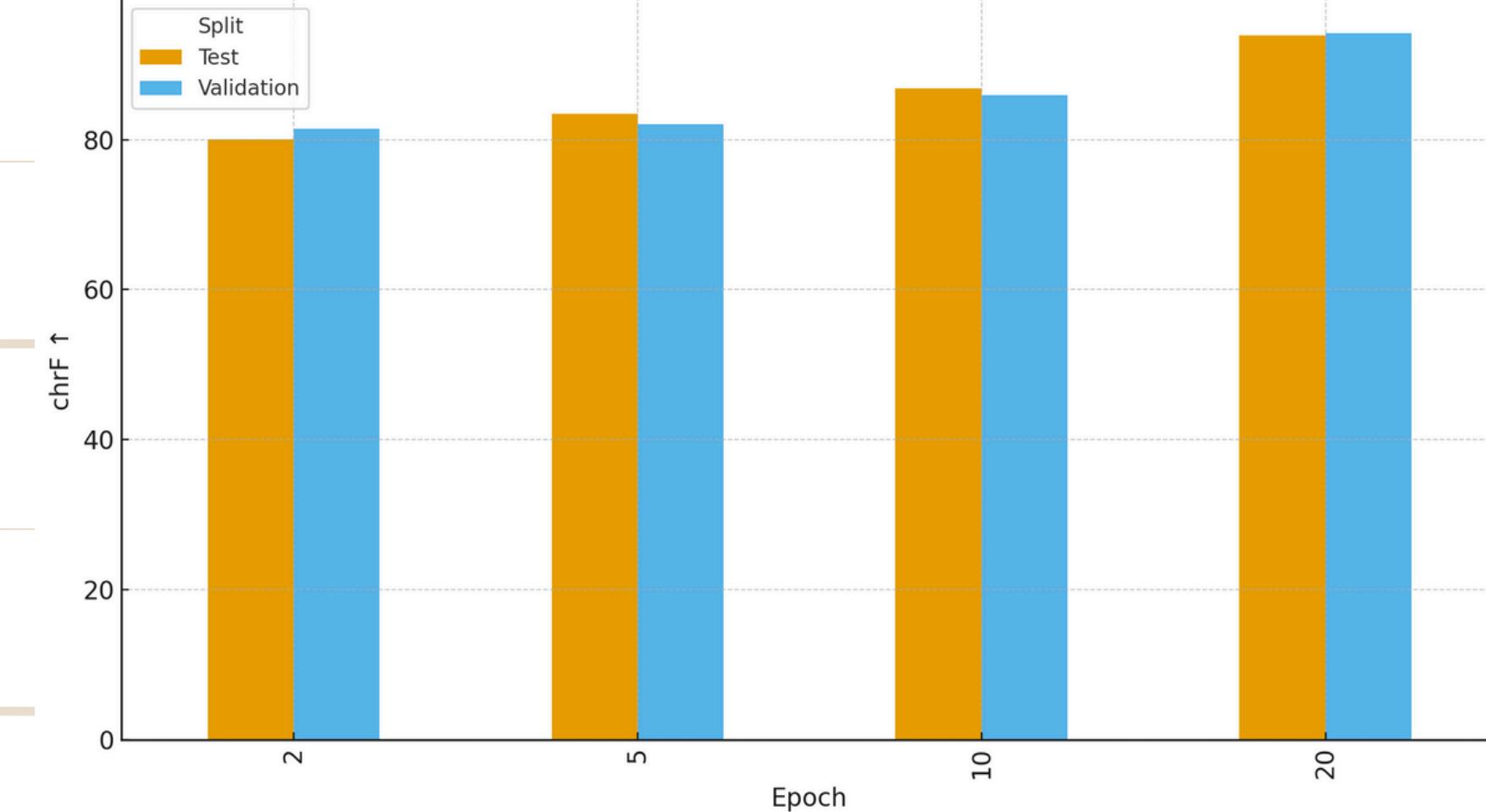
- **BLEU Score-** BLEU (Bilingual Evaluation Understudy) is a popular metric for text generation. Compares model output against the reference (correct sentences) based on n-gram overlap. Score ranges from 0 → 100.
- **ChrF (Character F-score)** measures similarity at the character level.
- **test_exact_match-** The strictest metric: percentage of predictions exactly identical to the reference sentence.
- **test_wer-** $(\text{Substitutions} + \text{Deletions} + \text{Insertions}) \div \text{Total Words}$. Lower is better.
- **test_cer-Character Error Rate (CER)** = errors at the character level \div total characters.

2

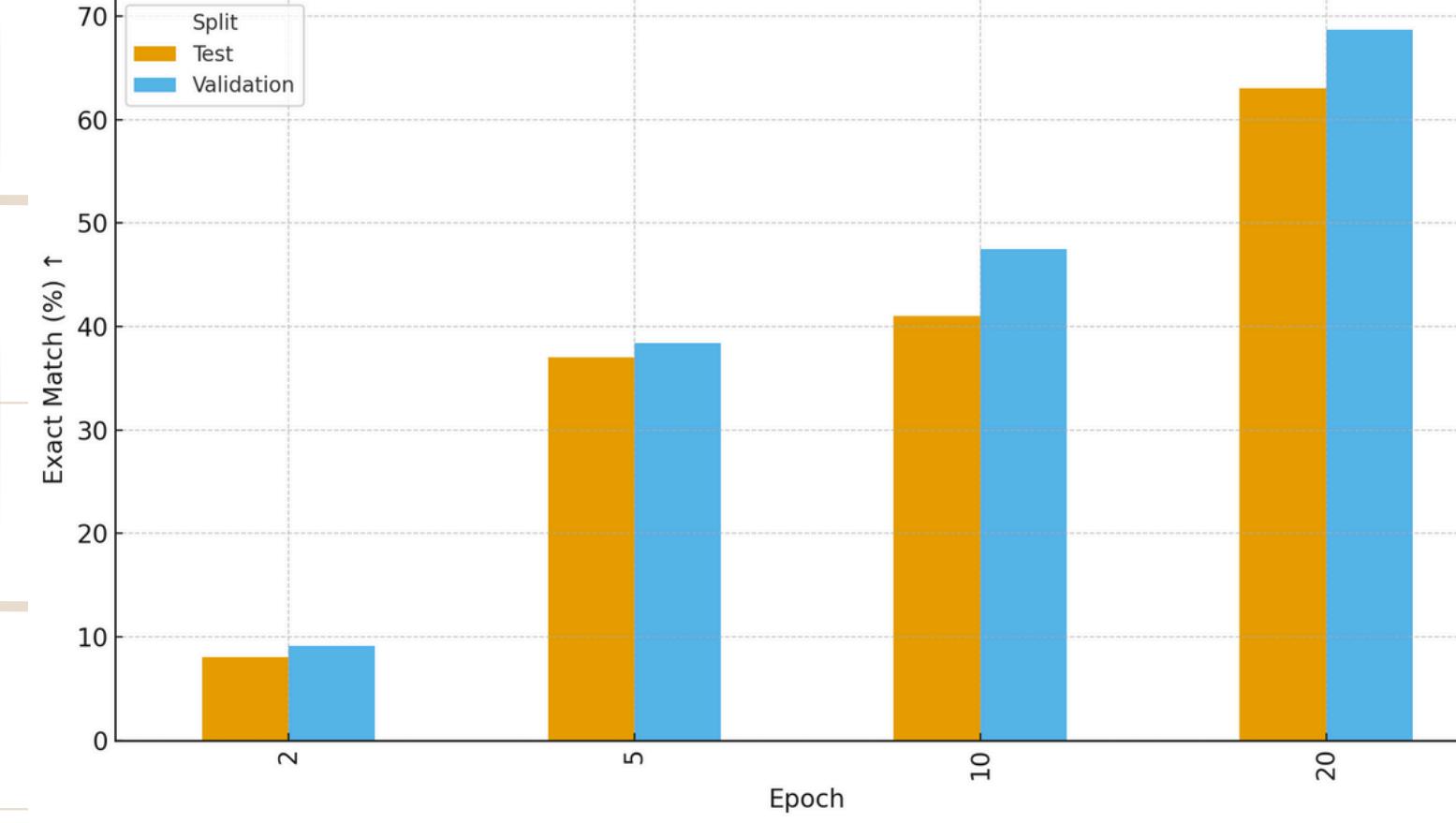
ByT5: BLEU by Epoch (Validation vs Test)



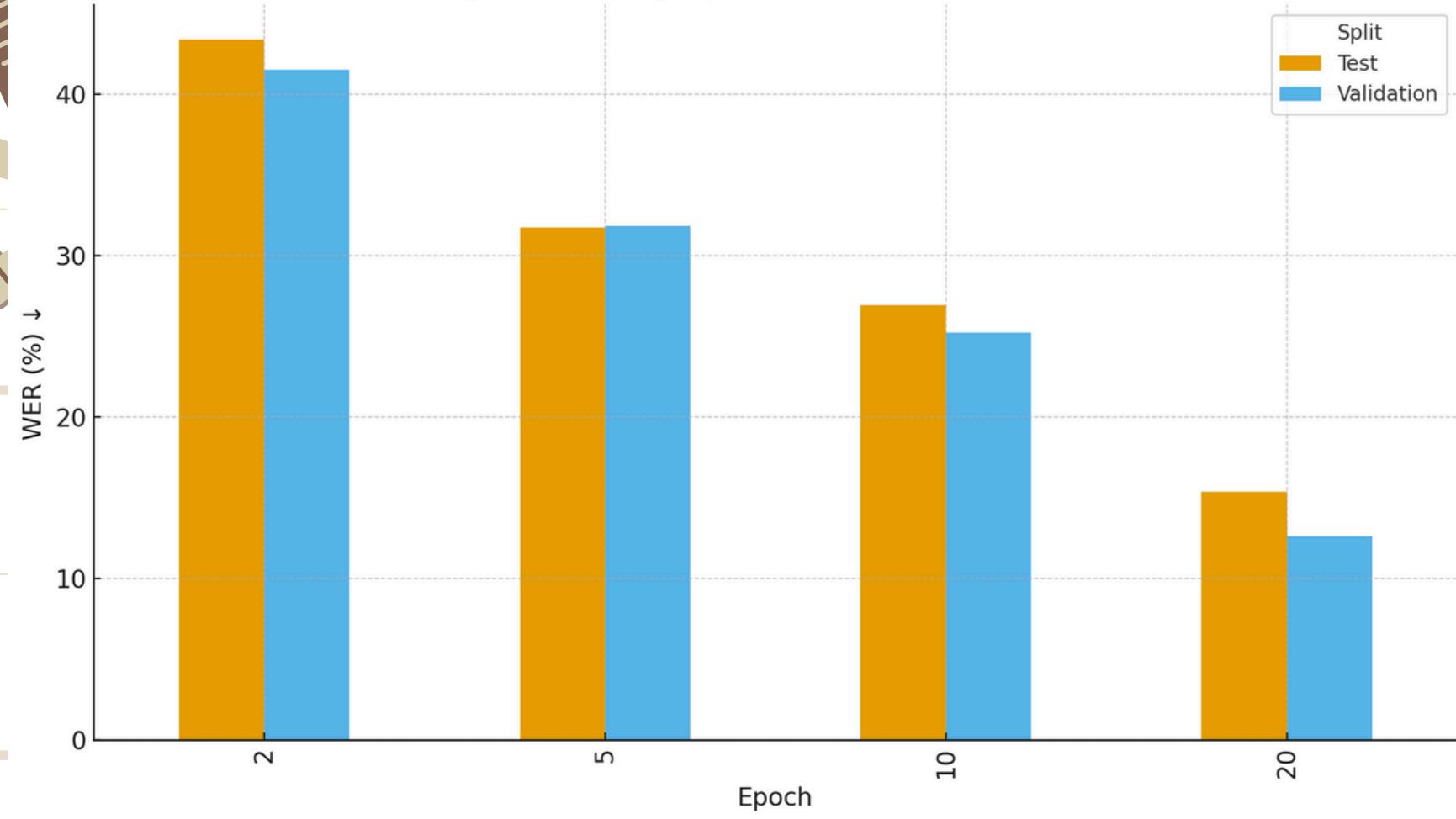
ByT5: chrF by Epoch (Validation vs Test)



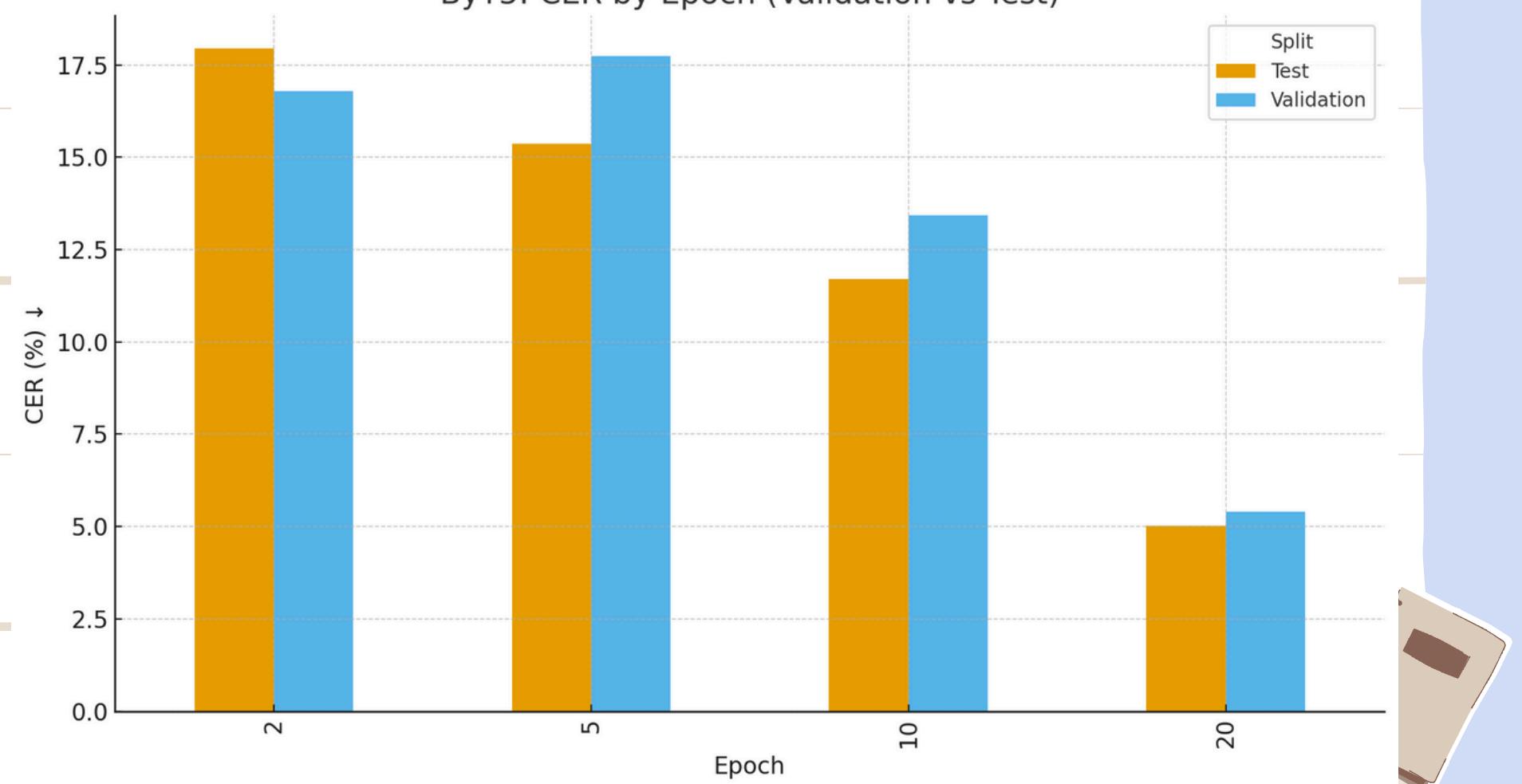
ByT5: Exact Match by Epoch (Validation vs Test)



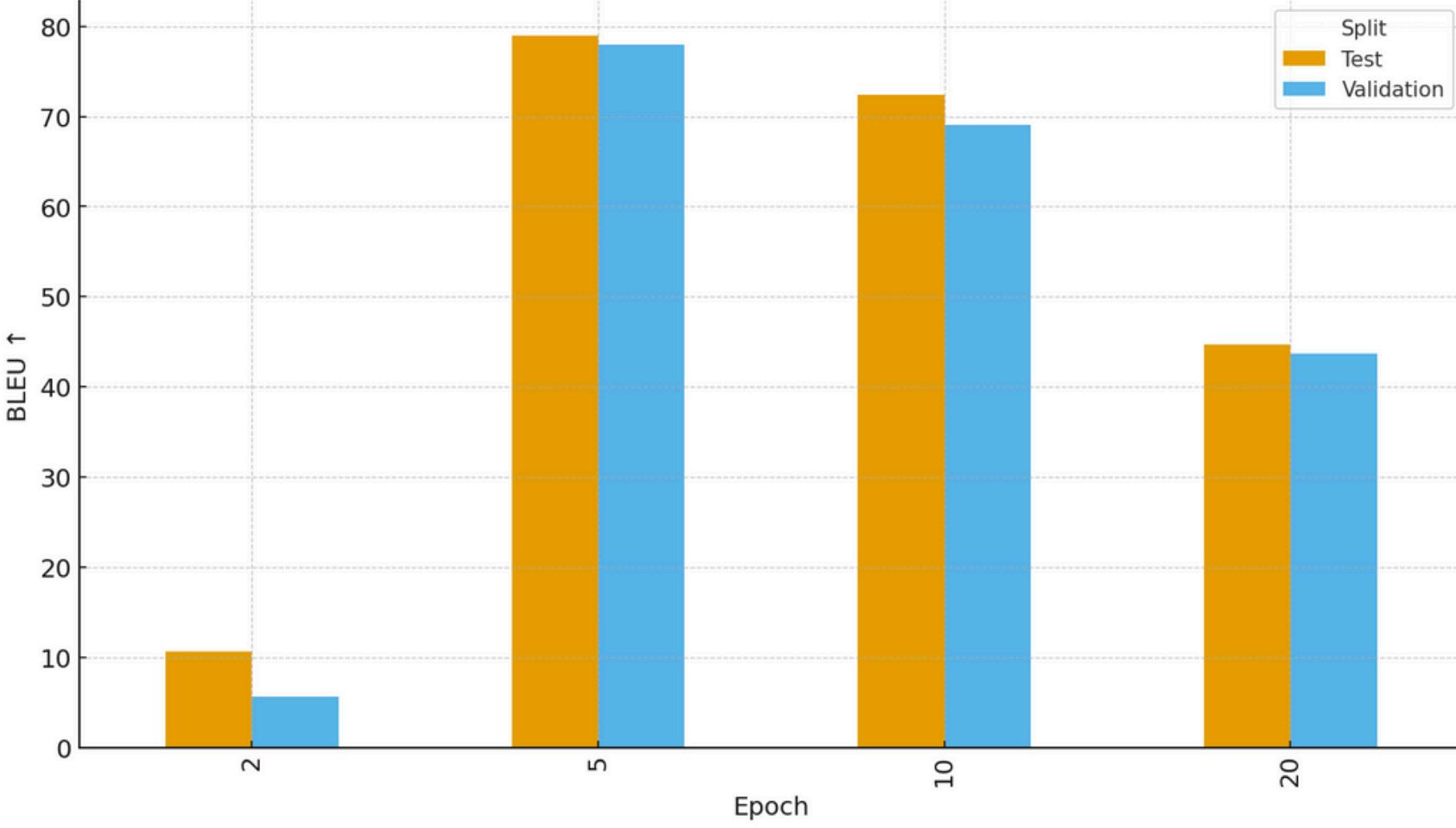
ByT5: WER by Epoch (Validation vs Test)



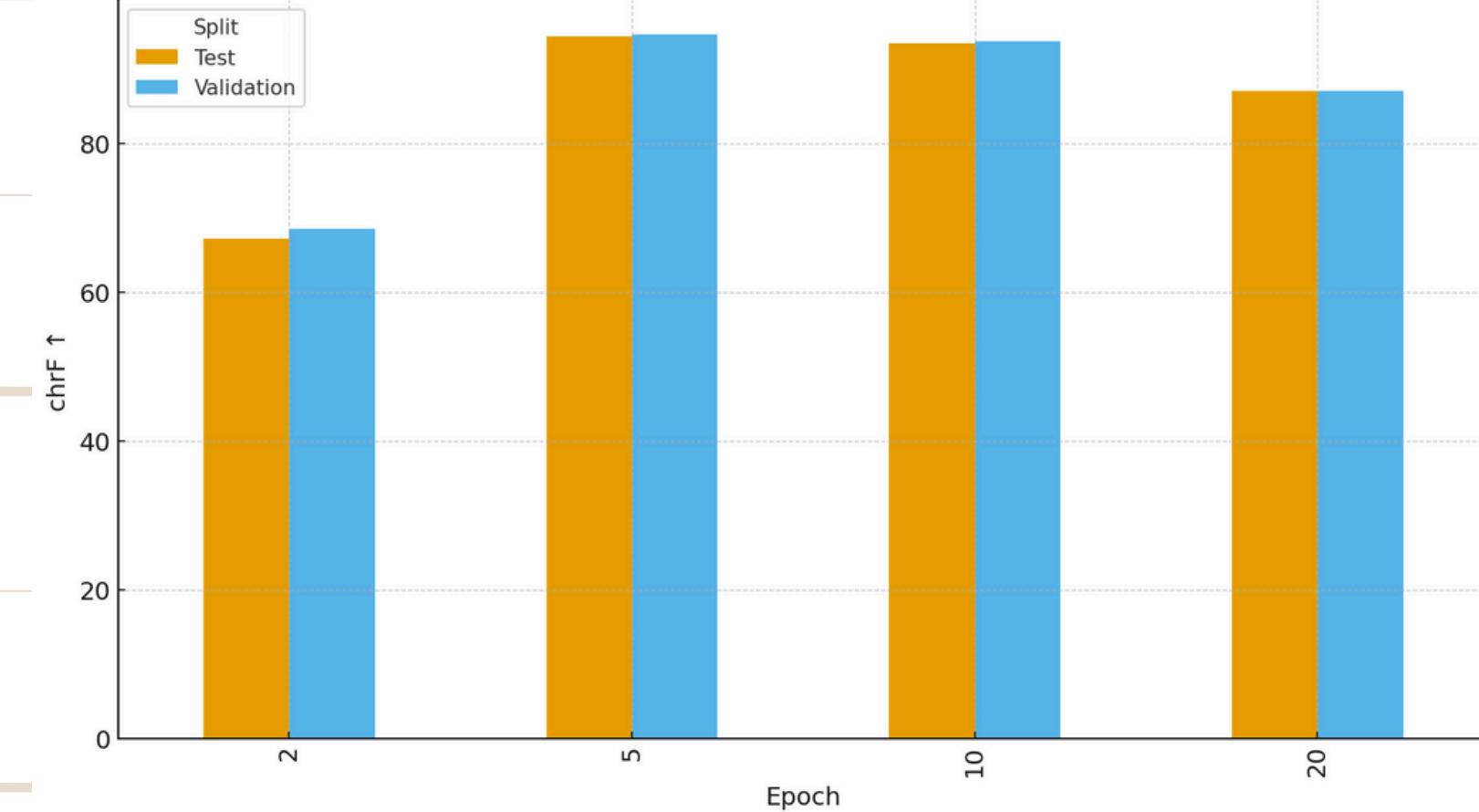
ByT5: CER by Epoch (Validation vs Test)



mT5: BLEU by Epoch (Validation vs Test)

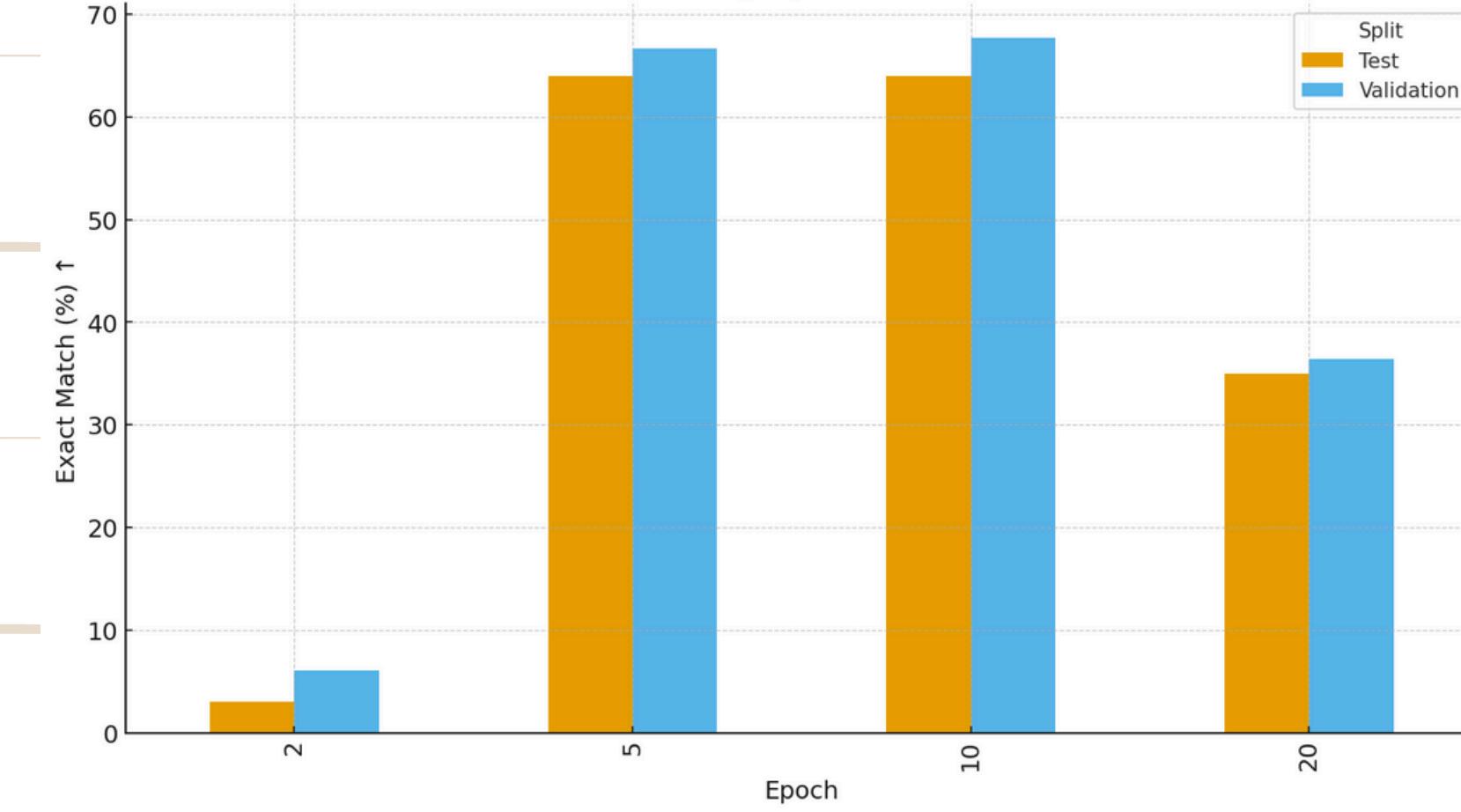


mT5: chrF by Epoch (Validation vs Test)



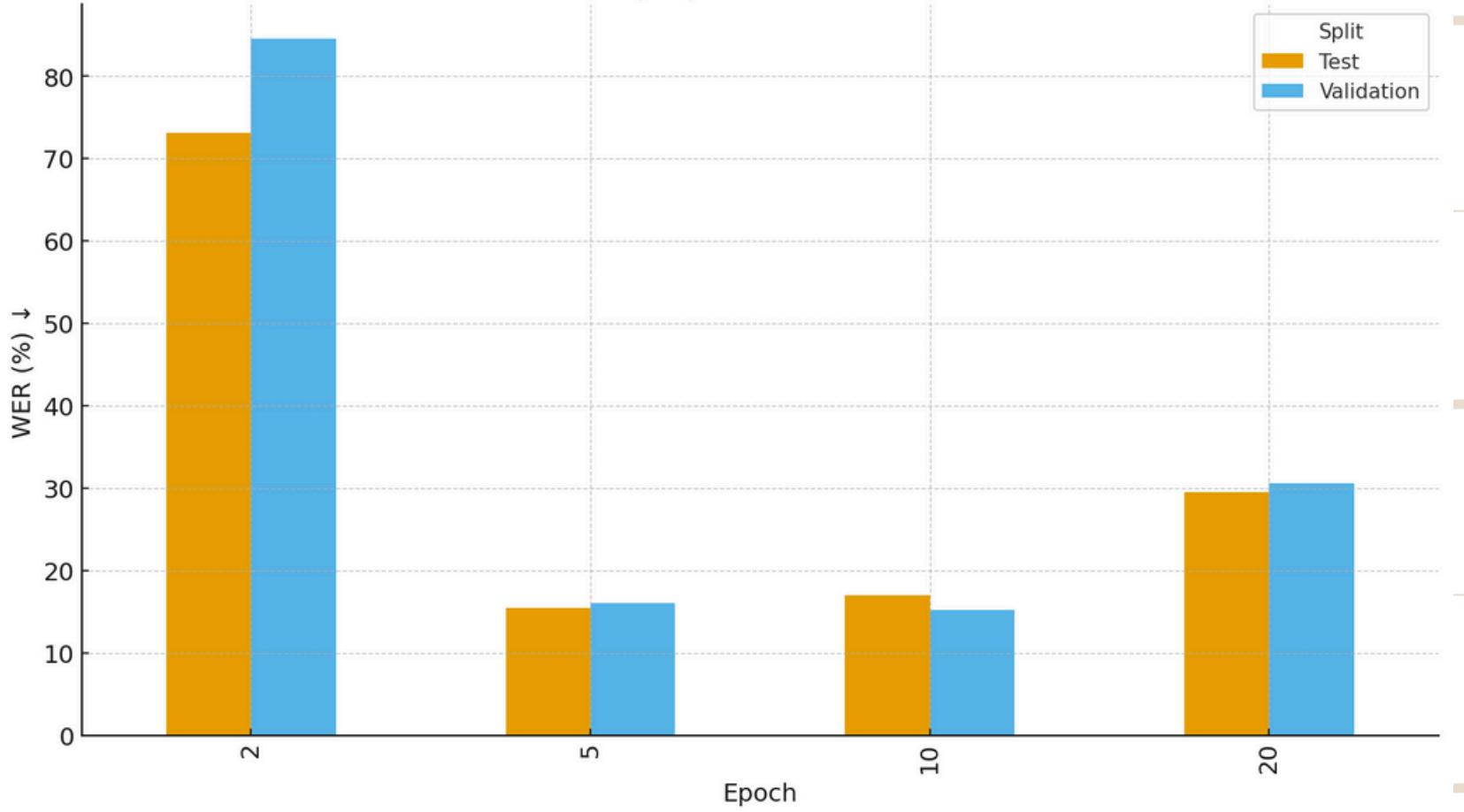
Epoch

mT5: Exact Match by Epoch (Validation vs Test)

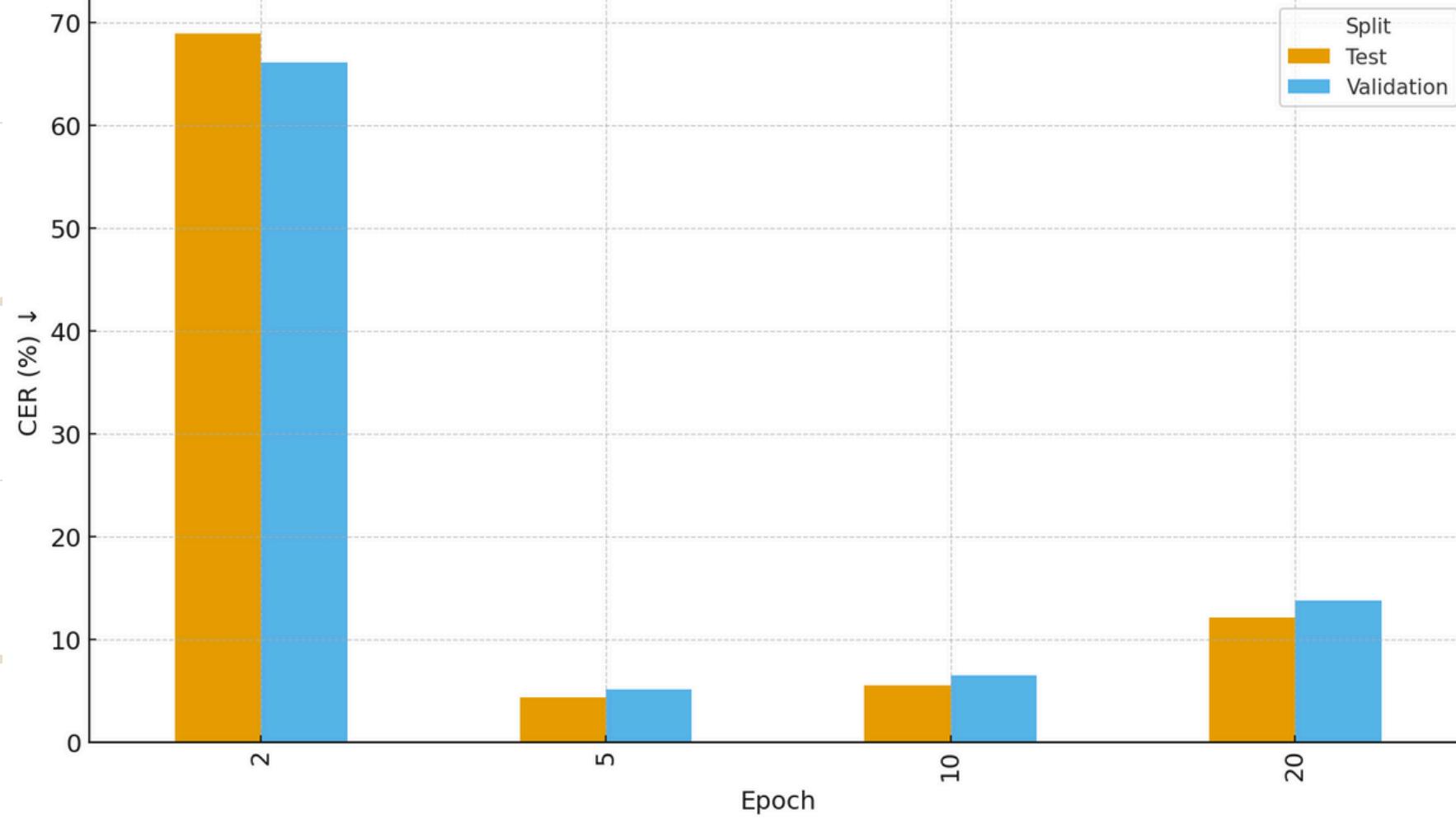


Epoch

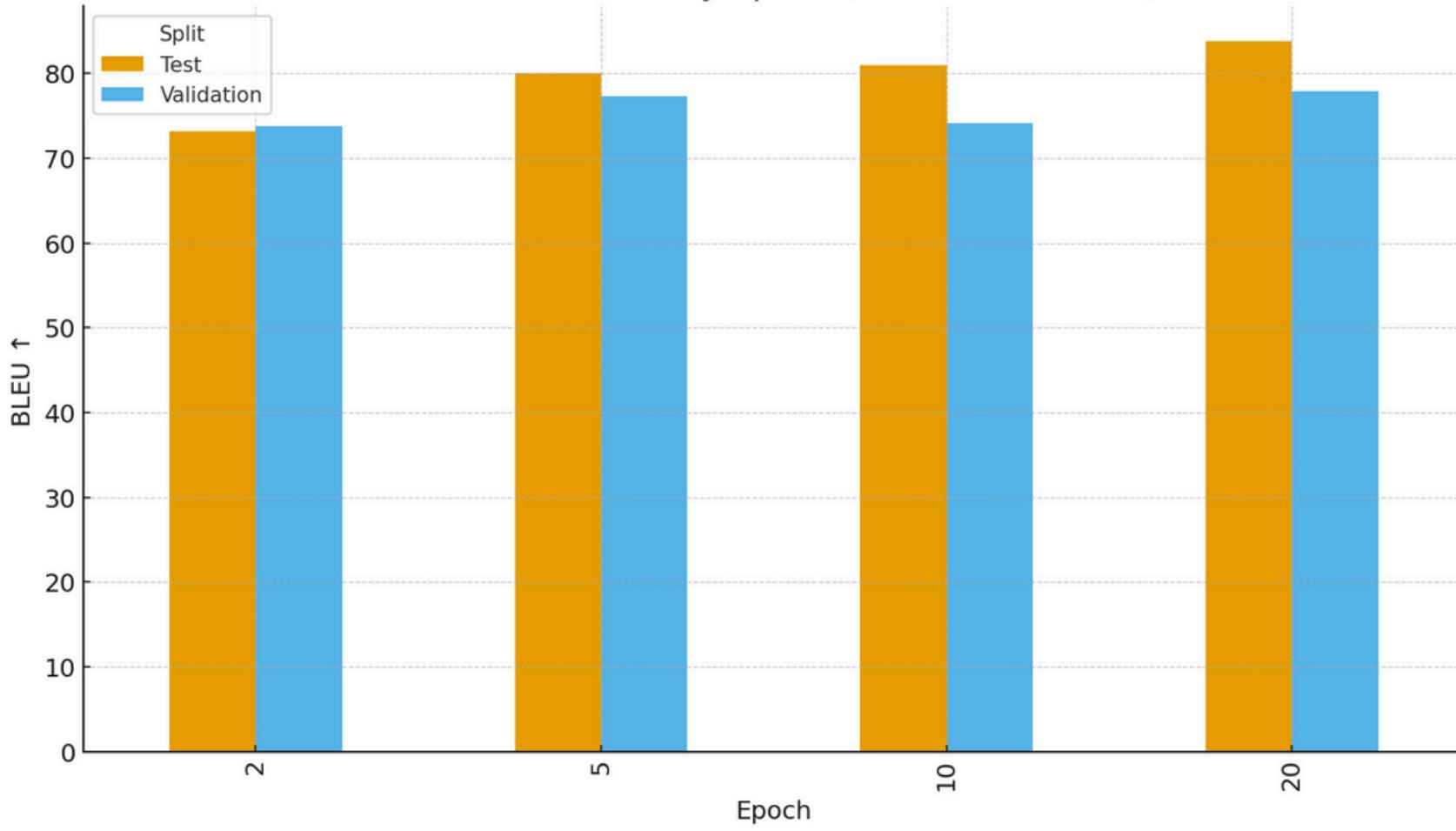
mT5: WER by Epoch (Validation vs Test)



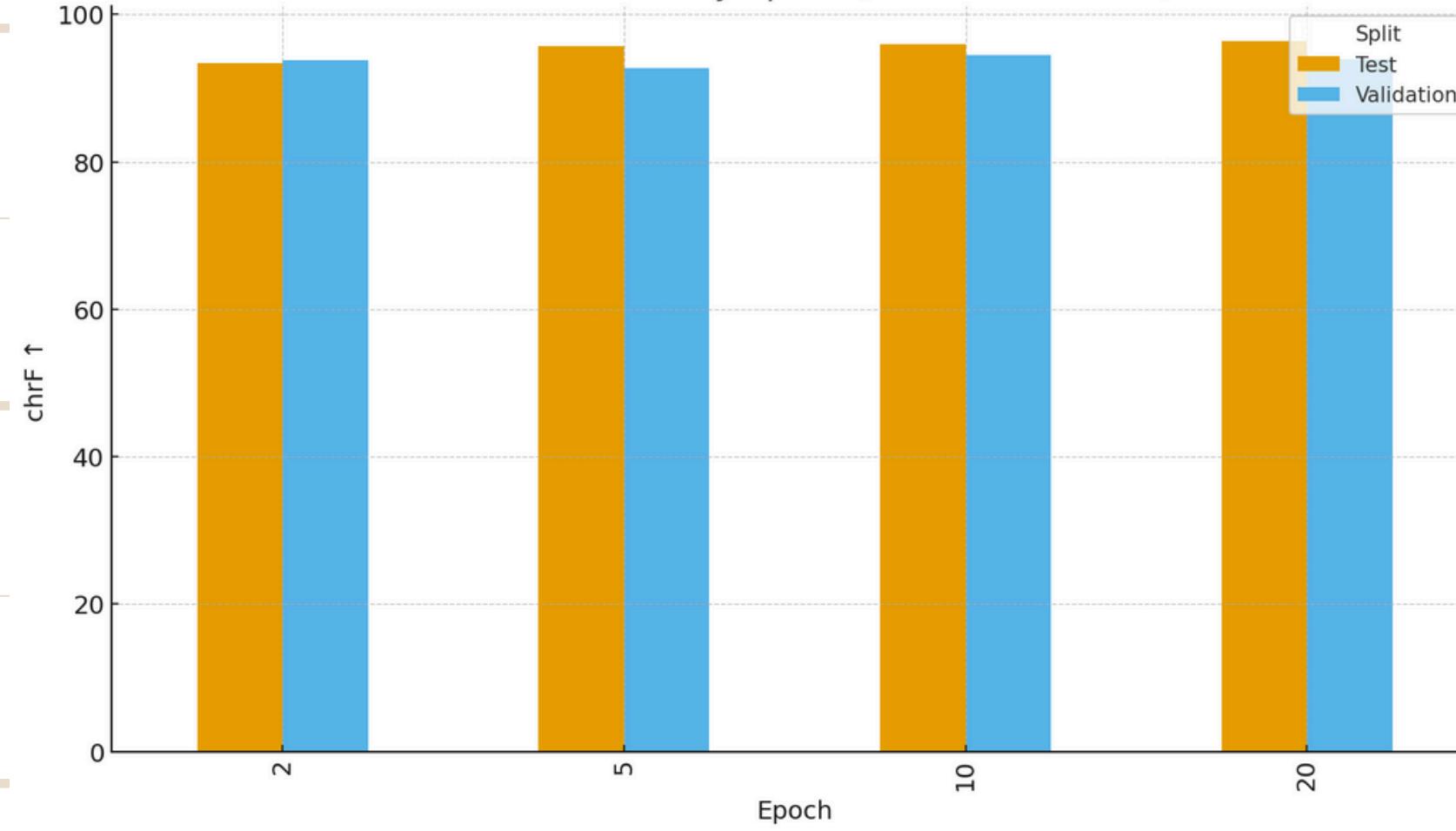
mT5: CER by Epoch (Validation vs Test)



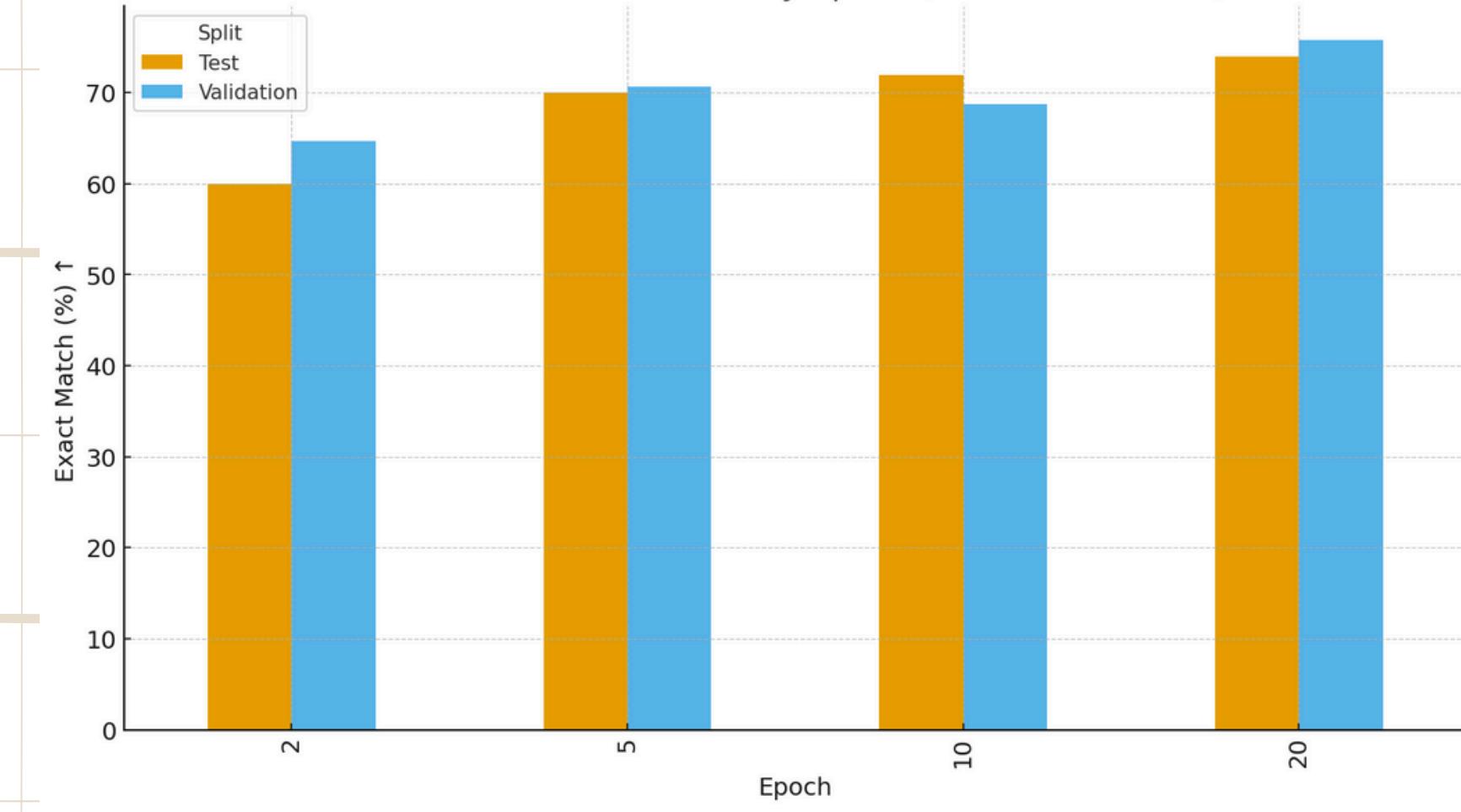
mBART-50: BLEU by Epoch (Validation vs Test)



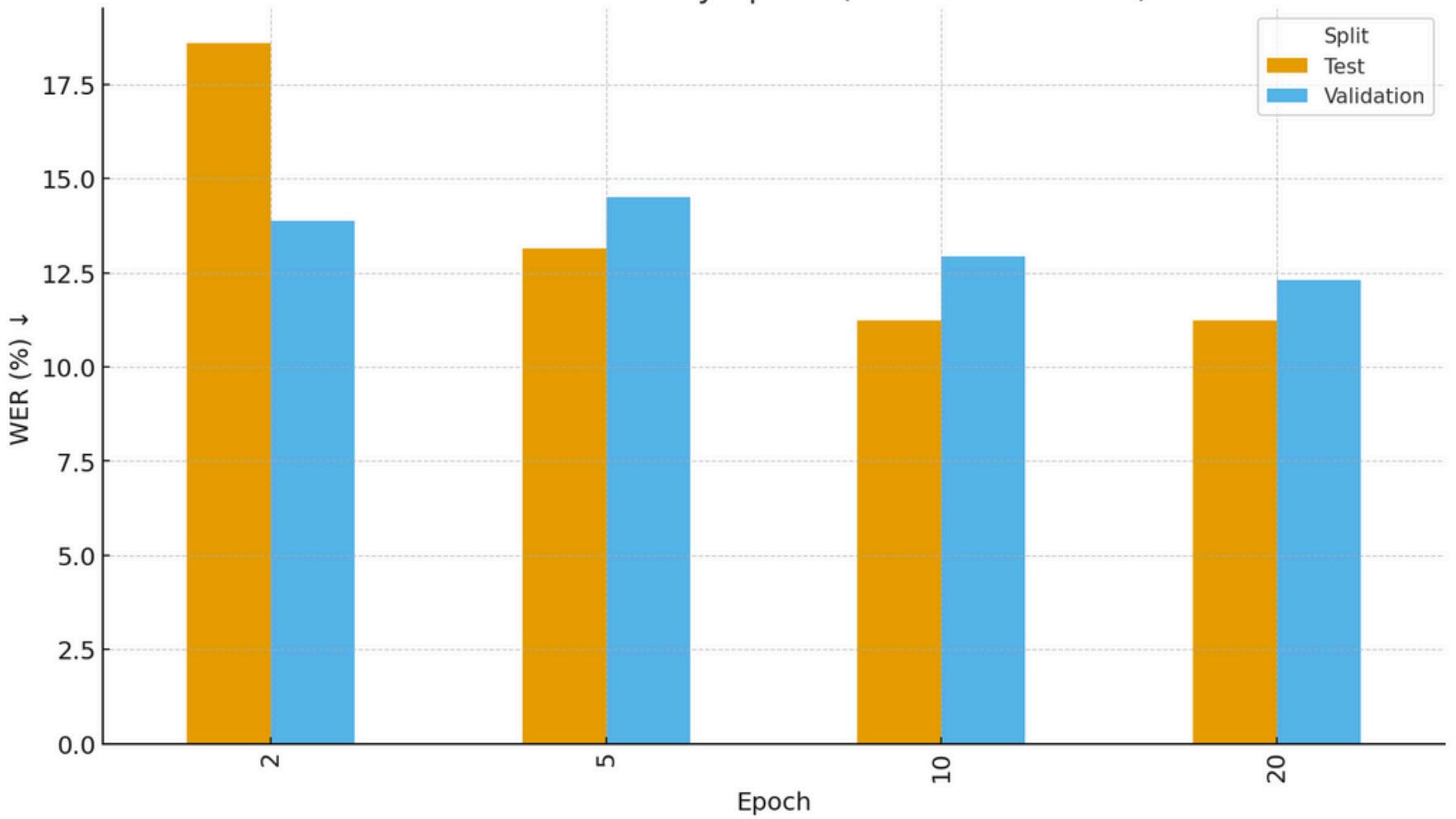
mBART-50: chrF by Epoch (Validation vs Test)



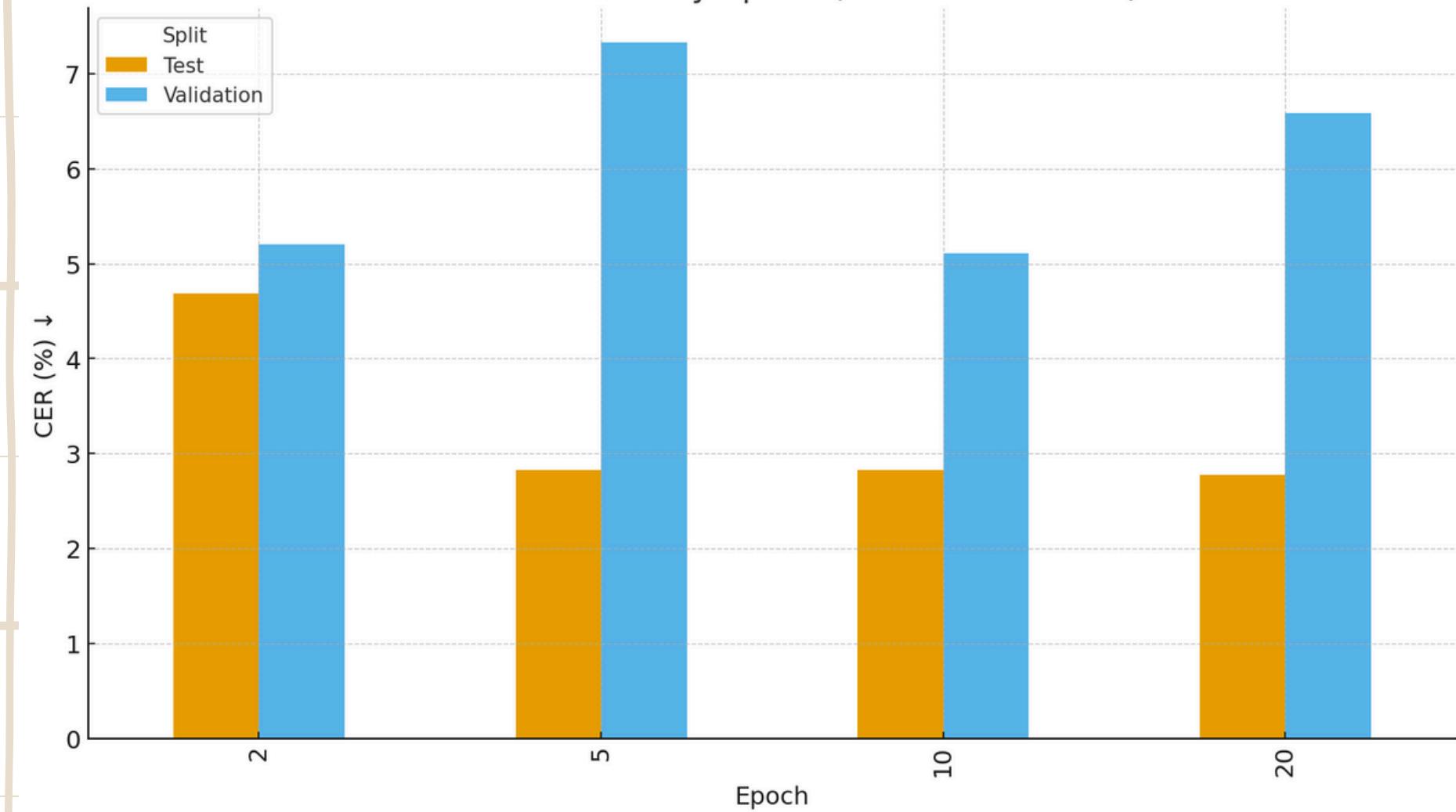
mBART-50: Exact Match by Epoch (Validation vs Test)



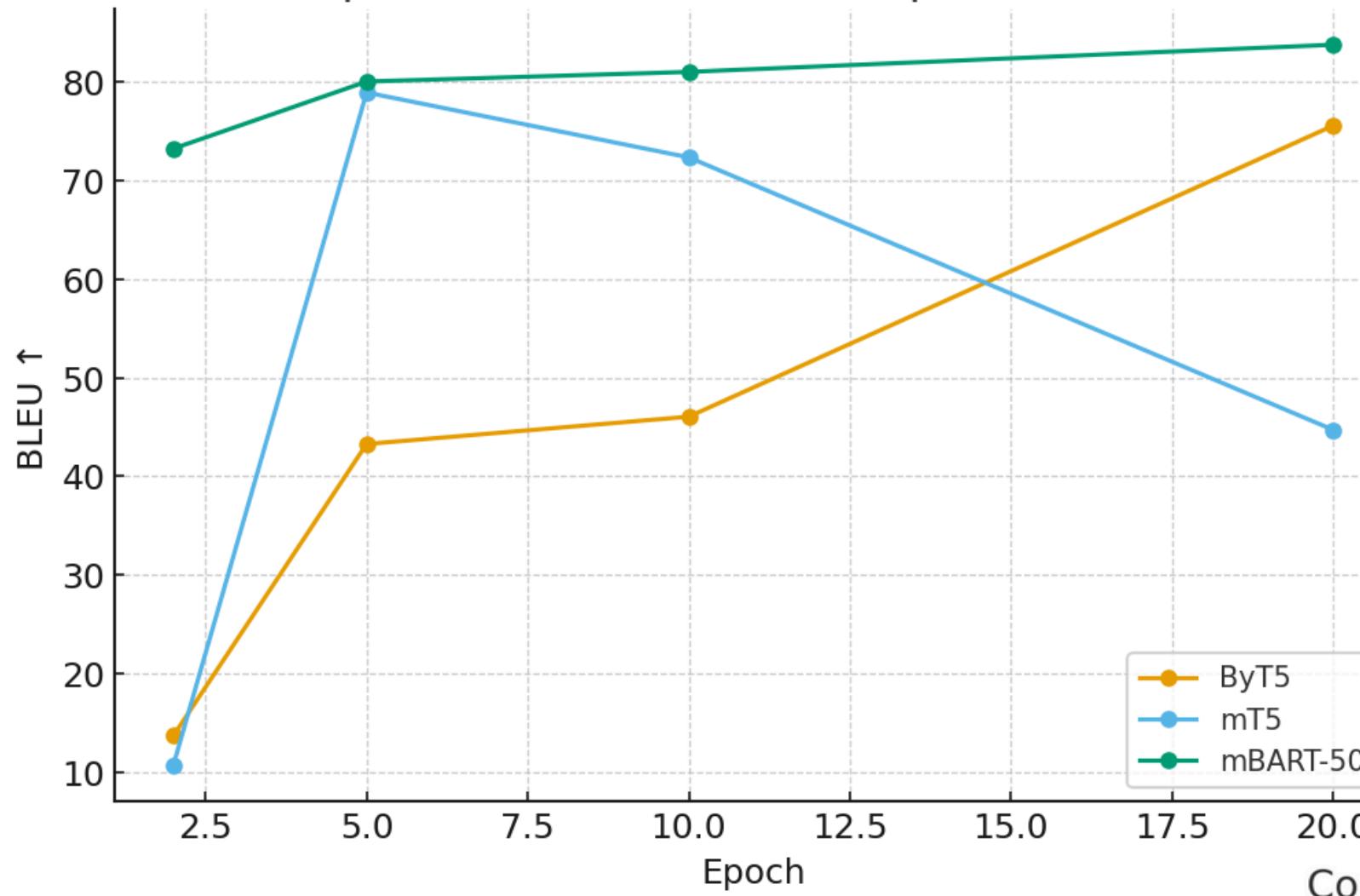
mBART-50: WER by Epoch (Validation vs Test)



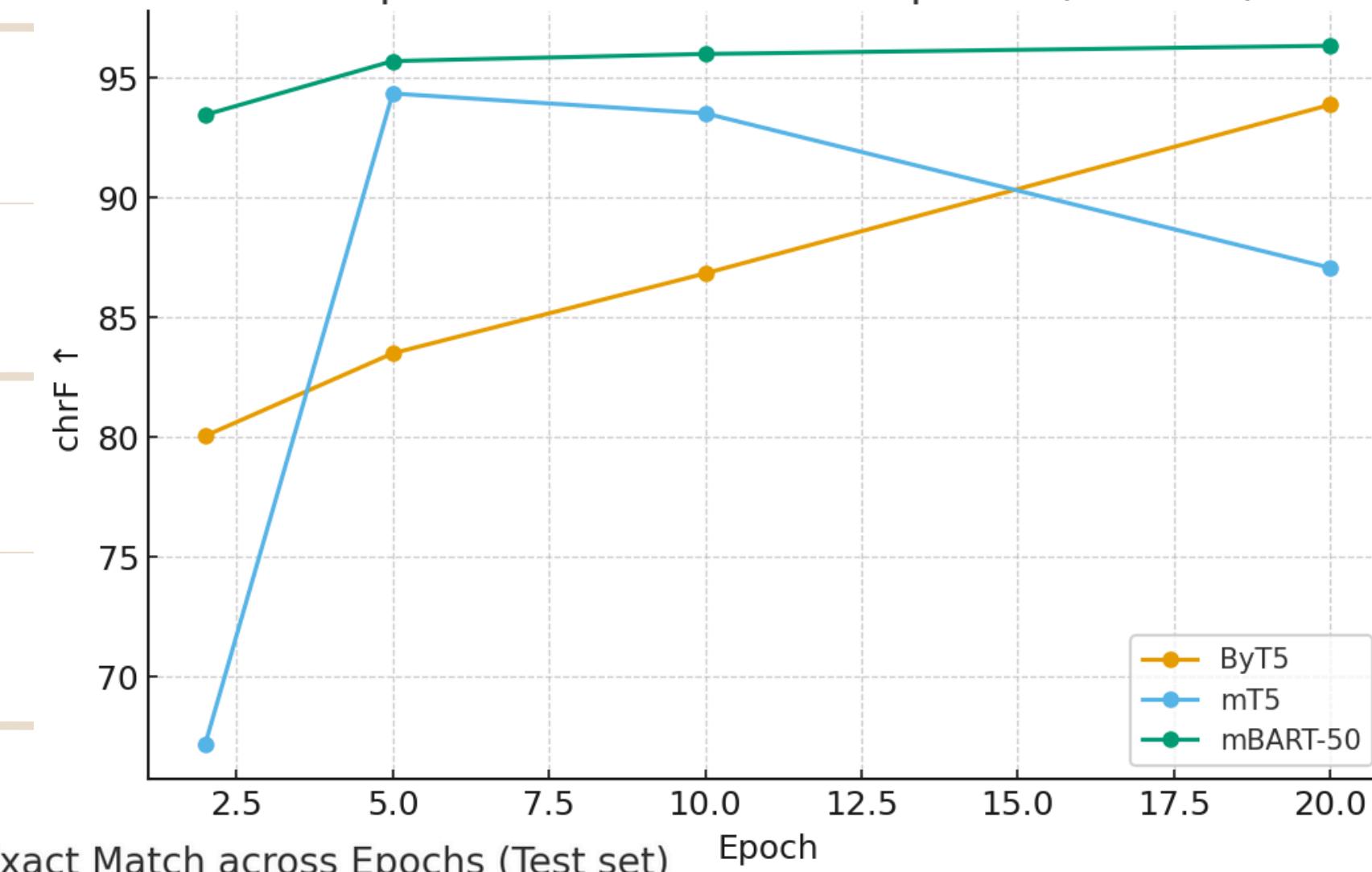
mBART-50: CER by Epoch (Validation vs Test)



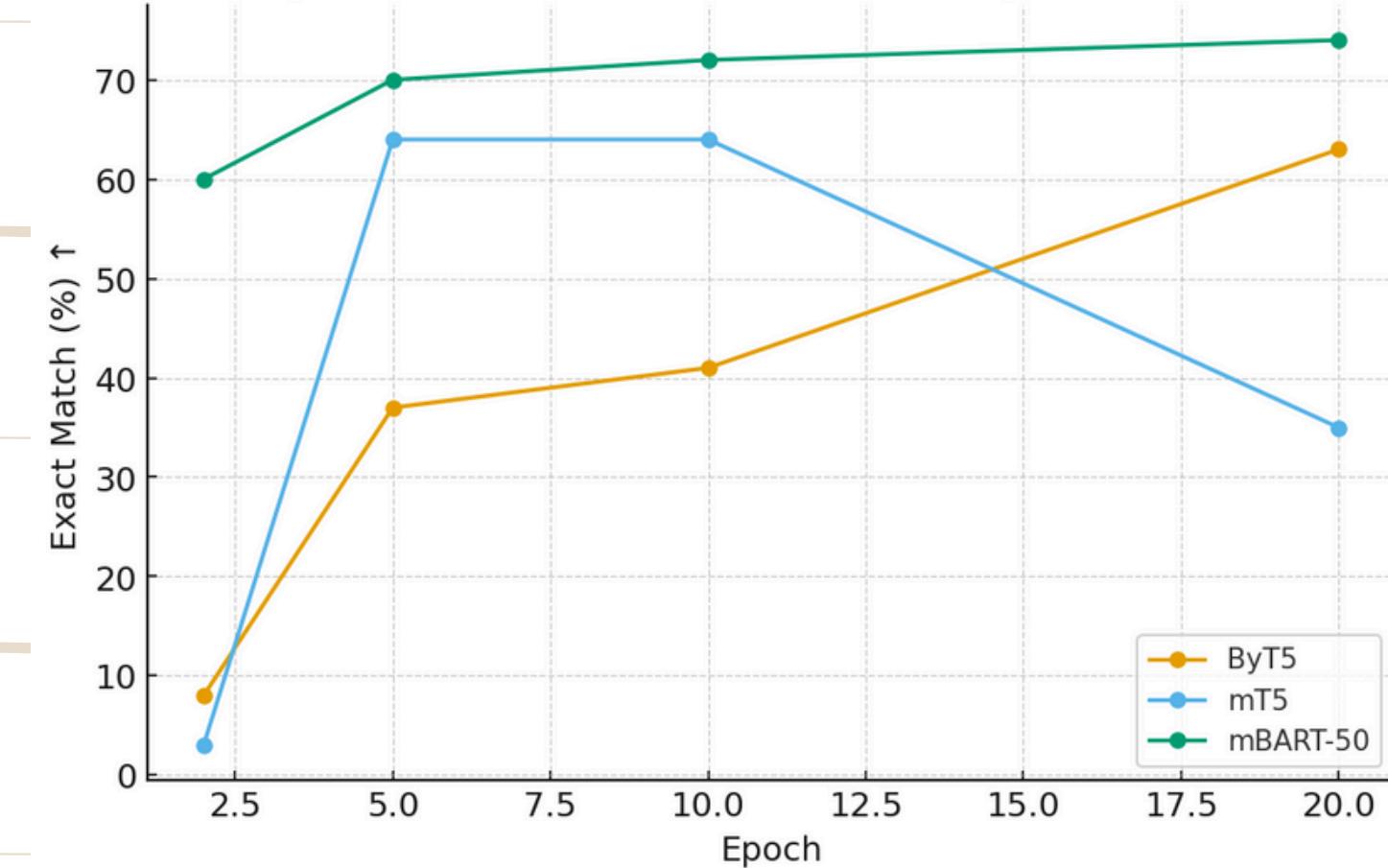
Comparison of BLEU across Epochs (Test set)



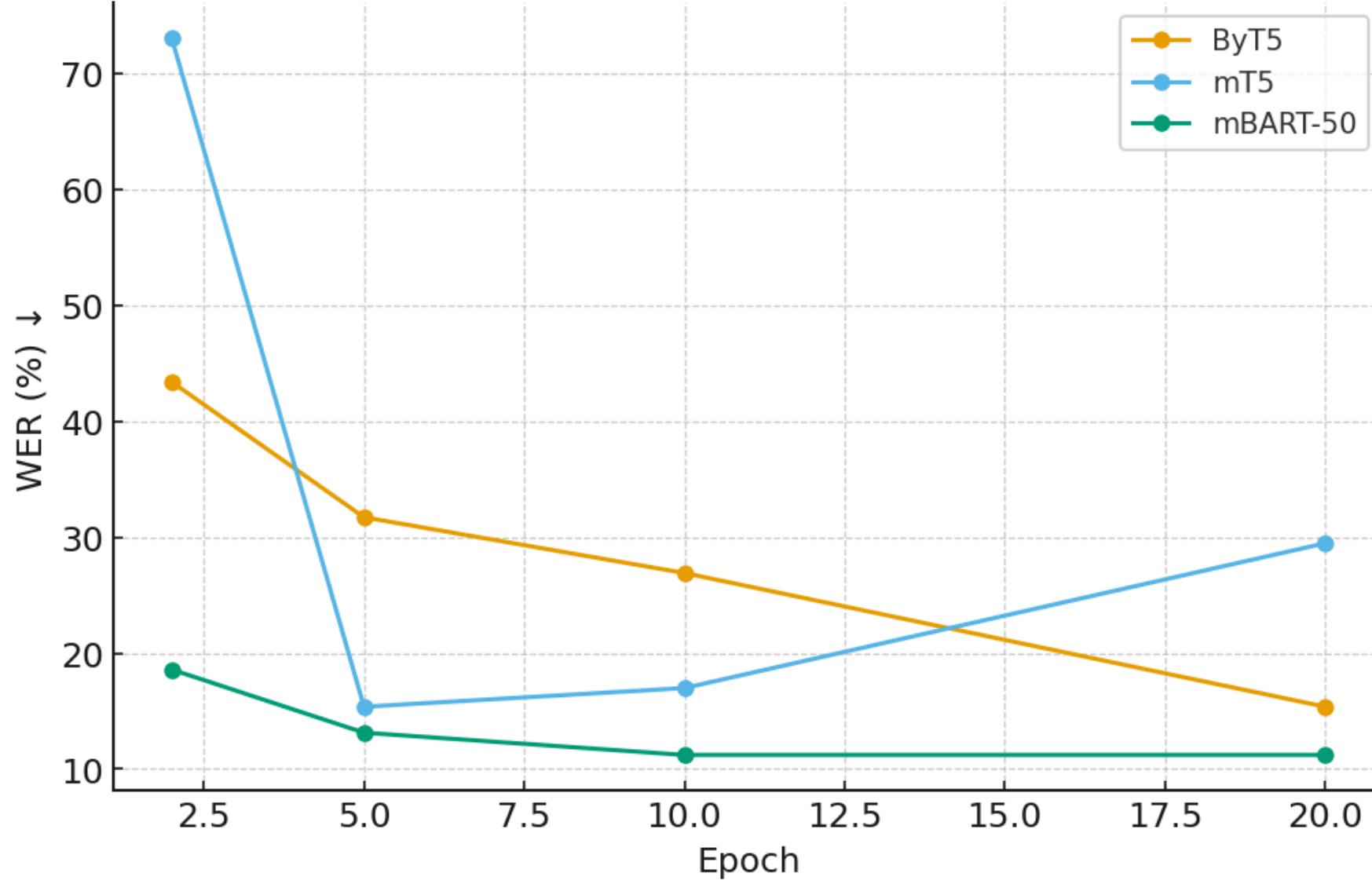
Comparison of chrF across Epochs (Test set)



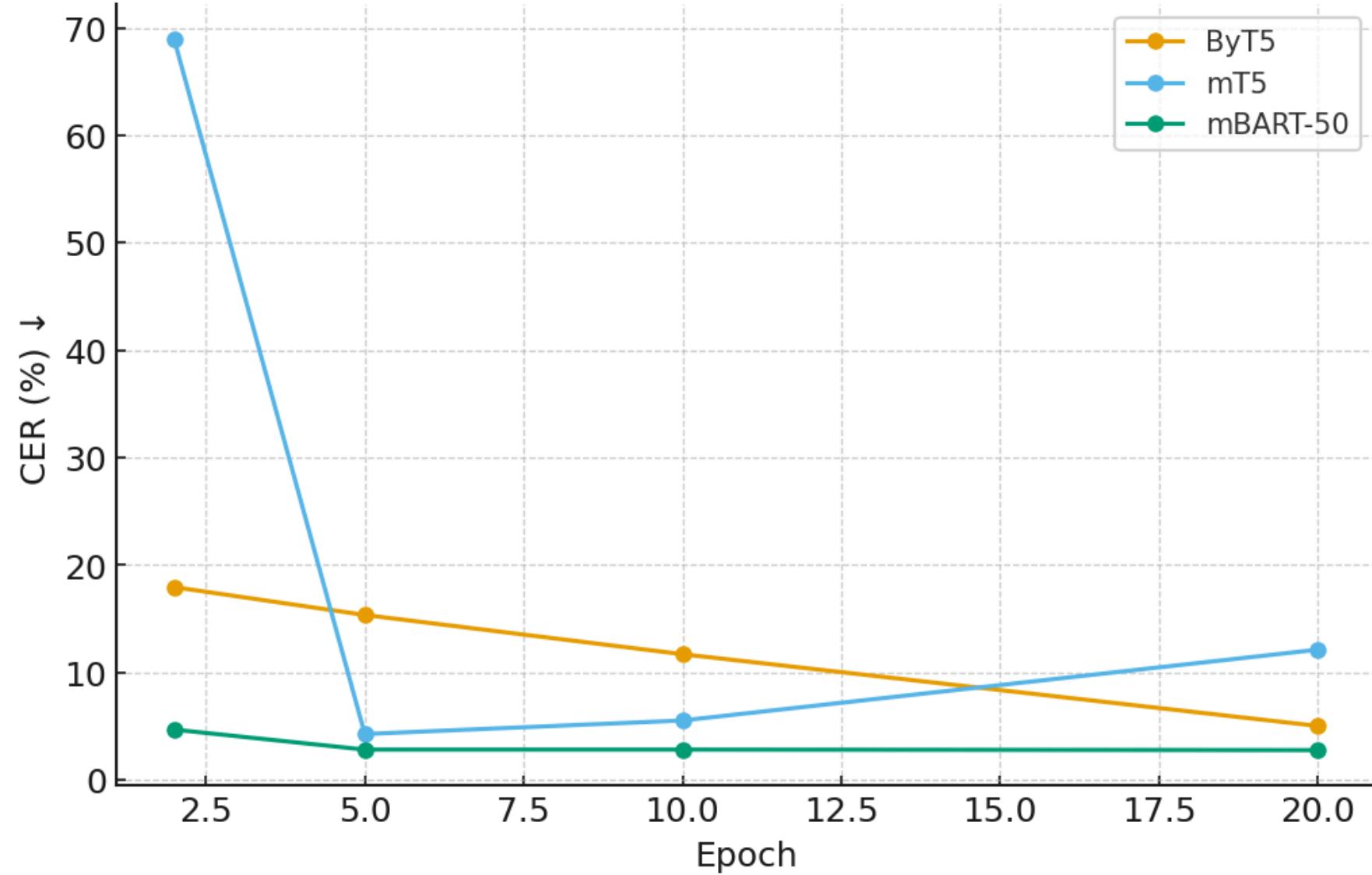
Comparison of Exact Match across Epochs (Test set)



Comparison of WER across Epochs (Test set)

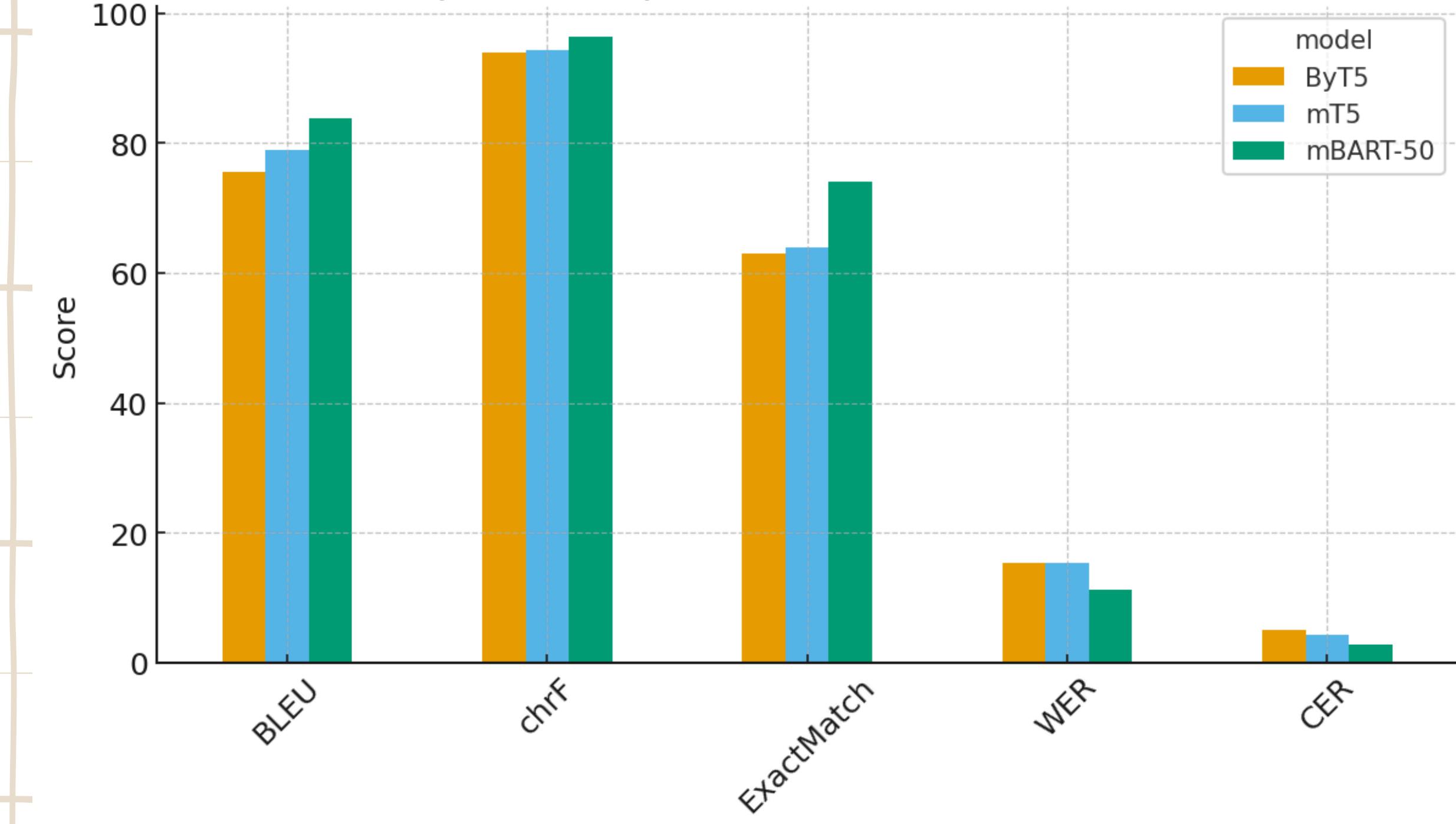


Comparison of CER across Epochs (Test set)



Comparison Aee Model Over Best Epoch

Best Epoch Comparison: Models vs Metrics (Test)



Conclusion

mBART-50 is the most consistent and best overall: it steadily improves with epoch and ends up highest on BLEU/chrF/Exact Match and lowest on WER/CER

ByT5 shows the largest improvement from early to late epochs and finishes close, but still trails mBART-50 on Exact Match and character-level quality

mT5 peaks early (\approx epoch-5) and then decrease by epoch-20 across most metrics

Future Work

In the future, want to try more Burmese grammar sentence structure using another models like XLM-RoBERTa + GECToR-style Tagger. Moreover, i want to use many training dataset for improving the model accuracy.

Thank
You!

