TDDE09 Project

CEFR-Oriented Controllable Sentence Simplification in Swedish

Using LLM-Generated Paraphrases:

A Low-Cost Approach to Controllable Swedish Text Simplification for Language Learners

Wenlin Fan, William Janowsky, Sergio Marín Muñoz, Xin Li

March 2026

Overview

01 Problem & Motivation

CEFR | Controllable Text Simplification | Problem Definition

02 Dataset Construction

Corpus | LLM simplification | prompt with attribution

03 Model Approach

Base model | Fine-tune approach

04 Evaluation

SARI | BERTScore | Complexity score | …

05 Results & Analysis

Model comparison and trade-off analysis

06 Future Work

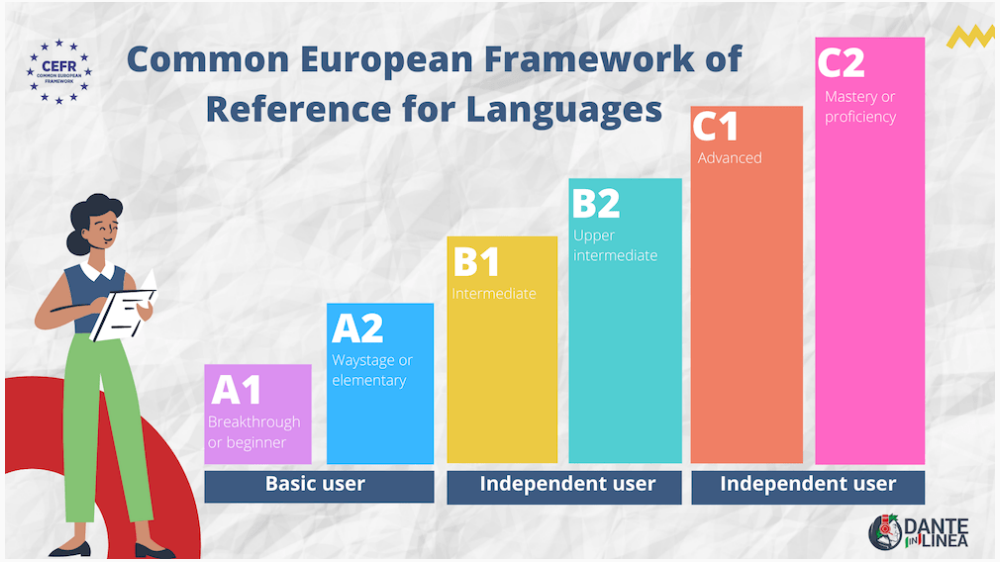
01 Problem & Motivation

CEFR | Controllable Text Simplification | Problem Definition

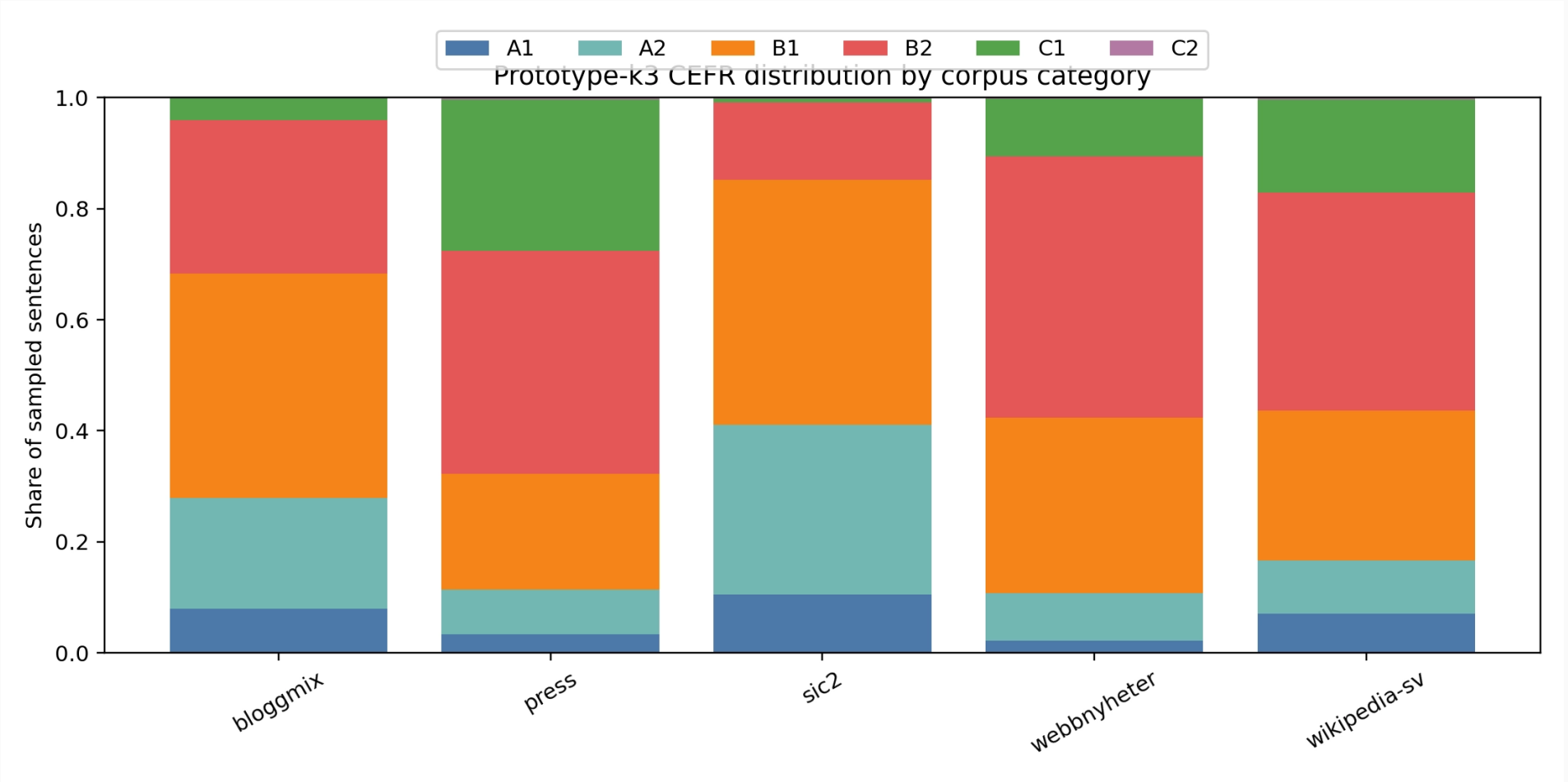
01 Problem & Motivation

Why This Research Matters

> **CEFR-Oriented Controllable Sentence Simplification in Swedish**
>
> **CEFR**(The Common European Framework of Reference for Languages) is a framework commonly used to measure proficiency levels of language learners.
>
> Public accessible content are mostly in B1/B2.
>
> Language resources targeting different **proficiency levels** are important for **language learners**
>
> **Text Simplification** could help.



Why This Research Matters



Research Problem Definition

First Problem Definition:

CEFR-Oriented Controllable Sentence Simplification in Swedish

From any levels to any levels

Straight forward Approach

Bummer!

corrupted content, meaning drifting, lack of accuracy, …

Research Problem Definition

Final Problem Definition:

CEFR-Oriented Controllable Sentence Simplification in Swedish

- Simplification Step: One level down each time
- Range: C1->B2, B2->B1, B1->A2

02 Dataset Construction

Corpus l LLM simplification l prompt with attribution

Dataset Construction Pipeline V1

Corpus Selection

SVT News
Daily news articles in Swedish

COC/TAILL
Textbook resources

SUC 3.0
Balanced Swedish corpus

Dataset Construction Pipeline V1

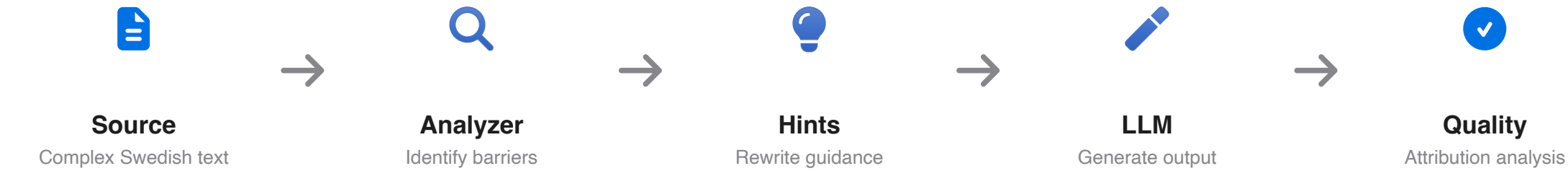
Corpus Selection

SVT News
Daily news articles in Swedish

COC/TAILL
Textbook resources

SUC 3.0
Balanced Swedish corpus

Dataset Construction Pipeline V2



Example Rewrite Hints

Passive → Active

Simplify Relative Clauses

Shorter Wording

....

Simpler Vocabulary

Analyzer Model & Prompt Design

Attribute with **a Logistic Regression Classifier**

Feature Extraction

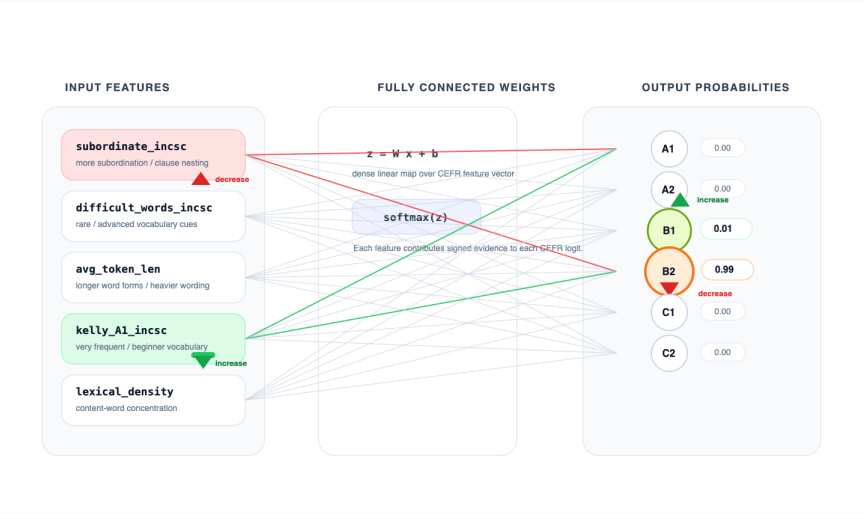
Extract 70 linguistic features from input sentence

Level Prediction

Probability of level c: $score_c(x) = w_c^\top x + b_c$

Feature Attribution

Contribution: $\Delta_j = (w_{s,j} - w_{t,j})x_j$



Prompt Example

System Prompt

You are rewriting Swedish source sentences to a precise CEFR A2 target. Keep every fact and relation from the source.

Target-level requirements:

- …

User Prompt

Rewrite from CEFR B1 to A2:

"Boken som jag lånade på biblioteket var mycket intressant."

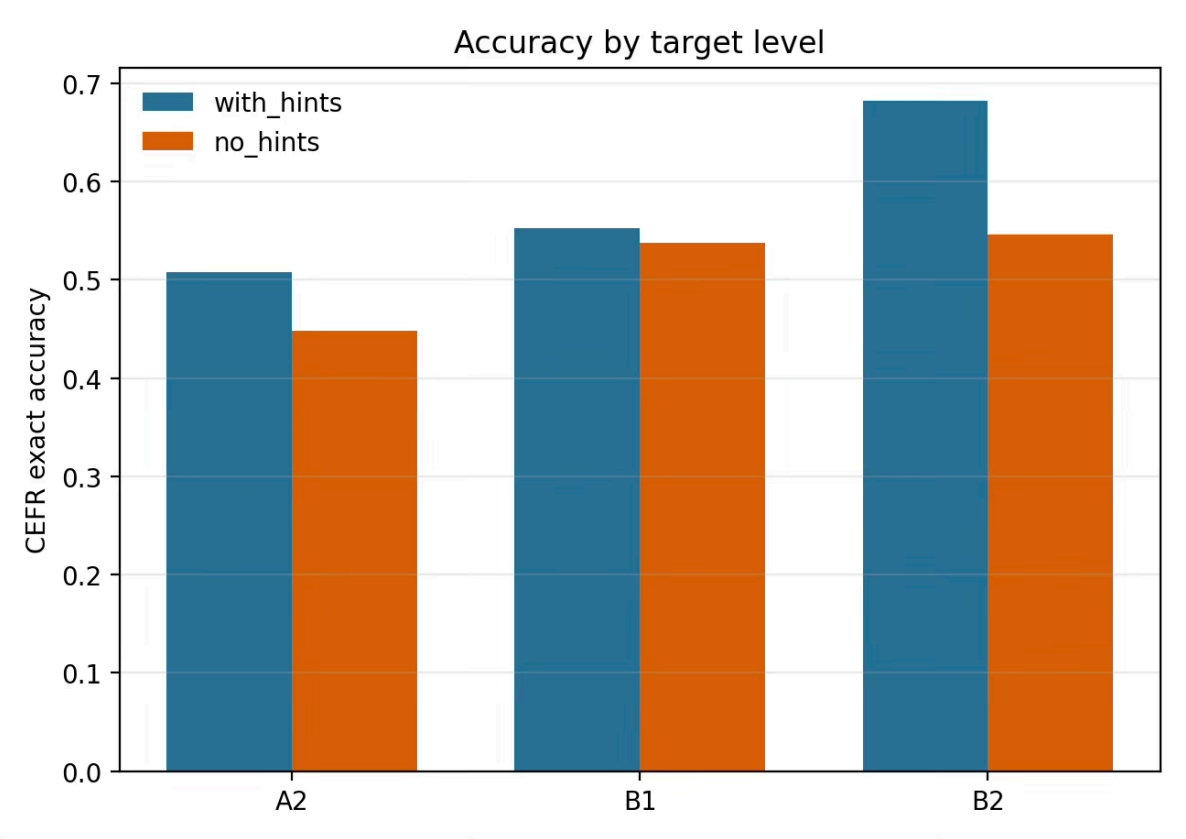
Rewrite Hints:

1. Rewrite the **relative clause** "som jag lånade på biblioteket" as a simpler main clause or a separate sentence.
2. …

Expected Output JSON format:

{"text": "…"}

Analyzer Model & Prompt Design

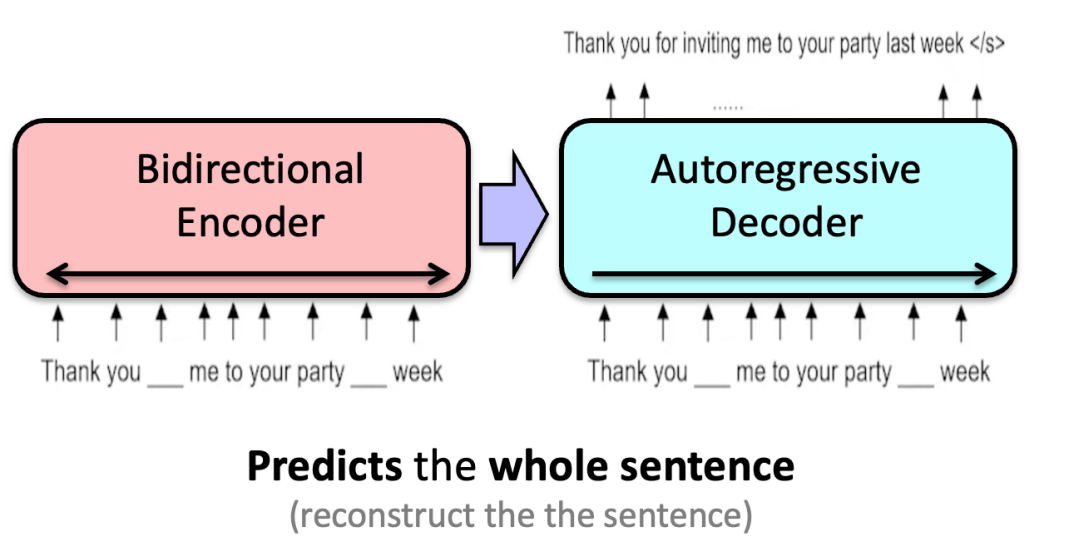


03 Model Approach

Base model | Fine-tune approach

Base Model

BART

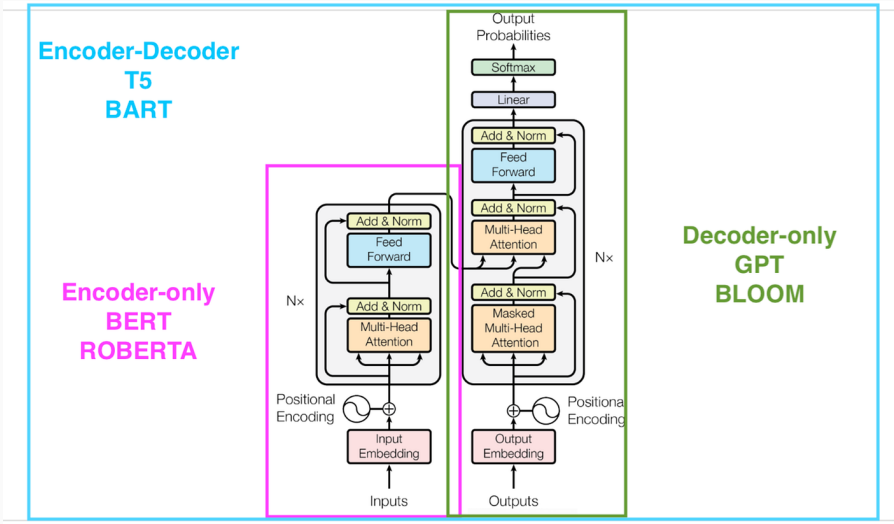


Seq2seq architecture

Pretrained to recover noised sentence

Suitable for text generation tasks

GPT



Pretrained to generate text

Base Model

Fine-tune GPT

Supervised fine-tuning with concatenated prompt prefix

Base Model:

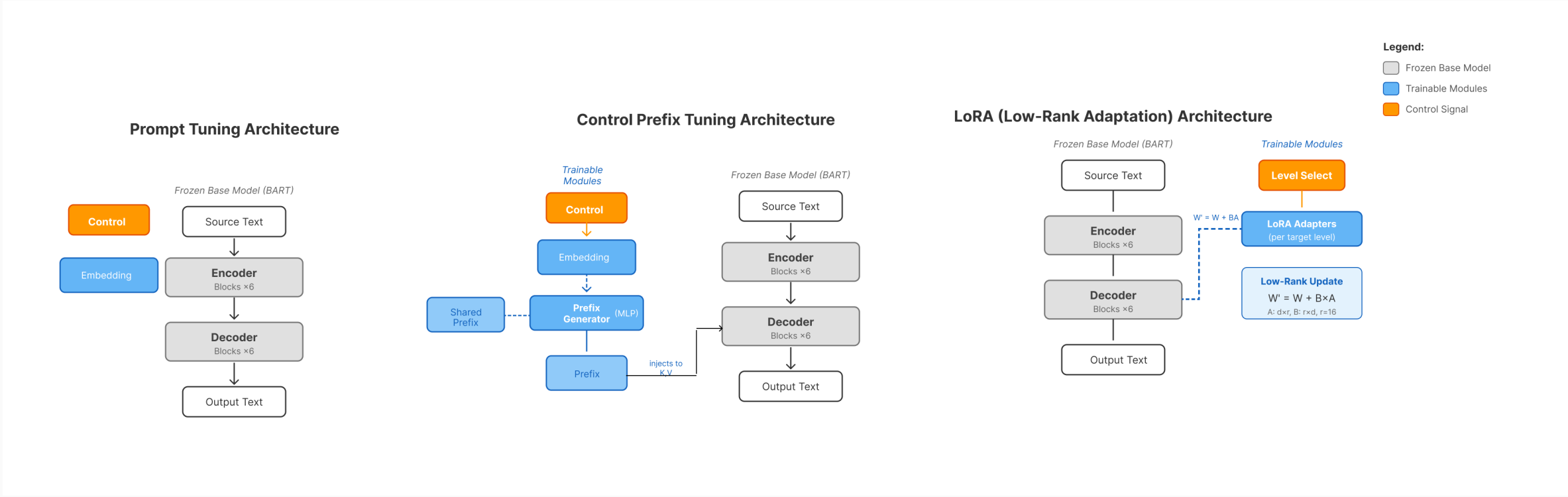
Llama 3.2 3b: an open-sourced pretrained multi-lingual model from Meta

GPT-sw3 6.7b: an open-sourced model pretrained on Swedish

03 Model Approach

Base Model

Fine-tune BART



Base Model: KBLab/bart-base-swedish-cased

A pretrained Swedish BART model

04 Evaluation

SARI | BERTScore | Complexity score | …

Evaluation Framework

General Metrics

SARI

Simplification quality (add, keep, delete)

BERTScore

Semantic preservation

LIX

Readability score

Control Metrics

Control accuracy

Exact target level hit rate

Ordered-probit Complexity score

A finer scale of complexity

Convertible to CEFR-levels

Evaluate simplification direction, targeting distance, …

Diagnostics

Copy Rate

Source text copying

PPL

Perplexity (fluency)

Similarity

Source-target similarity

04 Evaluation

SARI & BERTScore

SARI: System Against References & Input

Based on n-gram (1-4 gram) edit operations from source text, divided into three sets:

KEEP
Preserve n-grams

DELETE
Remove n-grams

ADD
Insert n-grams

Example:

Source:
Hon påbörjade arbetet tidigt.

Reference:
Hon började arbetet tidigt.

Operations:
- Hon, arbetet, tidigt
- påbörjade
- började

BERTScore: Semantic Preservation

Measures semantic similarity using contextual word embeddings:

$$S \in R^{m \times n}, S_{ij} = \cos(\tilde{x}_i, \tilde{y}_j)$$

Cosine similarity between word embeddings

$$\text{Recall} = \frac{1}{n} \sum_{j=1}^{n} \max_{1 \le i \le m} S_{ij}$$

Coverage of reference

$$\text{Precision} = \frac{1}{m} \sum_{i=1}^{m} \max_{1 \le j \le n} S_{ij}$$

Accuracy of prediction

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Harmonic mean of precision and recall

04 Evaluation

Control Metrics

Control Accuracy

Based on a CEFR-level classifier

$$Accuracy = \frac{\sum (cls(predicted) = cls_{target})}{|S_{test}|}$$

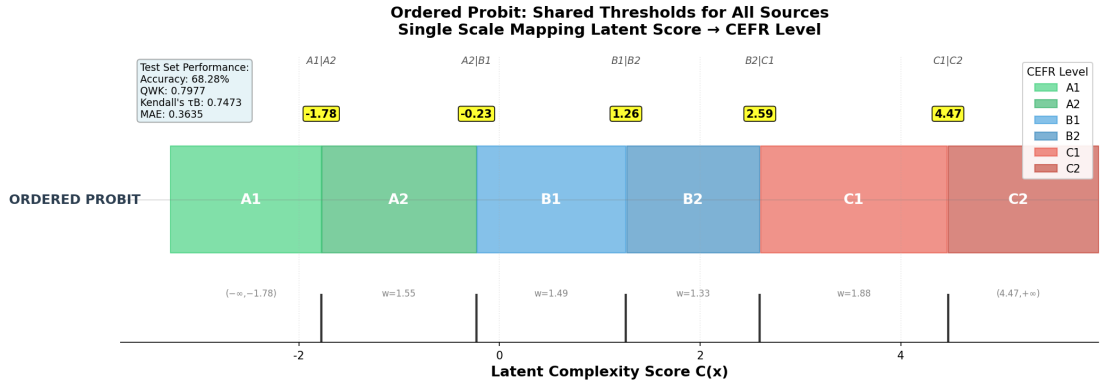
Ordered-Probit Complexity Score

Measurement on a continuous score C(x)

Target Complexity Error $\quad TCE = |C_{output} - C_{target\_mid}|$

Direction Accuracy $\quad DA = \frac{1}{N}\sum_{i=1}^{N}[C^{(i)}_{\text{output}} < C^{(i)}_{\text{source}}]$

Expected Level Error, Over-Simplification Rate, Simplification Gain Score, …



05 Results & Analysis

Model comparison and trade-off analysis

Main Results: Model Comparison

Compact Step-Down V2 Raw-Metric Table
Bold green = best, pale yellow = second-best, SARI* = secondary metric

route	model	SARI*	control accuracy	within tgt	TCE	ELE	src sim	copy	exact copy	PPL
BART-LoRA	bart-lora r64-e3	50.30	0.388	0.405	1.017	0.682	0.954	0.685	0.085	170.3
BART-LoRA	bart-lora r16-e3	49.70	0.362	0.379	1.053	0.706	0.960	0.716	0.100	170.6
Decoder-only	sw3-6.7b decoder-ft	38.45	0.592	0.402	1.011	0.679	0.850	0.180	0.000	84.7
Decoder-only	llama3.2-3b decoder-ft	36.05	0.493	0.402	0.986	0.662	0.833	0.189	0.000	163.4
BART-discrete	bart-discrete s30-c20	44.34	0.369	0.386	1.163	0.772	0.913	0.697	0.150	223.4
BART-prompt	bart-prompt vtok10-lr0.2	41.56	0.376	0.368	1.103	0.744	0.929	0.712	0.095	224.2

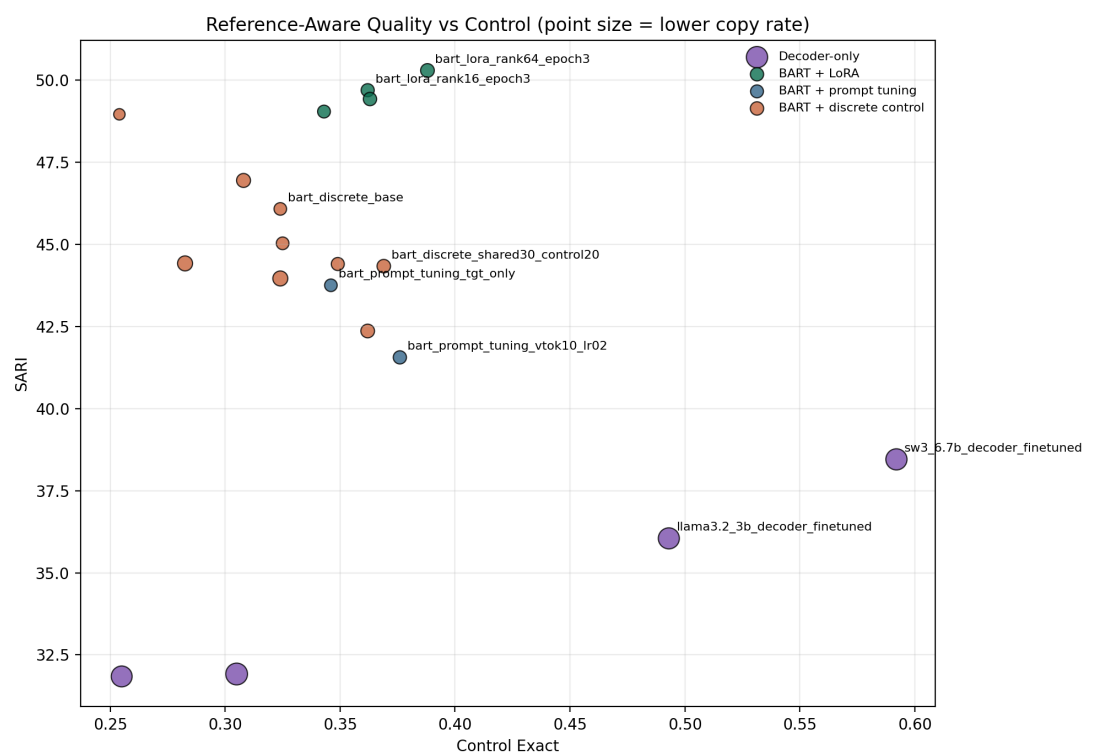
Conclusions:

1. BART-LoRA outperformed other fine-tuned BART models in SARI
2. Comparing to GPT, BART tend to rewrite conservatively (with minimum edition) and copy more.
3. GPT simplifications achieve higher accuracy to targeting CEFR-levels but tend to drift in meaning

Further Questions:

1. Is BART-LoRA really better than GPT?
2. Why GPT can achieve higher accuracy while hitting a low SARI score?

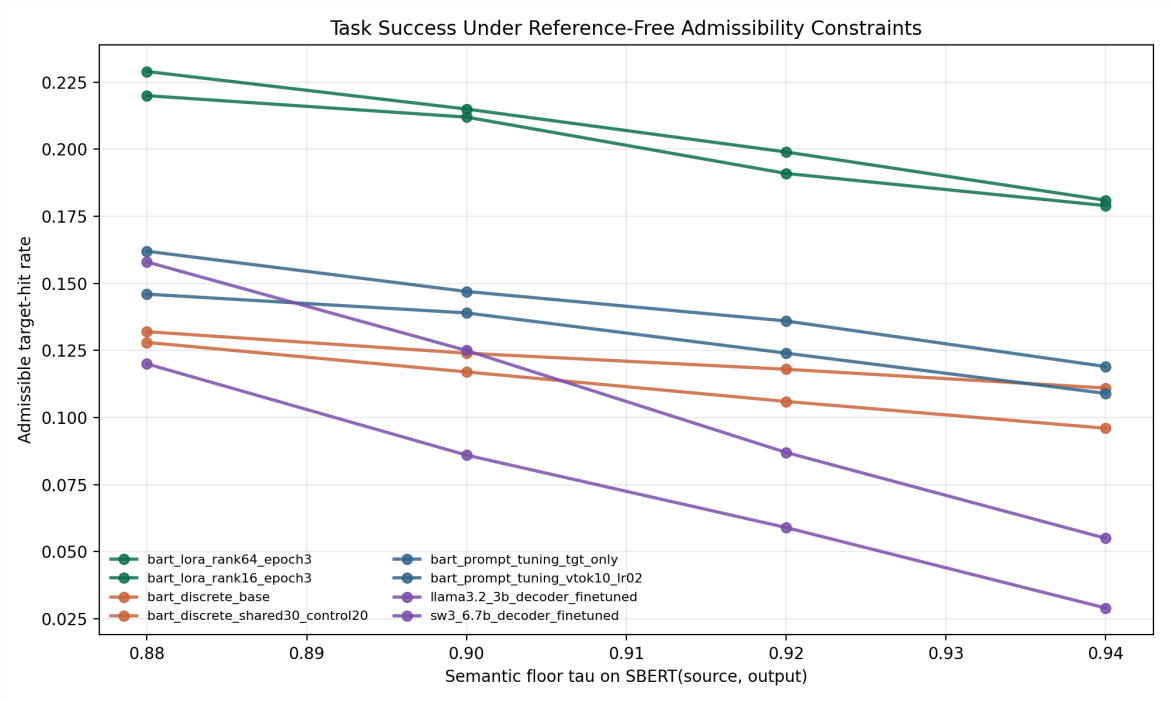
Further analysis: SARI vs Control Accuracy



Observations:

SARI and Control Accuracy seem to be controversial.

Further analysis: Admissible Accuracy



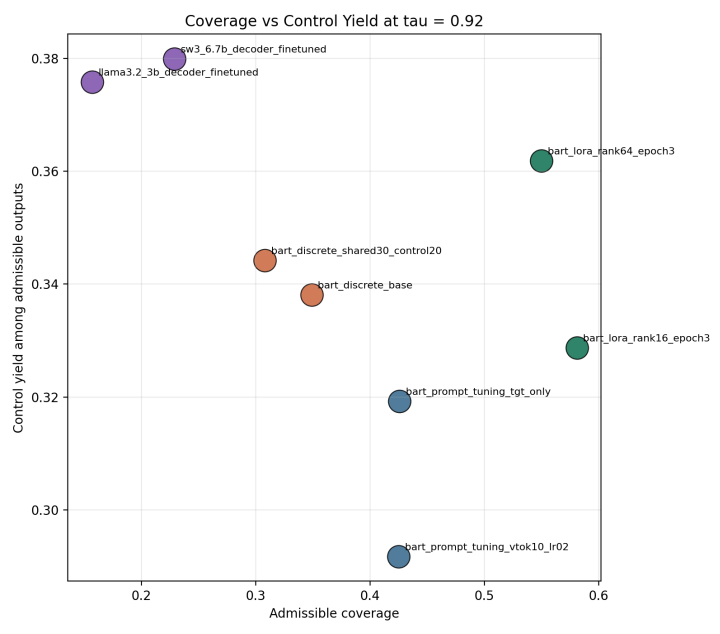
Add Constraints:

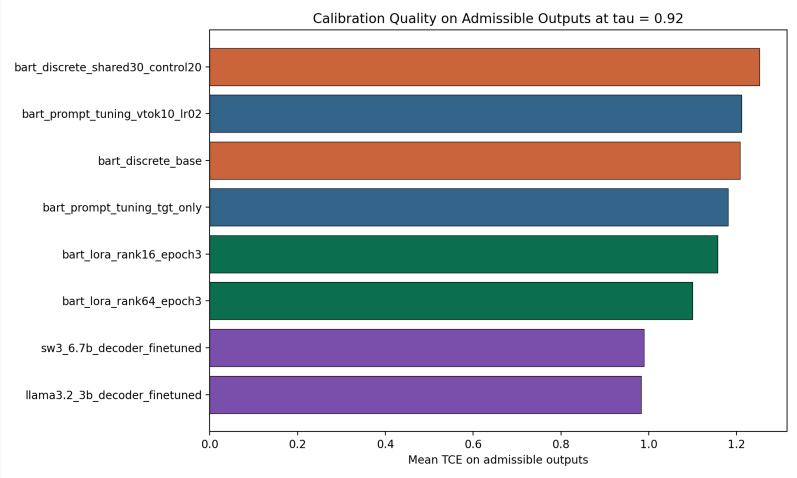
Admissible: SBert(source, output) > $threshould

Observations:

GPT models are more sensitive to threshold.

Further analysis: Accuracy vs Admissibility





Measurements:

Admissible coverage: Proportion of admissible simplification at specific tau.

Control Yield: Control Accuracy within admissible simplifications.

Observations:

GPT models hit a low admissible coverage while achieving a high accuracy.

Within admissible outputs, GPT models achieve lower TCE.

Further analysis

Compact Step-Down V2 Table
Bold green = best, pale yellow = second-best, SARI* = secondary metric

route	model	control accuracy	within tgt	cov.@0.92	task succ.@0.92	yield@0.92	safe TCE	src sim	SARI*
BART-LoRA	bart-lora r64-e3	0.388	0.405	0.550	0.199	0.362	1.100	0.954	50.30
BART-LoRA	bart-lora r16-e3	0.362	0.379	0.581	0.191	0.329	1.157	0.960	49.70
Decoder-only	sw3-6.7b decoder-ft	0.592	0.402	0.229	0.087	0.380	0.989	0.850	38.45
Decoder-only	llama3.2-3b decoder-ft	0.493	0.402	0.157	0.059	0.376	0.983	0.833	36.05
BART-discrete	bart-discrete s30-c20	0.369	0.386	0.308	0.106	0.344	1.252	0.913	44.34
BART-prompt	bart-prompt vtok10-lr0.2	0.376	0.368	0.425	0.124	0.292	1.211	0.929	41.56

Conclusions:

1. BART-LoRA achieves the best in end2end success rate.
2. Within admissible outputs, GPT models still keep a high control accuracy.

Examples

keep | delete | add

BART side	GPT / decoder-only side
Det är sent också, och du ska till skolan i morgon. it is late also , and you shall to school in morning. **Det är sent nu. Du ska till skolan i morgon.** it is late now . you shall to school in morning.	**Det är sent också, och du ska till skolan i morgon.** it is late also , and you shall to school in morning . **Det är sent nu. Du ska gå till skolan imorgon.** it is late now . you shall go to school tomorrow .
Vi kan hyra sådana där stora partytält, så gör det inget om det regnar. we can rent those there large party-tents , then makes it nothing if it rains. **Vi kan hyra stora partytält. Då gör det inget om det regnar.** we can rent large party-tents . then makes it nothing if it rains.	**Vi kan hyra sådana där stora partytält, så gör det inget om det regnar.** we can rent those there large party-tents , then makes it nothing if it rains. **Vi kan hyra ett stort tält. Då spelar det ingen roll om det regnar ute.** we can rent a large tent . then plays it no role if it rains outside .
År 1964 kom hans första LP-skiva Ballader och oförskämdheter ut. in 1964 came his first LP-record Ballads and Impudences out . **År 1964 kom hans första skiva. Den hette Ballader och oförskämdheter.** in 1964 came his first record . it was-called Ballads and Impudences.	**År 1964 kom hans första LP-skiva Ballader och oförskämdheter ut.** in 1964 came his first LP-record Ballads and Impudences out . **Hans första skiva kom år 1970. Den heter Ballader och oförskämt.** his first record came in 1970 . it is-called Ballads and impolitely .

06 Future Work

Further Working Directions:

Reinforcement learning

CoT(Chain-of-Thought) Training with model attributions, suggestions

Passage-level simplification

Thank You

Any Questions?