Assessing the Readability of German Sentences with Transformer Ensembles

AComplexity at GermEval 2022

Patrick Gustav Blaneck
IT Center RWTH Aachen,
Aachen University of Applied Sciences

Tobias Bornheim
ORDIX AG,
Aachen University of Applied Sciences

Niklas Grieger
Institute for Data-Driven Technologies,
Aachen University of Applied Sciences

Stephan Bialonski
Institute for Data-Driven Technologies,
Aachen University of Applied Sciences

Dataset / Task

- 1000 labelled German sentences
- Sourced from 23 Wikipedia articles
- Evaluated by German native-speakers on scale from 1-7

- Task: Predict the mean complexity score of a given sentence.

What Motivated Our Approach

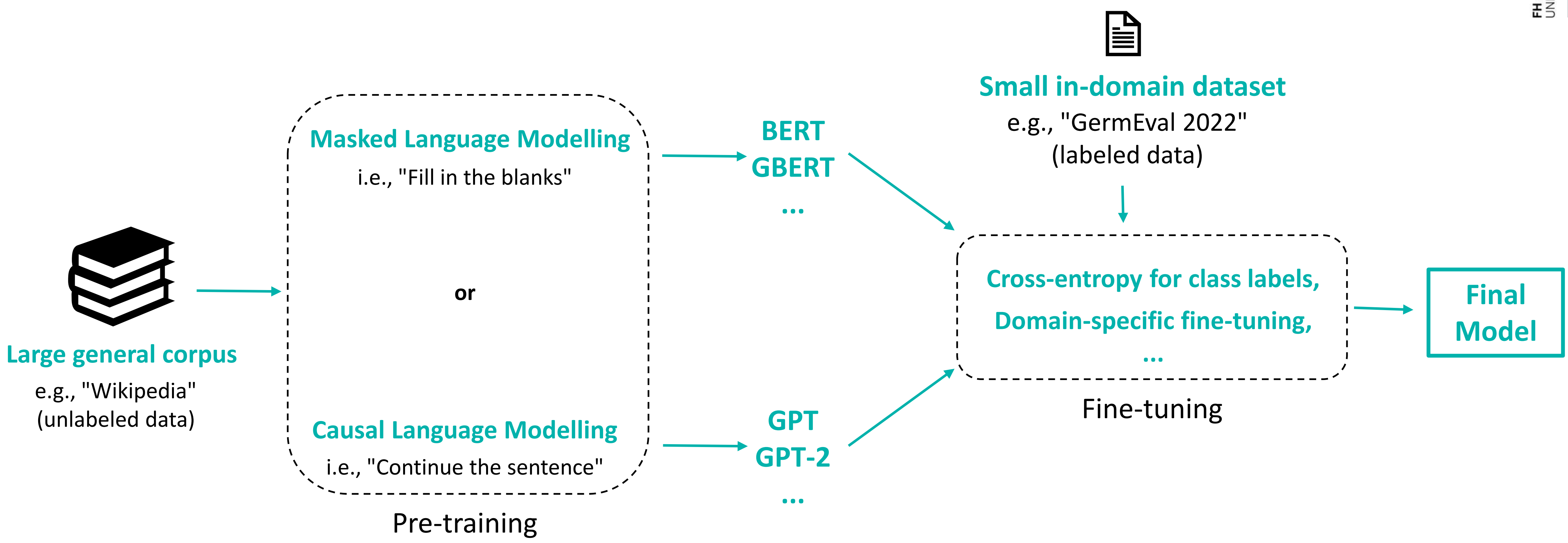
Q1: How does the size of an LLM effect readability assessment?

Q2: Can traditional readability features be incorporated to boost performance?

Does combining multiple models in ensembles increase performance? If so:

- Q3: What happens if different model types are combined?
- Q4: How many models should an ensemble contain?

Large Language Models



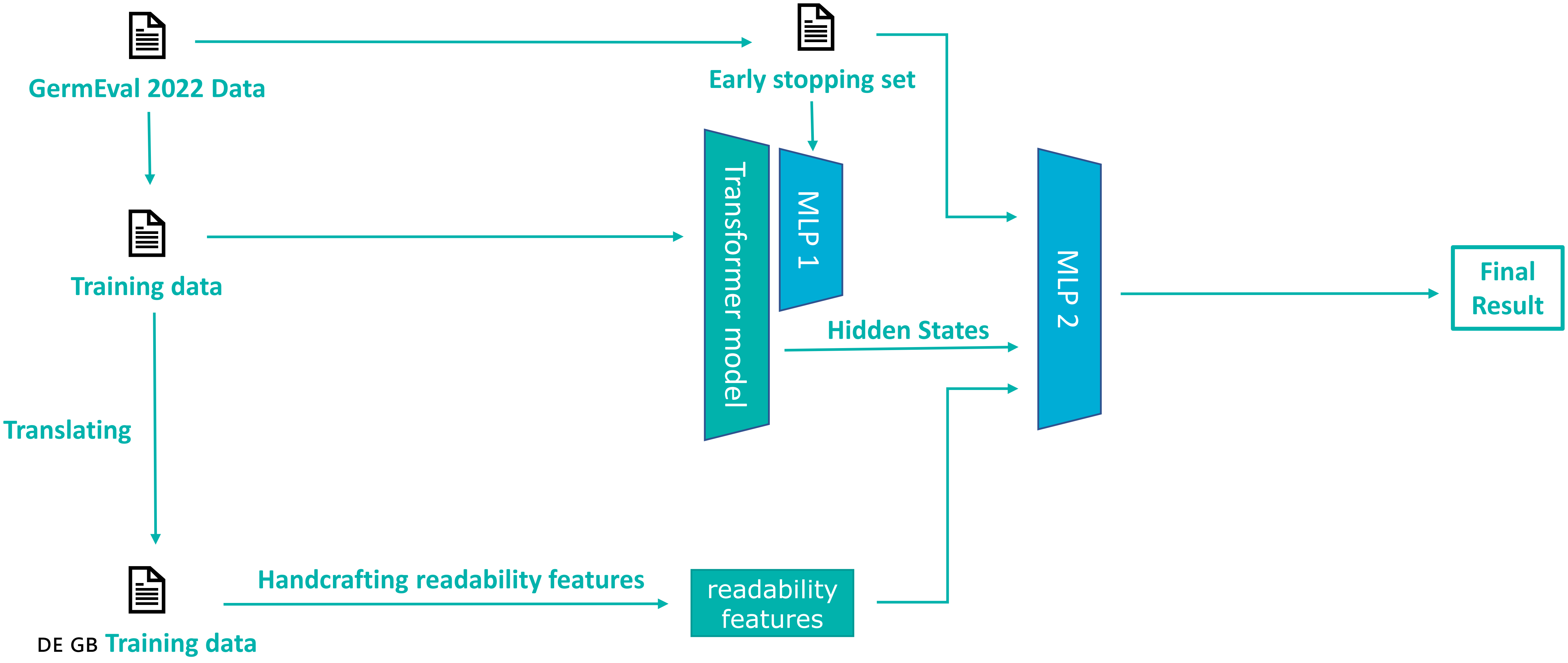
Large Language Models

- GBERT:[1]
  - Pre-trained weights: [deepset/gbert-large]
  - # of parameters: 340M
  - File size: approx. 1.4 GiB
- GPT-2-Wechsel:[2]
  - Pre-trained weights: [malteos/gpt2-xl-wechsel-german]
  - # of parameters: 1.5B
  - File size: approx. 6.3 GiB

HUGGING FACE

[1] Chan et al., German's Next Language Model (2020) [GBERT by deepset]
[2] Minixhofer et al., WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language (2021)

Training



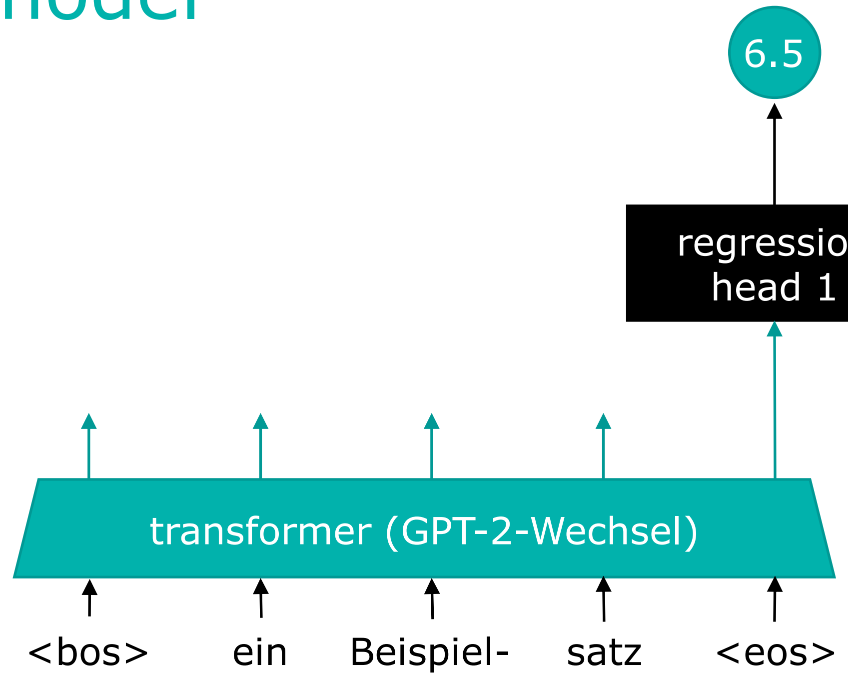
GermEval 2022 Data
Early stopping set
Training data
Transformer model
MLP 1
MLP 2
Hidden States
Final Result
Translating
Handcrafting readability features
readability features
DE GB Training data

Readability features

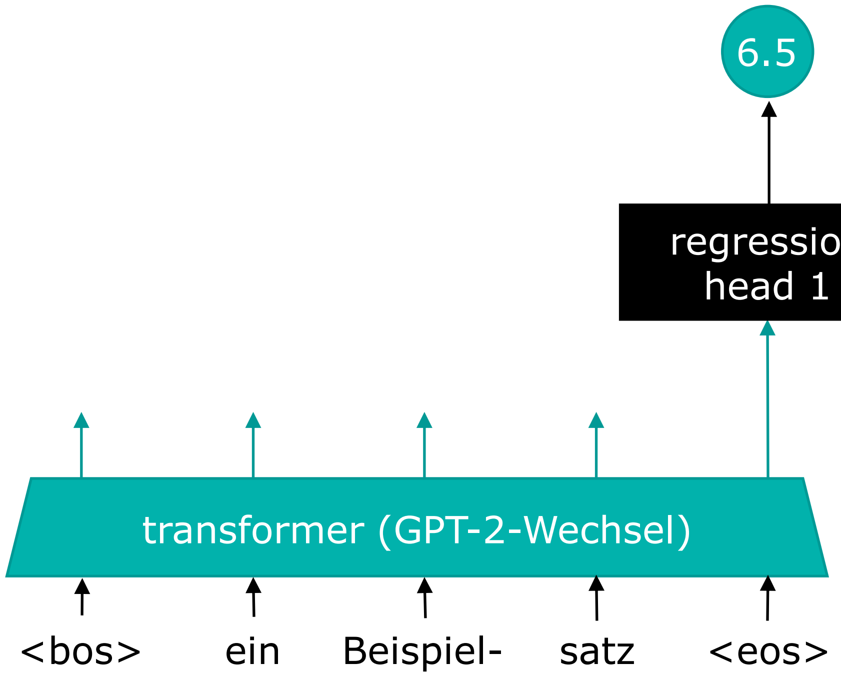
- Surface-based:
  - sample size, vocabulary size
- Sentence-based:
  - length, punctuation
- Dependency-based:
  - dependency distance, dependencies per token
- Constituency-based:
  - height syntax tree, clauses per sentence
- ...

Training — Finetuning

- transformer + regression head
- regression head: multi-layer perceptron (MLP)
- finetuning of the whole model

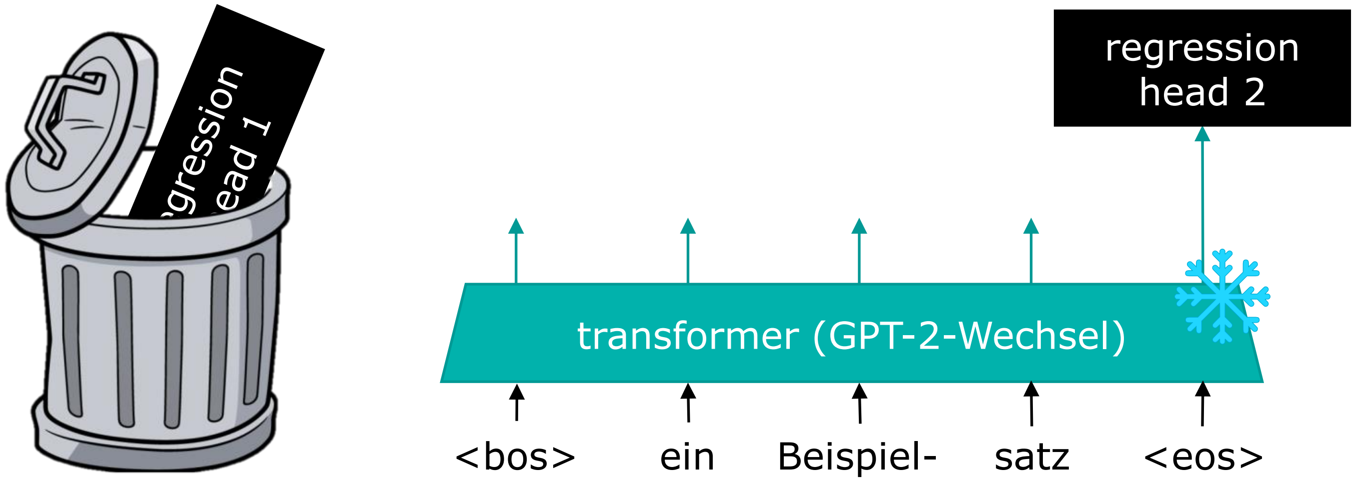


Training — Readability Features



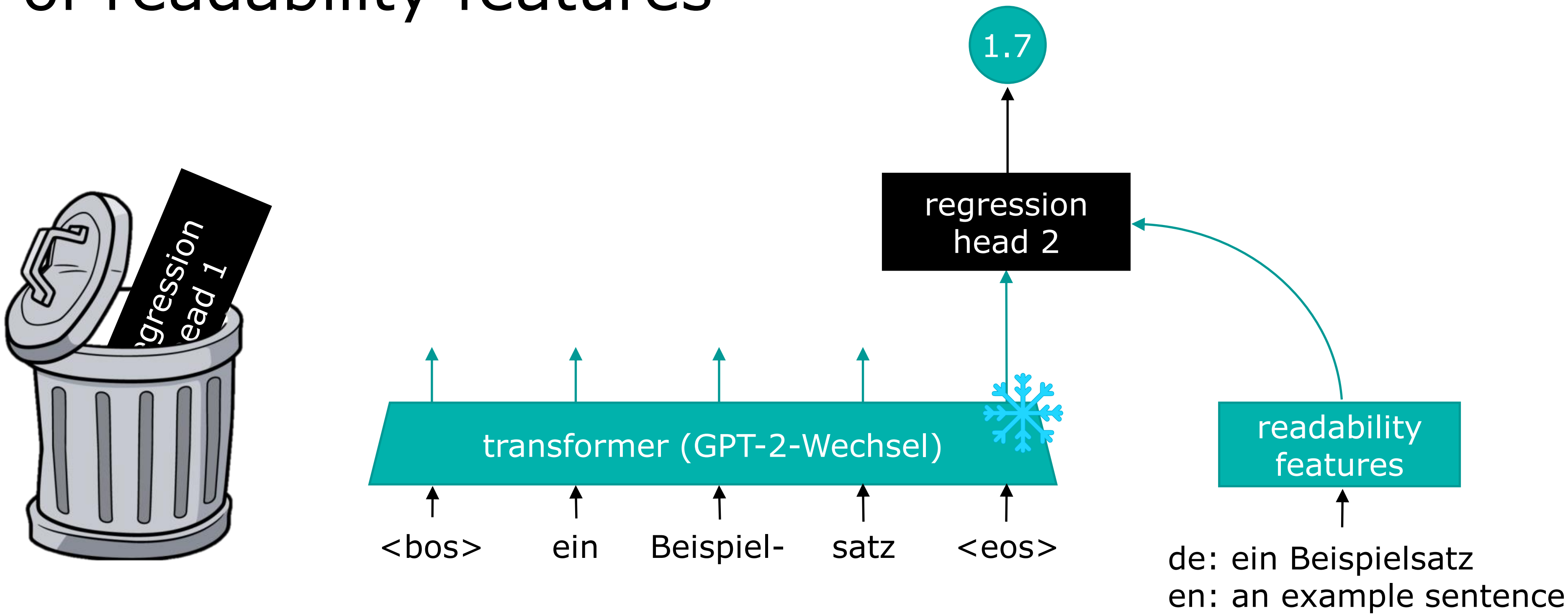
Training — Readability Features

- transformer is frozen
- second regression head: non-linear MLP



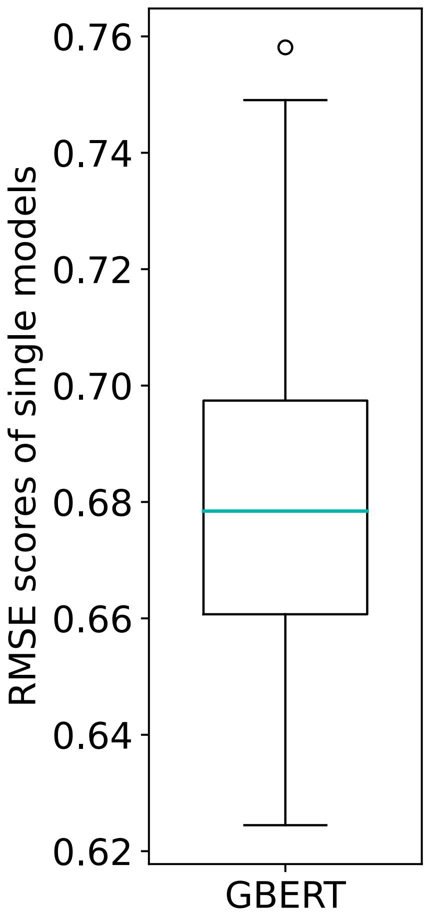
Training — Readability Features

- transformer is frozen
- second regression head: non-linear MLP
- incorporation of readability features



Training — Ensembling

- goal: reduce overfitting and high variance of models
- ensembles average predicted MOS scores
- ensemble members differed in weight initialization and training data



Risch & Krestel, [Bagging BERT Models for Robust Aggression Identification]() (2020)

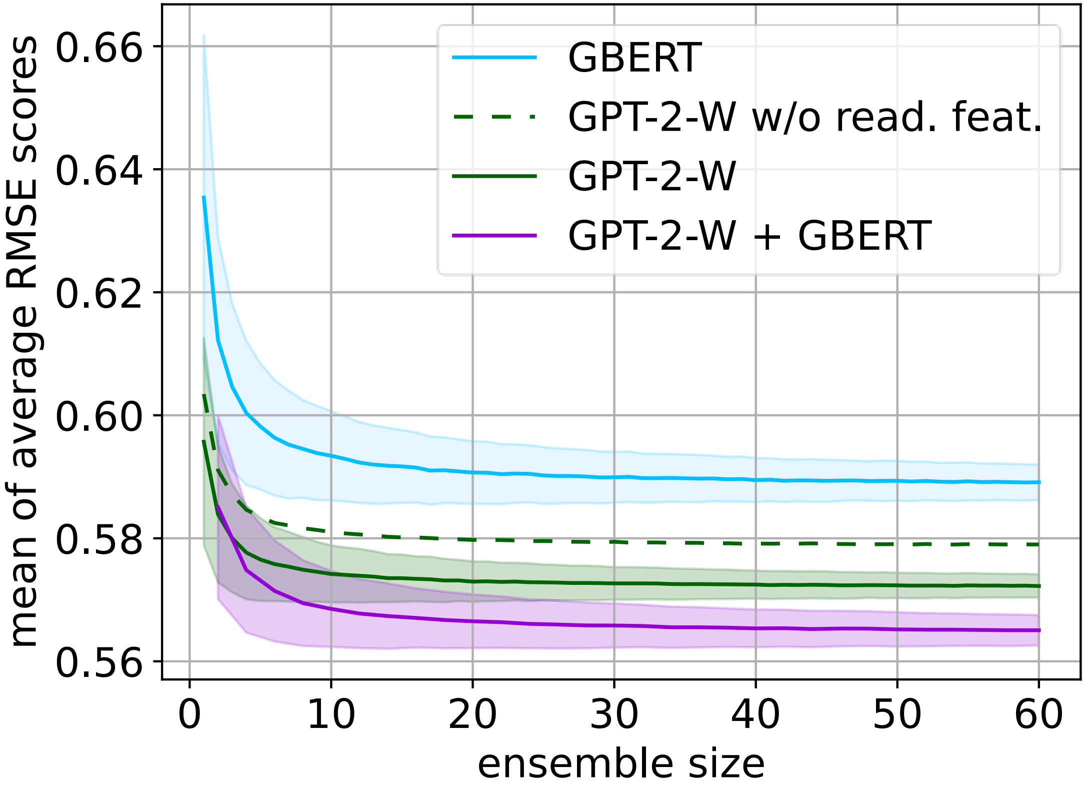
Answering Our Questions

Q1: size of the LLM (How does the size of an LLM effect readability assessment?)

Q2: readability features (Can traditional readability features be incorporated to boost performance?)

Q3: ensemble composition (What happens if different model types are combined?)

Q4: ensemble size (How many models should an ensemble contain?)



Submissions

two submitted ensembles

	RMSE	mapped RMSE
340 GPT-2-Wechsel	0.461	0.454
100 GPT-2-Wechsel + 100 GBERT	**0.442**	**0.435**

➔ ensemble composition is important

➔ big LLMs seem to benefit from traditional readability features

Any questions?

You can find our code at
https://github.com/dslaborg/tcc2022

Readability Features

- we used two public libraries[1,2]
- features were based on
  - readability grades (various metrics and indices)
  - sentence info (number of characters/words, number of long/complex words, …)
  - POS tags (lexical density, word rarity)
  - word usage (verbs, prepositions, …)
  - …
- not all features were appropriate (e.g., some may need longer input texts to be useful)

[1] https://github.com/andreasvc/readability
[2] https://github.com/tsproisl/textcomplexity

Postprocessing

- during inference on the test data we found that some of our models predicted MOS scores <1.0
- MOS scores were created using a scale from 1.0 to 7.0

→ we discarded all predictions that were smaller than 1.0

Hypothesis: distribution shift between training and test data (also reported by other participants)

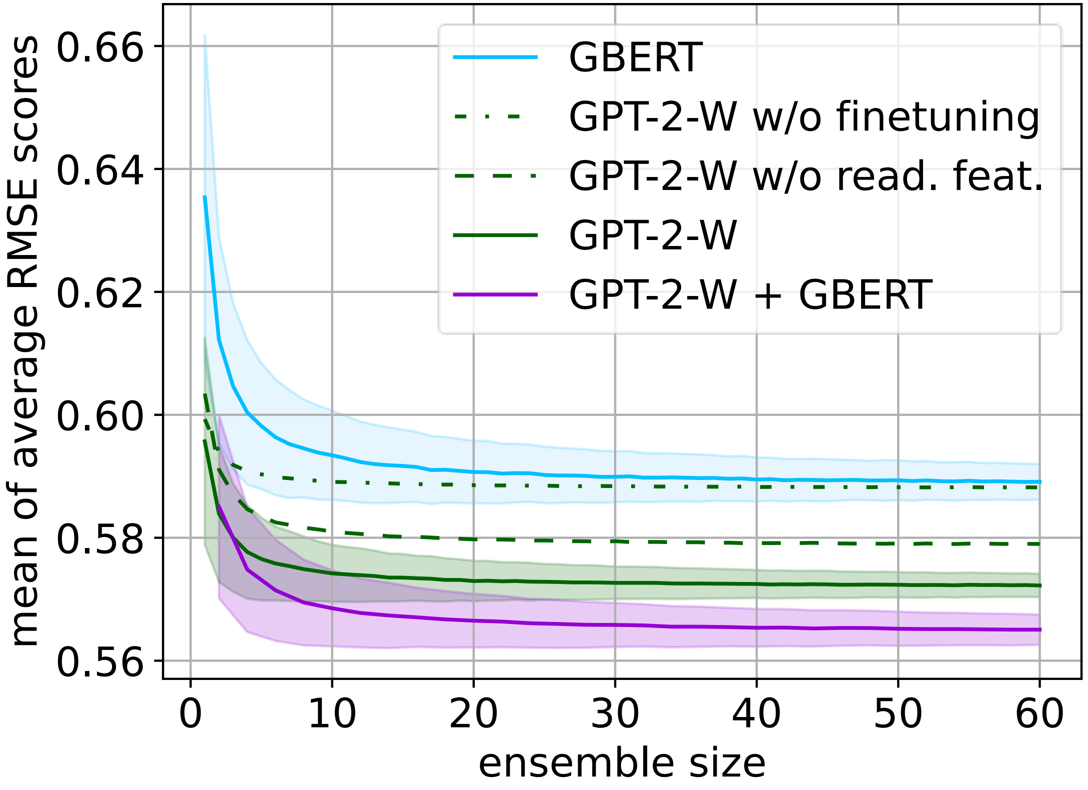
Multilingual Models

- we explored two large multilingual LLMs
- models were not finetuned
- Luminous by Aleph Alpha[1] was trained on 5 languages and has an estimated 40-80 billion parameters (proprietary model)
- XLM-RoBERTa$_{XXL}$[2] was trained on 2.5 TB of CommonCrawl data containing 100 languages, the model has 10.7 billion parameters

[1]https://www.aleph-alpha.com/technology
[2]https://huggingface.co/facebook/xlm-roberta-xxl

Bootstrapping

1. for every setup, a pool of 100 models was trained
2. for each ensemble size, a subset of models was randomly picked from the pool and combined into an ensemble
3. step 2 was repeated 1000 times



GPT-2 / WECHSEL