ReadMe++: Benchmarking Multilingual LMs for Multi-domain Readability Assessment

Tarek Naous

Michael J. Ryan

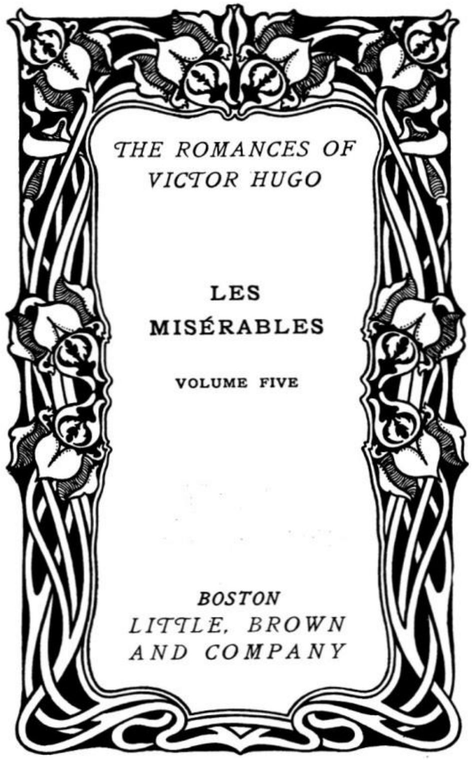
Anton Lavrouk

Mohit Chandra

Wei Xu

EMNLP 2024

Georgia Tech



THE ROMANCES OF
VICTOR HUGO

LES
MISÉRABLES

VOLUME FIVE

BOSTON
LITTLE, BROWN
AND COMPANY

In the uncoerced slowness of its gait, suppleness and agility were discernible.

In its voluntary slow movement, its flexibility and agility were noticeable.

In its voluntary slow movement, you could still see how flexible and quick it is.

Enabling content accessibility *for various audiences requires* ***reliable readability predictors****!*

multi-domain *multilingual*

Human-annotated Resources: (Arase et al. 2022, Brunato et al. 2018, and more)

A man driving a red and black go-kart with number " 11 " on it

***CEFR Scale**: Common European Framework of Reference for Languages*

Level	Description	Rating
A1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words	**1**
A2	Can understand short, simple texts on familiar matters of a concrete type	**2**
B1	Can read straightforward factual texts on subjects related to his/her field and interest	**3**
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts	**4**
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her area of specialty	**5**
C2	Can understand and interpret critically virtually all forms of the written language	**6**

The metaphor of the dream navel, then, creates and supports a certain structure of meaning and inquiry

Past resources are mostly restricted to a few domains (Wikipedia, News, Books) and English

Human-annotated Resources: (Arase et al. 2022, Brunato et al. 2018, and more)

A man driving a red and black go-kart with number " 11 " on it

***CEFR Scale**: Common European Framework of Reference for Languages*

Level	Description	Rating
A1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words	1
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts	4
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her area of specialty	5
C2	Can understand and interpret critically virtually all forms of the written language	6

Need a diverse resource for **domain** and **language** generalization of readability methods

The metaphor of the dream navel, then, creates and supports a certain structure of meaning and inquiry

Past resources are mostly restricted to a few domains (Wikipedia, News, Books) and English

We Introduce ReadMe++

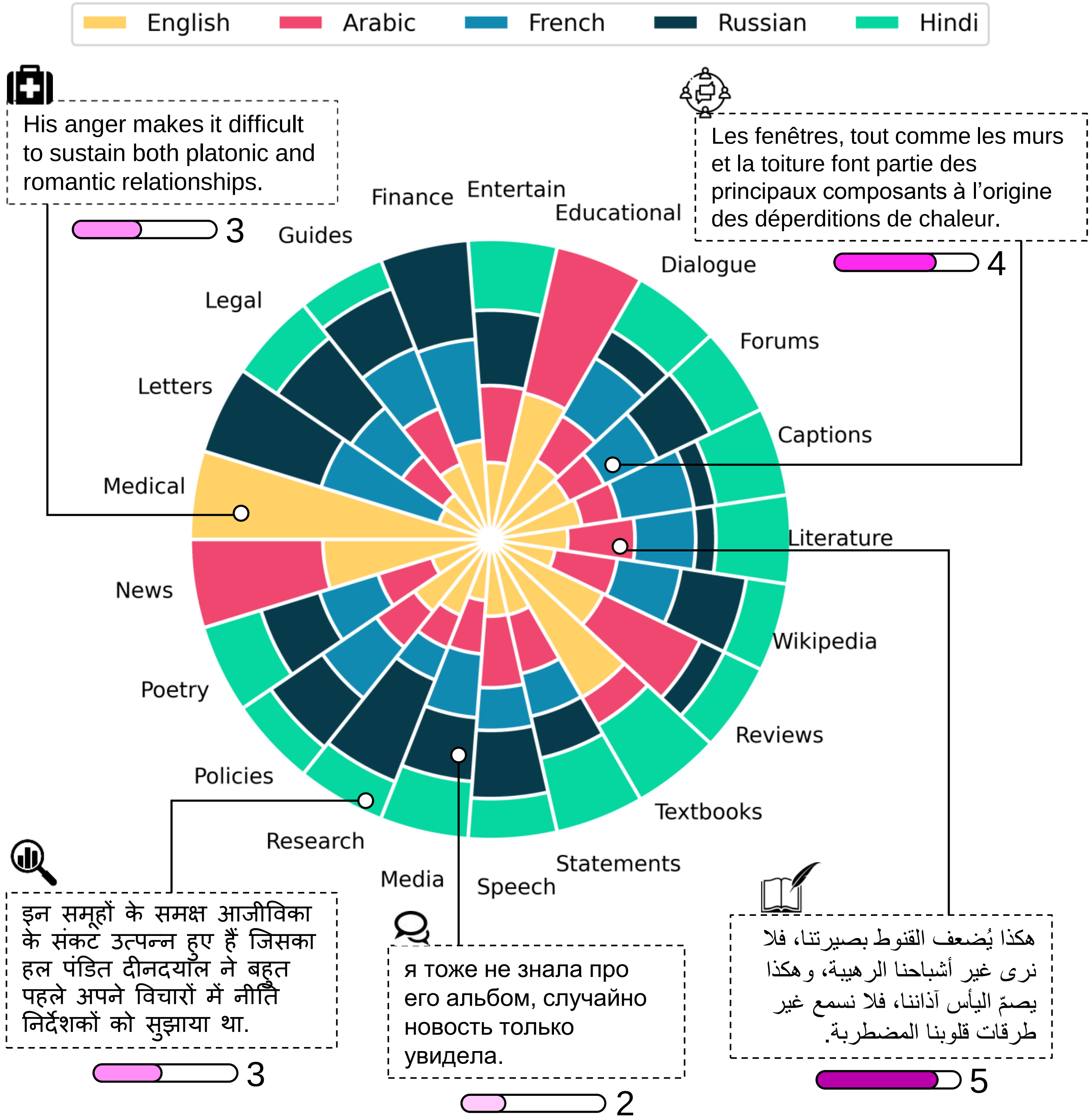
Massively diverse benchmark for readability

More language diversity

- 5 different languages
- 4 different writing scripts
- 9,757 human-annotated sentences

More domain diversity

- 21 top-level domains
- 112 data sources



His anger makes it difficult to sustain both platonic and romantic relationships.

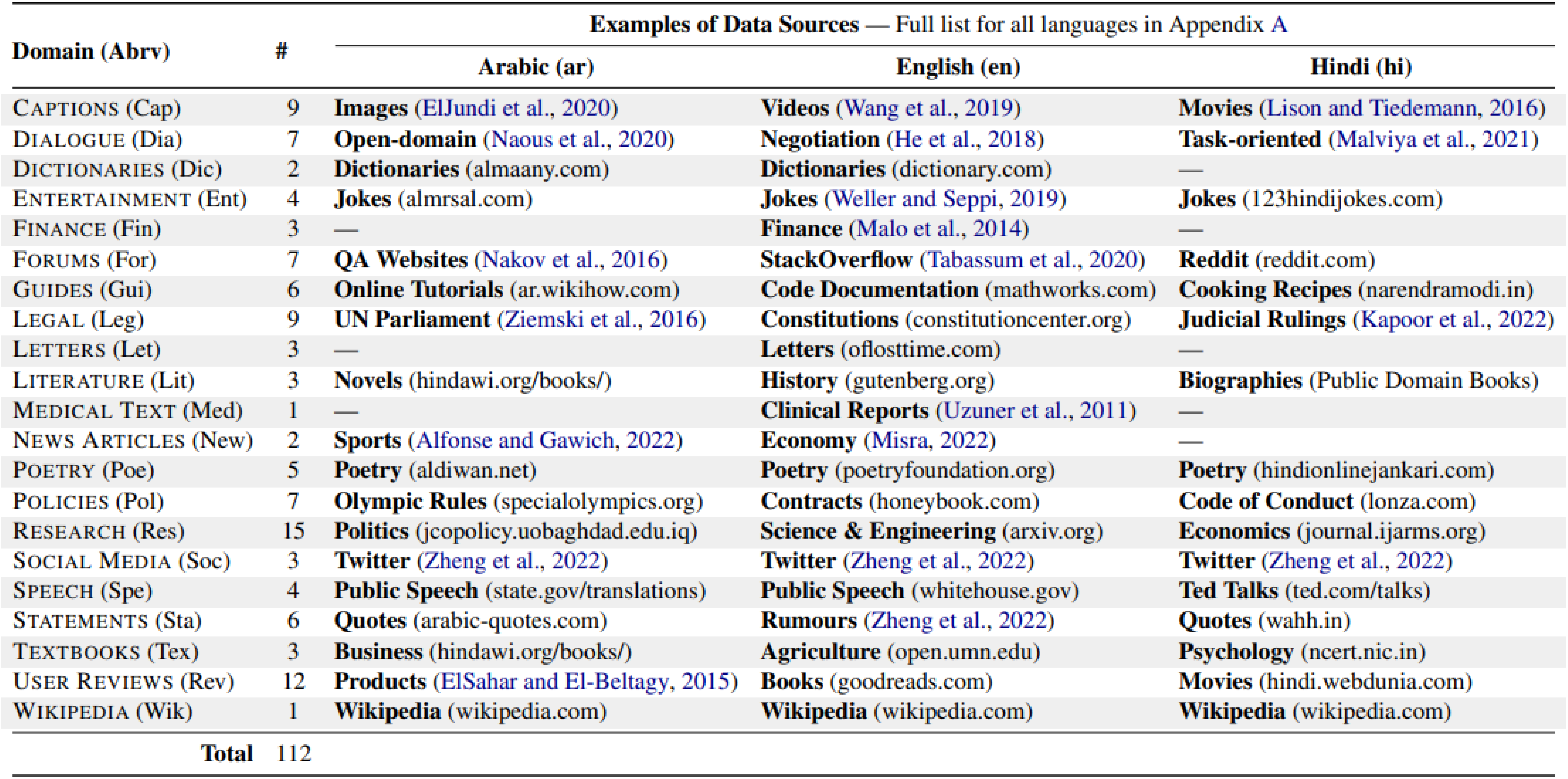
Les fenêtres, tout comme les murs et la toiture font partie des principaux composants à l'origine des déperditions de chaleur.

इन समूहों के समक्ष आजीविका के संकट उत्पन्न हुए हैं जिसका हल पंडित दीनदयाल ने बहुत पहले अपने विचारों में नीति निर्देशकों को सुझाया था.

я тоже не знала про его альбом, случайно новость только увидела.

هكذا يُضعف القنوط بصيرتنا، فلا
نرى غير أشباحنا الرهيبة، وهكذا
يصمّ اليأس آذاننا، فلا نسمع غير
طرقات قلوبنا المضطربة.

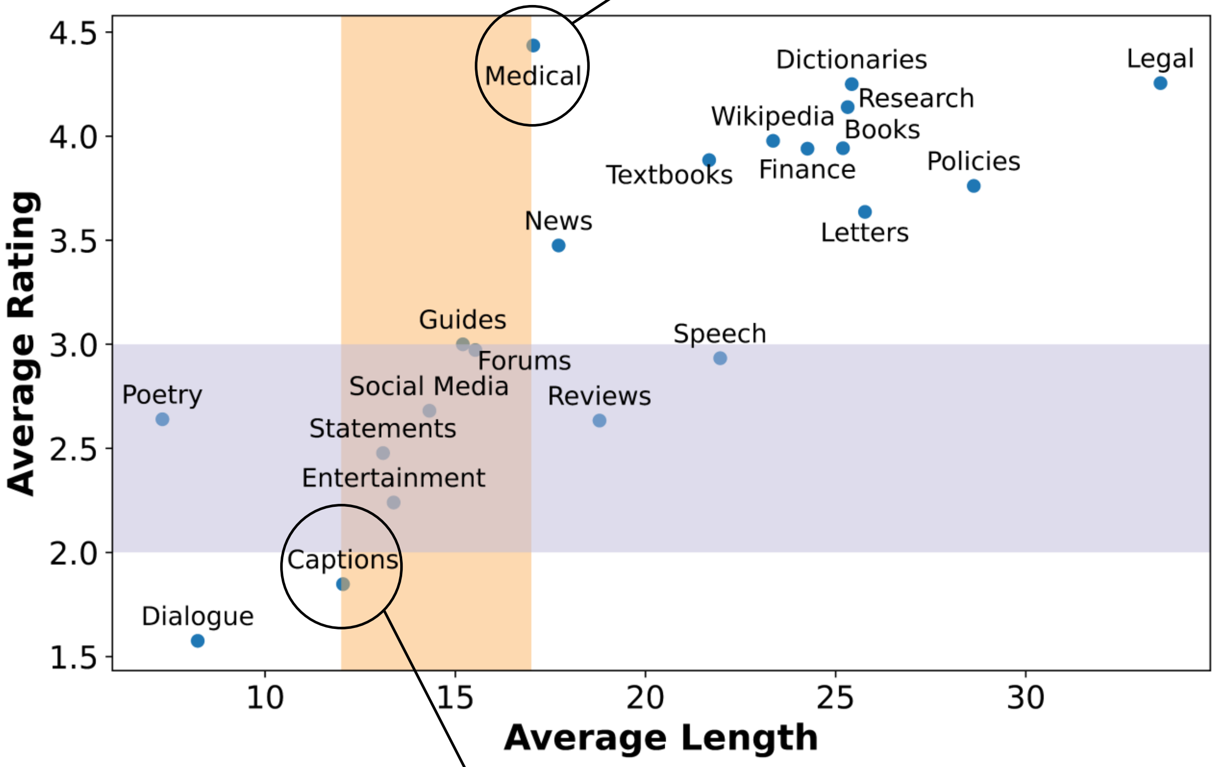
ReadMe++: Domains & Sources



Domain (Abrv)	#	Examples of Data Sources — Full list for all languages in Appendix A		
		Arabic (ar)	English (en)	Hindi (hi)
Captions (Cap)	9	**Images** (ElJundi et al., 2020)	**Videos** (Wang et al., 2019)	**Movies** (Lison and Tiedemann, 2016)
Dialogue (Dia)	7	**Open-domain** (Naous et al., 2020)	**Negotiation** (He et al., 2018)	**Task-oriented** (Malviya et al., 2021)
Dictionaries (Dic)	2	**Dictionaries** (almaany.com)	**Dictionaries** (dictionary.com)	—
Entertainment (Ent)	4	**Jokes** (almrsal.com)	**Jokes** (Weller and Seppi, 2019)	**Jokes** (123hindijokes.com)
Finance (Fin)	3	—	**Finance** (Malo et al., 2014)	—
Forums (For)	7	**QA Websites** (Nakov et al., 2016)	**StackOverflow** (Tabassum et al., 2020)	**Reddit** (reddit.com)
Guides (Gui)	6	**Online Tutorials** (ar.wikihow.com)	**Code Documentation** (mathworks.com)	**Cooking Recipes** (narendramodi.in)
Legal (Leg)	9	**UN Parliament** (Ziemski et al., 2016)	**Constitutions** (constitutioncenter.org)	**Judicial Rulings** (Kapoor et al., 2022)
Letters (Let)	3	—	**Letters** (oflosttime.com)	—
Literature (Lit)	3	**Novels** (hindawi.org/books/)	**History** (gutenberg.org)	**Biographies** (Public Domain Books)
Medical Text (Med)	1	—	**Clinical Reports** (Uzuner et al., 2011)	—
News Articles (New)	2	**Sports** (Alfonse and Gawich, 2022)	**Economy** (Misra, 2022)	—
Poetry (Poe)	5	**Poetry** (aldiwan.net)	**Poetry** (poetryfoundation.org)	**Poetry** (hindionlinejankari.com)
Policies (Pol)	7	**Olympic Rules** (specialolympics.org)	**Contracts** (honeybook.com)	**Code of Conduct** (lonza.com)
Research (Res)	15	**Politics** (jcopolicy.uobaghdad.edu.iq)	**Science & Engineering** (arxiv.org)	**Economics** (journal.ijarms.org)
Social Media (Soc)	3	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)	**Twitter** (Zheng et al., 2022)
Speech (Spe)	4	**Public Speech** (state.gov/translations)	**Public Speech** (whitehouse.gov)	**Ted Talks** (ted.com/talks)
Statements (Sta)	6	**Quotes** (arabic-quotes.com)	**Rumours** (Zheng et al., 2022)	**Quotes** (wahh.in)
Textbooks (Tex)	3	**Business** (hindawi.org/books/)	**Agriculture** (open.umn.edu)	**Psychology** (ncert.nic.in)
User Reviews (Rev)	12	**Products** (ElSahar and El-Beltagy, 2015)	**Books** (goodreads.com)	**Movies** (hindi.webdunia.com)
Wikipedia (Wik)	1	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)	**Wikipedia** (wikipedia.com)
Total	112			

ReadMe++: Sentence Diversity

With history, will go for cardiac catheterization evaluation.



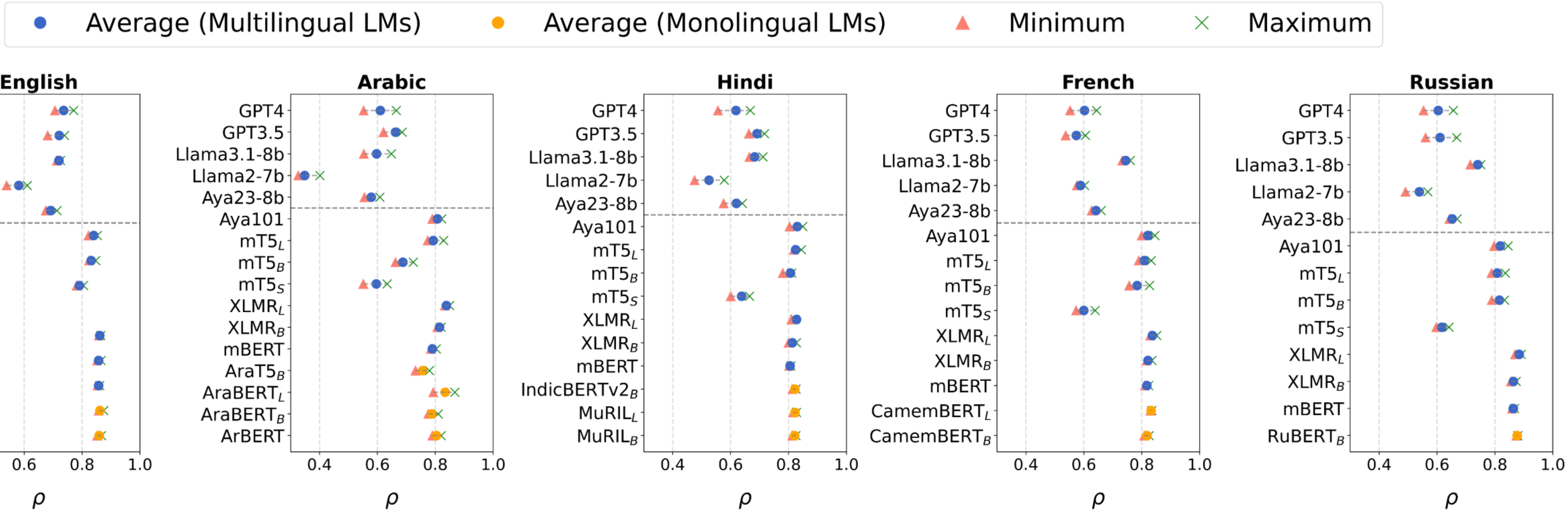
A young boy is indoors showing his family his dance moves.

1

Supervised & Prompting Methods

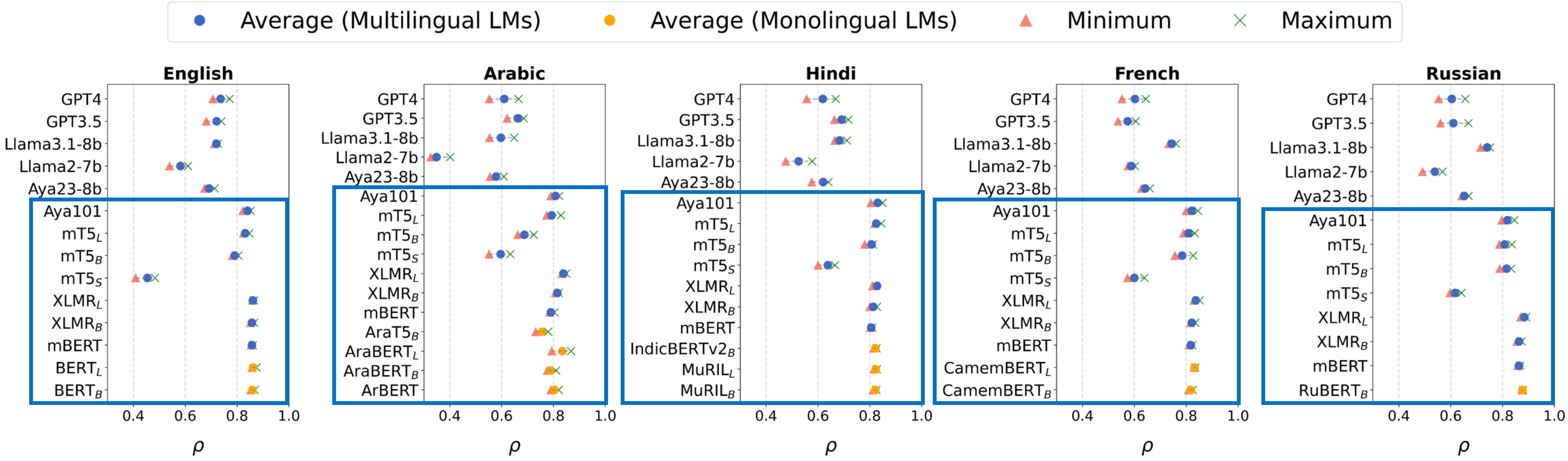
How good are fine-tuned and prompted LLMs are at predicting sentence readability?

Benchmarking Results



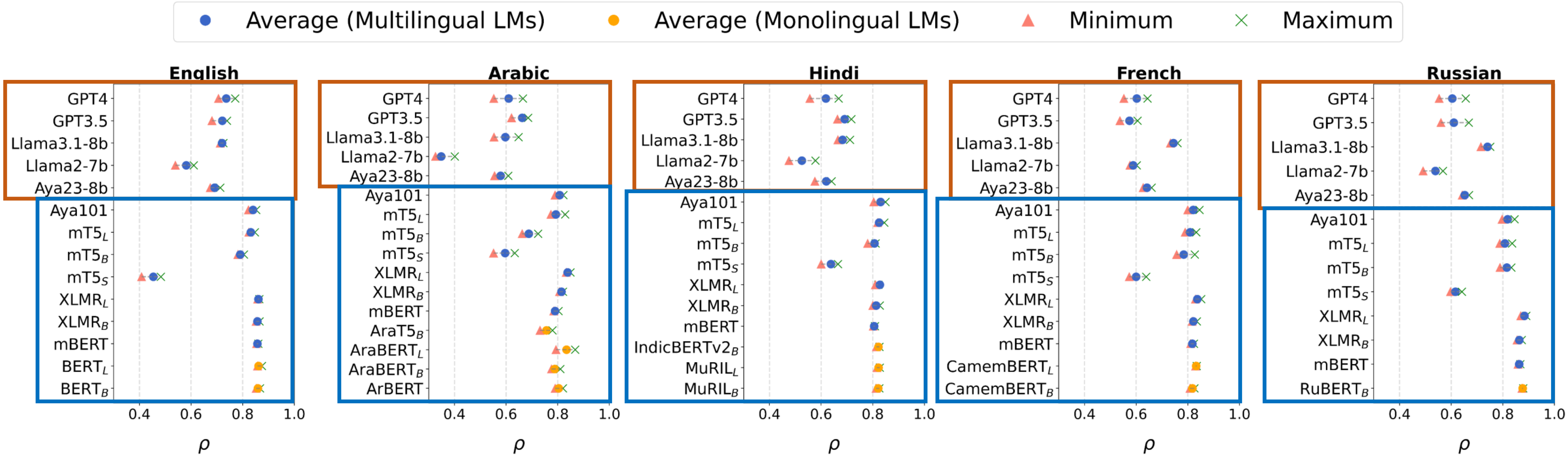
Data Split: random splitting per domain, ensuring all domains are covered in each train/val/test split

Benchmarking Results



Fine-tuned LMs (on all domains) achieve high correlations with human scores (0.8-0.9)

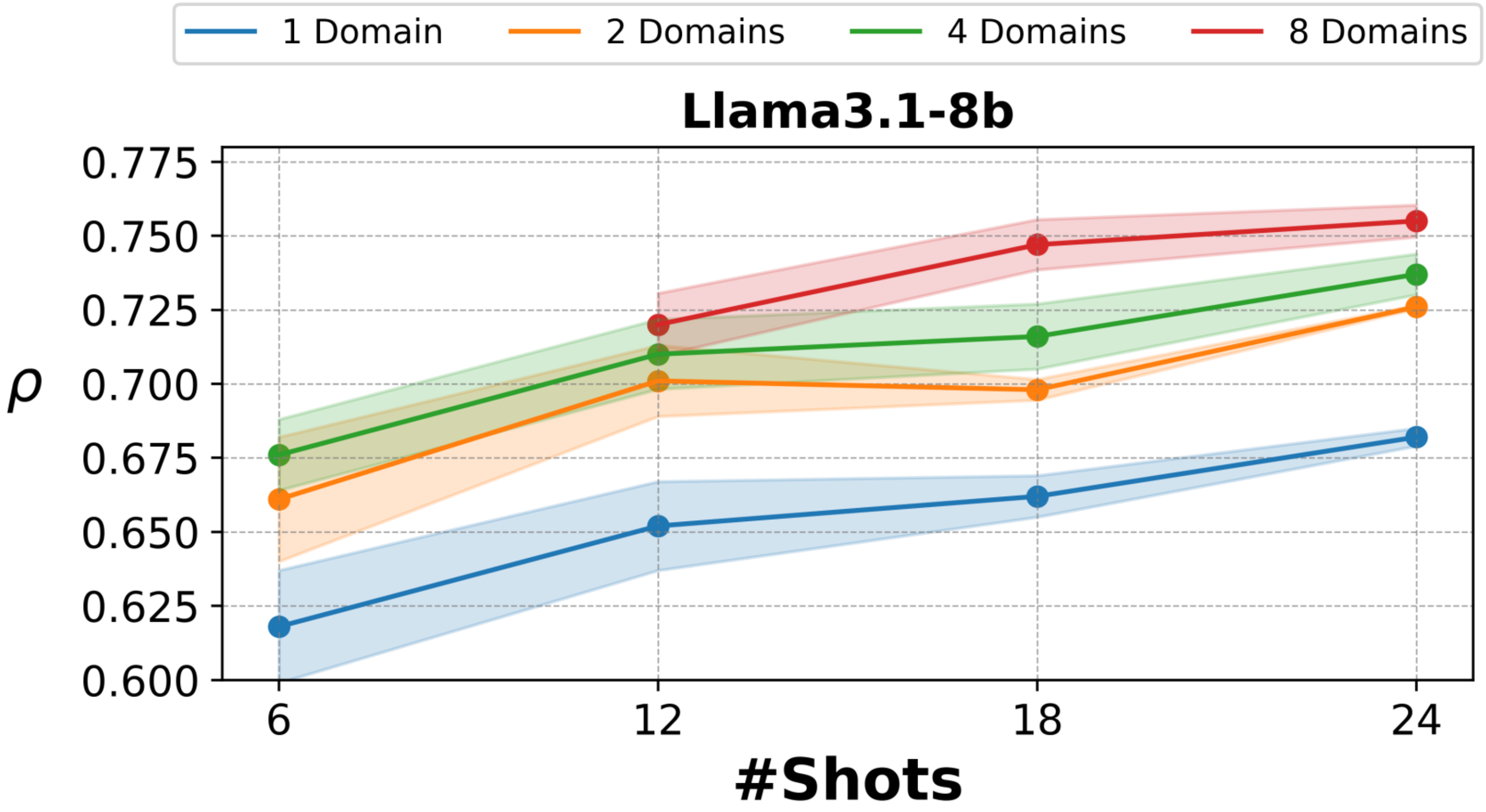
Benchmarking Results



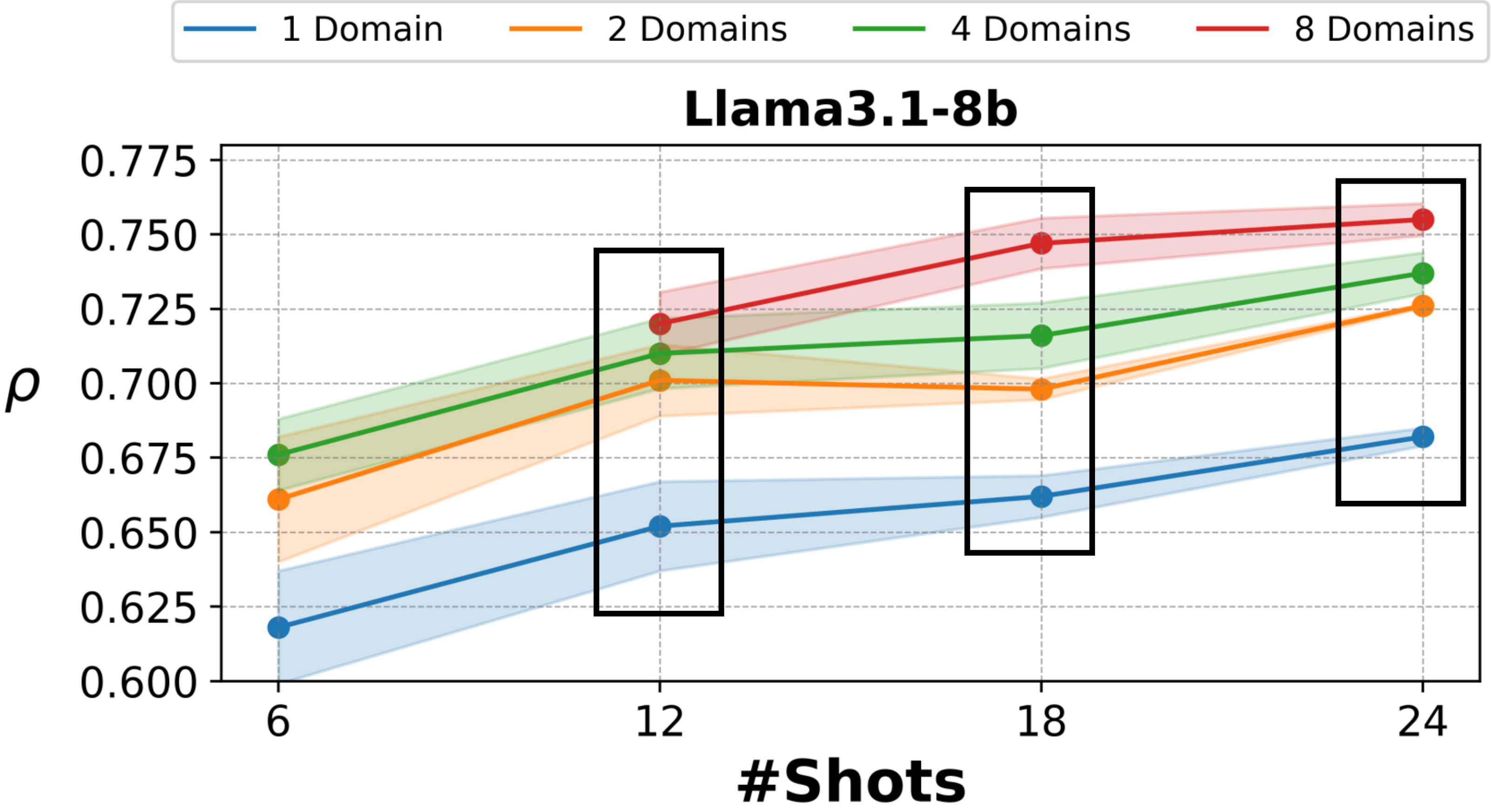
Fine-tuned LMs (on all domains) achieve high correlations with human scores (0.8-0.9)

Prompted LMs (5-shot random demonstrations) fall behind fine-tuned LMs

The Importance of Domain Diversity in Demonstrations



The Importance of Domain Diversity in Demonstrations

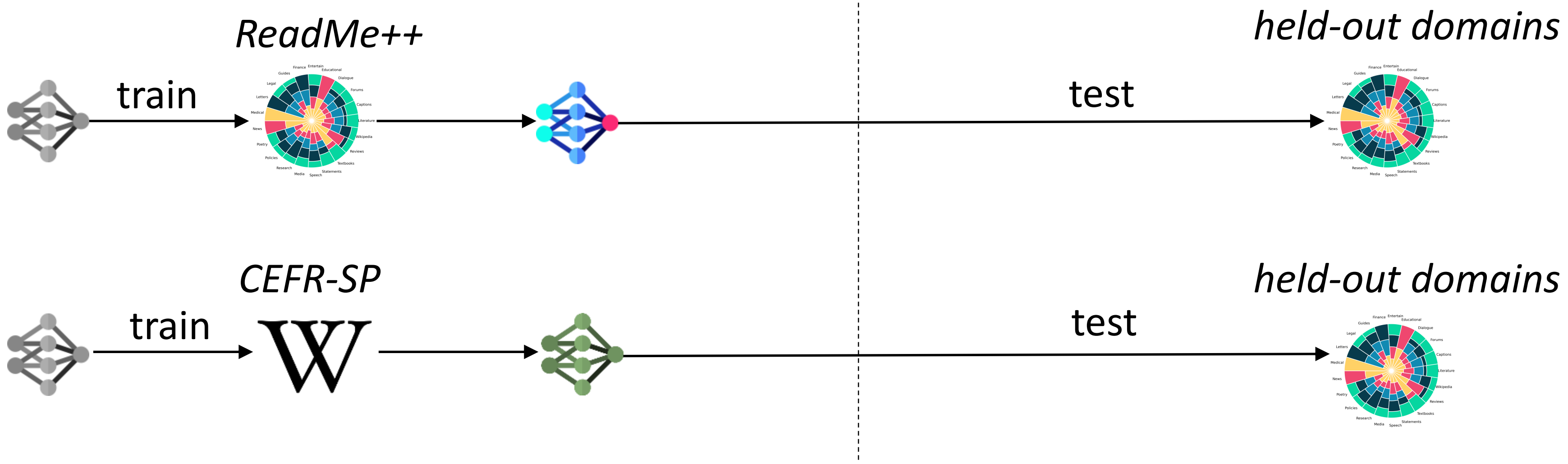


Prompting performance improves greatly as demonstrations are sampled from more domains

Domain Generalization

Fine-tuning

Evaluation on Unseen Domains



			Training Source	
# Unseen Domains	*#train/val*	*#test*	*ReadMe++*	*CEFR-SP*
2: Wikipedia, Research	1995/235	631	**0.611**	0.439
4: Letters, Social Media, Entertainment, Guides	2285/267	309	**0.761**	0.649
6: Research, Finance, Statements, Entertainment, Dialogue, News	1885/221	755	**0.780**	0.517
8: Policies, Captions, Statements, Research, Reviews, Legal, Social, Poetry	1653/191	1017	**0.828**	0.690

LMs trained on ReadMe++ perform better on unseen domains compared to single-domain datasets

1 *Supervised & Prompting Methods*

How good are fine-tuned and prompted LLMs are at predicting sentence readability?

2 *Unsupervised Methods*

How do traditional metrics and unsupervised LM-based methods compare?

Traditional Metrics

Sentence Length, FKGL *(Flesch-Kincaid Grade Level)*, ARI *(Automated Readability Index)*

LM-based Metric *(Martinc et al. 2021)*

Order Rank

bigger weight for difficult words

***W**ord **N**egative **L**og **L**oss*

lower probability → higher loss

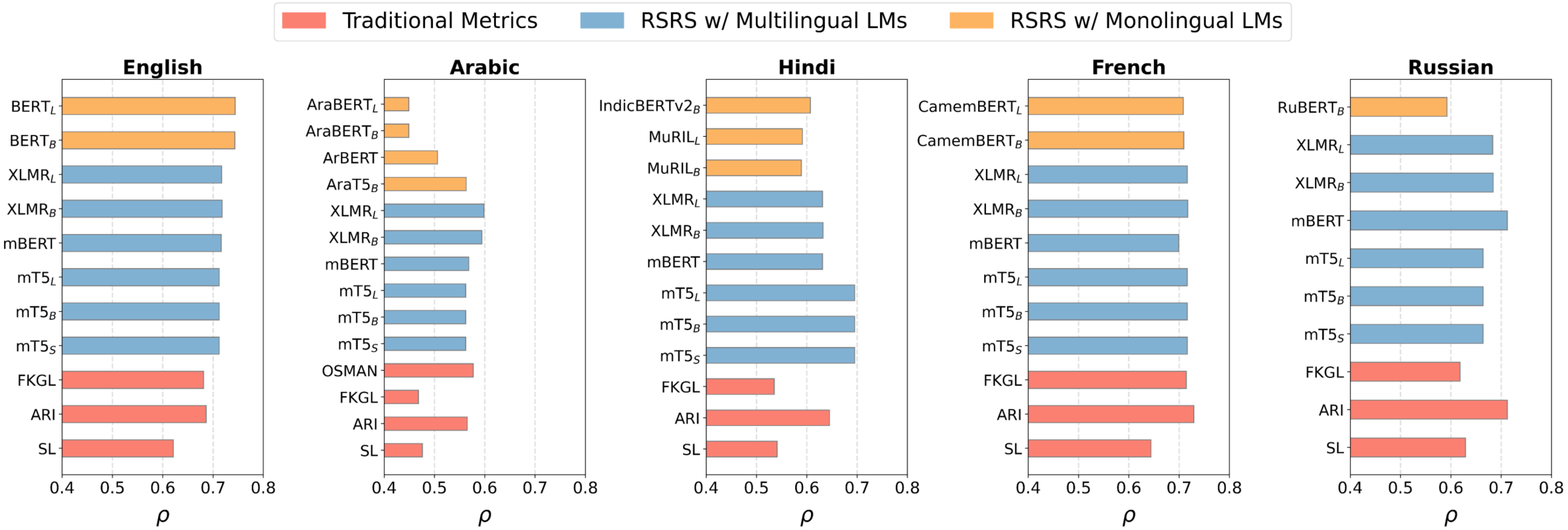
$$\mathrm{RSRS} = \frac{\sum_{i=1}^{|S|} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Number of tokens

$$\underbrace{\text{In}}_{1.1}\ \underbrace{\text{the}}_{0.3}\ \underbrace{\text{voluntary}}_{3.6}\ \underbrace{\text{slowness}}_{4.5}\ \underbrace{\text{of}}_{1}\ \underbrace{\text{its}}_{0.2}\ \underbrace{\text{gait}}_{11.2}$$

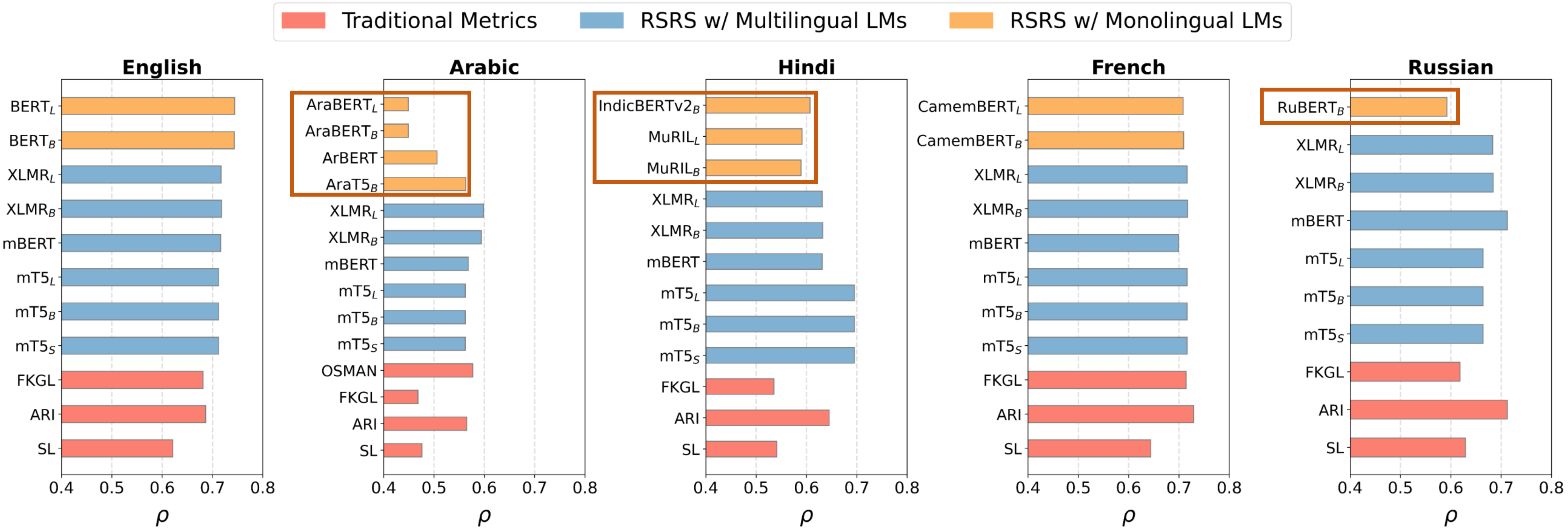
$$\underbrace{\text{In}}_{1.1}\ \underbrace{\text{the}}_{0.3}\ \underbrace{\text{voluntary}}_{3.6}\ \underbrace{\text{slowness}}_{4.5}\ \underbrace{\text{of}}_{1}\ \underbrace{\text{its}}_{0.2}\ \underbrace{\text{movement}}_{4.5}$$

Benchmarking Results



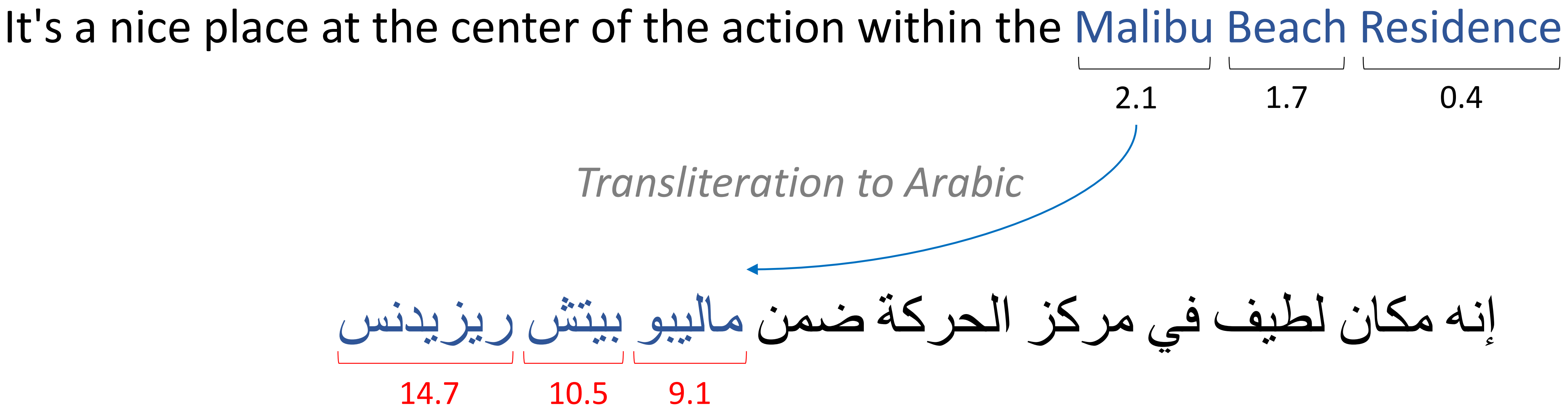
RSRS is competitive with traditional feature-based metrics, outperforms them in some cases

Benchmarking Results



RSRS w/ monolingual LMs performs poorer compared with multilingual LMs in non-latin scripts

The Impact of Transliterations on RSRS in Non-Latin Script Languages



It's a nice place at the center of the action within the Malibu Beach Residence

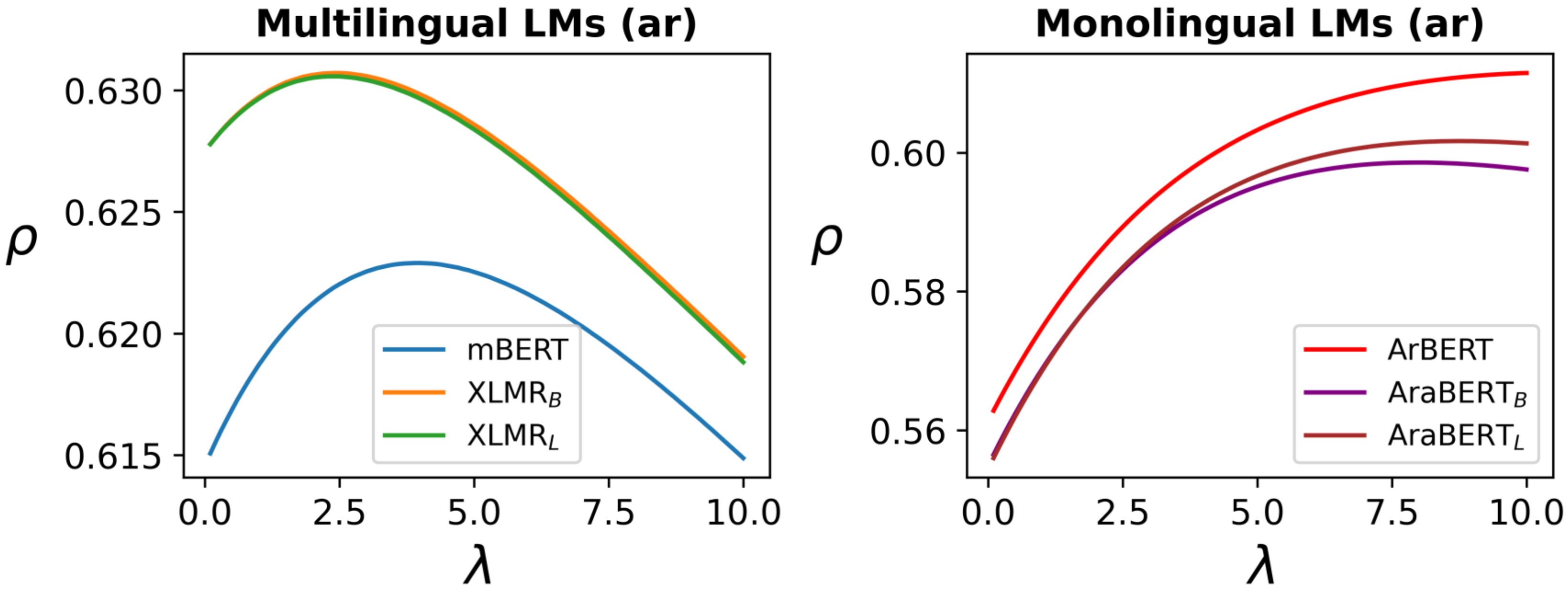
إنه مكان لطيف في مركز الحركة ضمن ماليبو بيتش ريزيدنس

Transliteration in non-latin script treated as rare words in RSRS = high word losses

***Not all types of rare words increase difficulty**, transliterations can inflate RSRS scores*

The Impact of Transliterations on RSRS in Non-Latin Script Languages

Penalize RSRS scores by λ for sentences containing transliterations and check correlation with humans



Jumps in correlation (7-8%) for monolingual LMs as RSRS scores are decreased

Takeaways

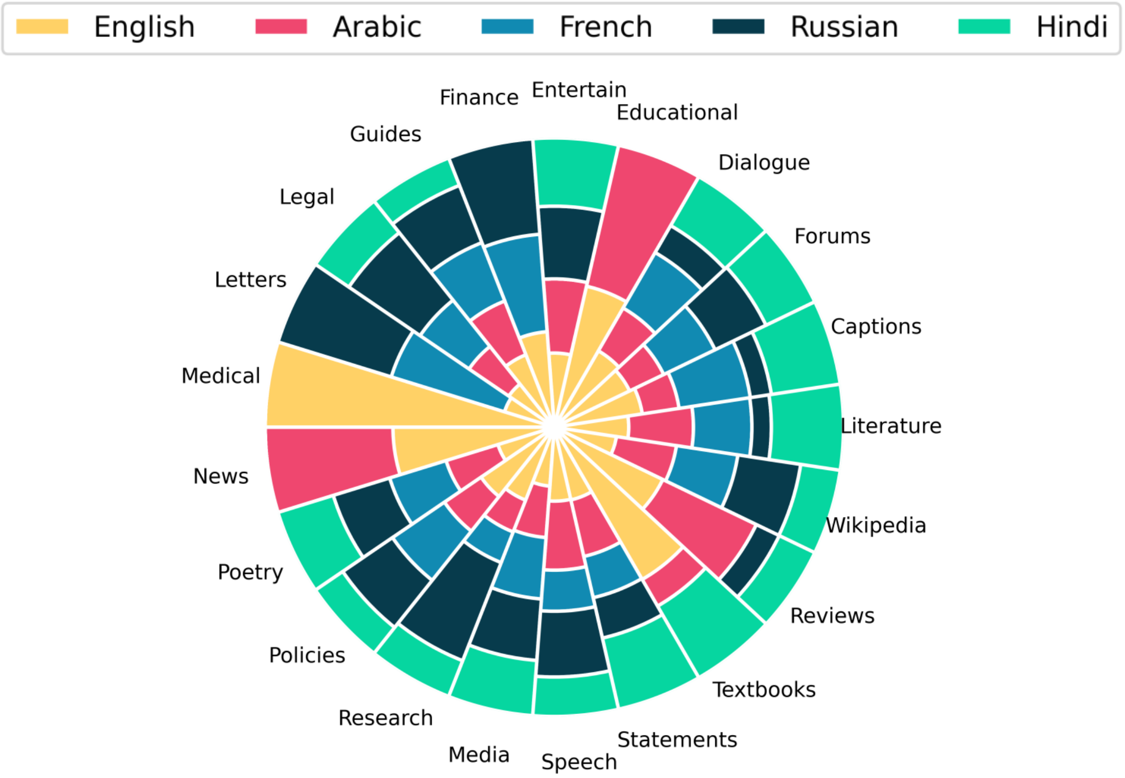
Domain diversity is important for generalizable predictors

- More efficient fine-tuning, prompting, generalization to unseen domains/languages

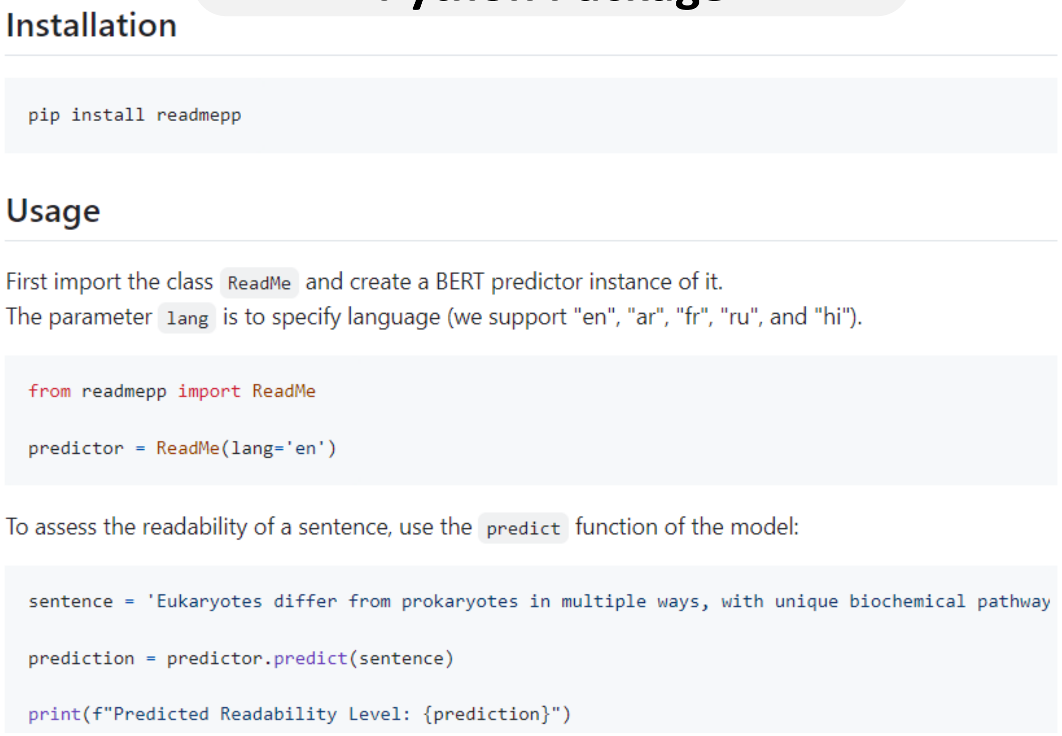
Language diversity needed to influence design of better metrics

- Languages with other writing systems hold their own challenges

ขอบคุณ Merci 谢谢 धन्यवाद Asante Teşekkürler شكرا
ありがとう Gracias متشكرم நன்றி Obrigado Thank You



Python Package



Installation

```
pip install readmepp
```

Usage

First import the class `ReadMe` and create a BERT predictor instance of it.
The parameter `lang` is to specify language (we support "en", "ar", "fr", "ru", and "hi").

```
from readmepp import ReadMe

predictor = ReadMe(lang='en')
```

To assess the readability of a sentence, use the `predict` function of the model:

```
sentence = 'Eukaryotes differ from prokaryotes in multiple ways, with unique biochemical pathway

prediction = predictor.predict(sentence)

print(f"Predicted Readability Level: {prediction}")
```

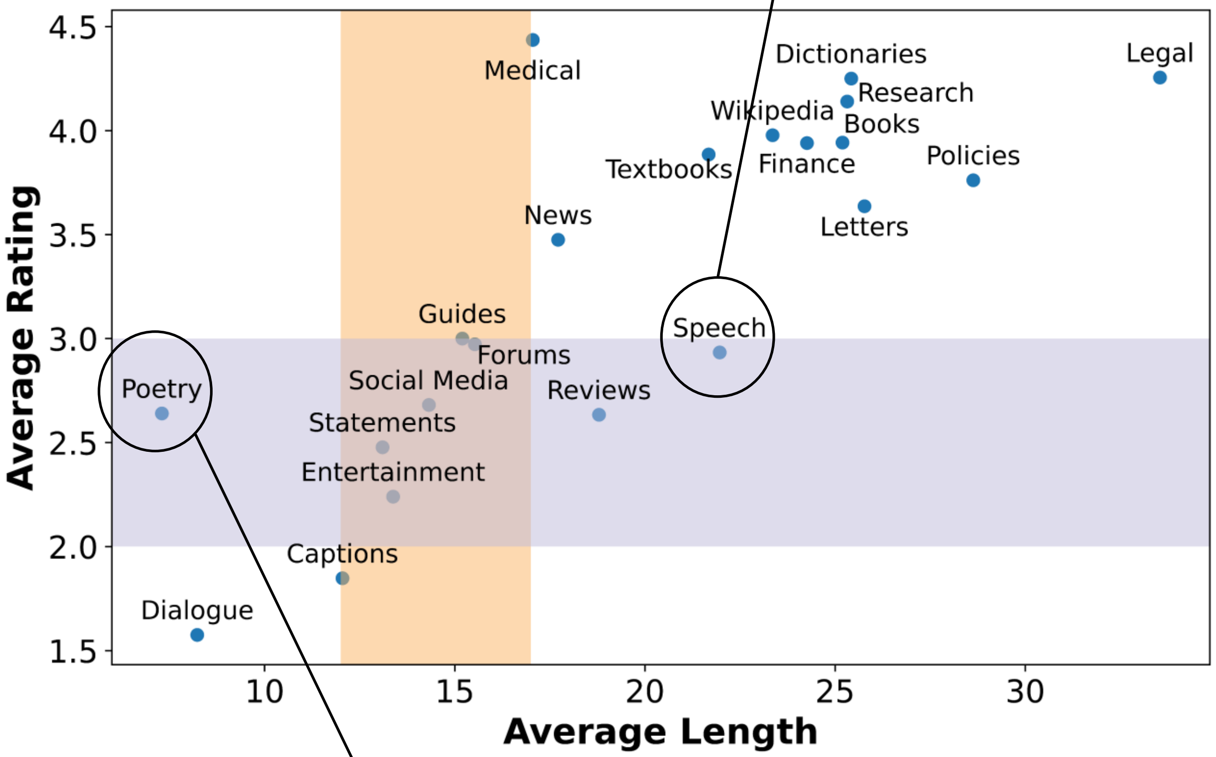
ReadMe++ is available at: https://github.com/tareknaous/readme

Feel free to follow up with me on @tareknaous

Additional Slides

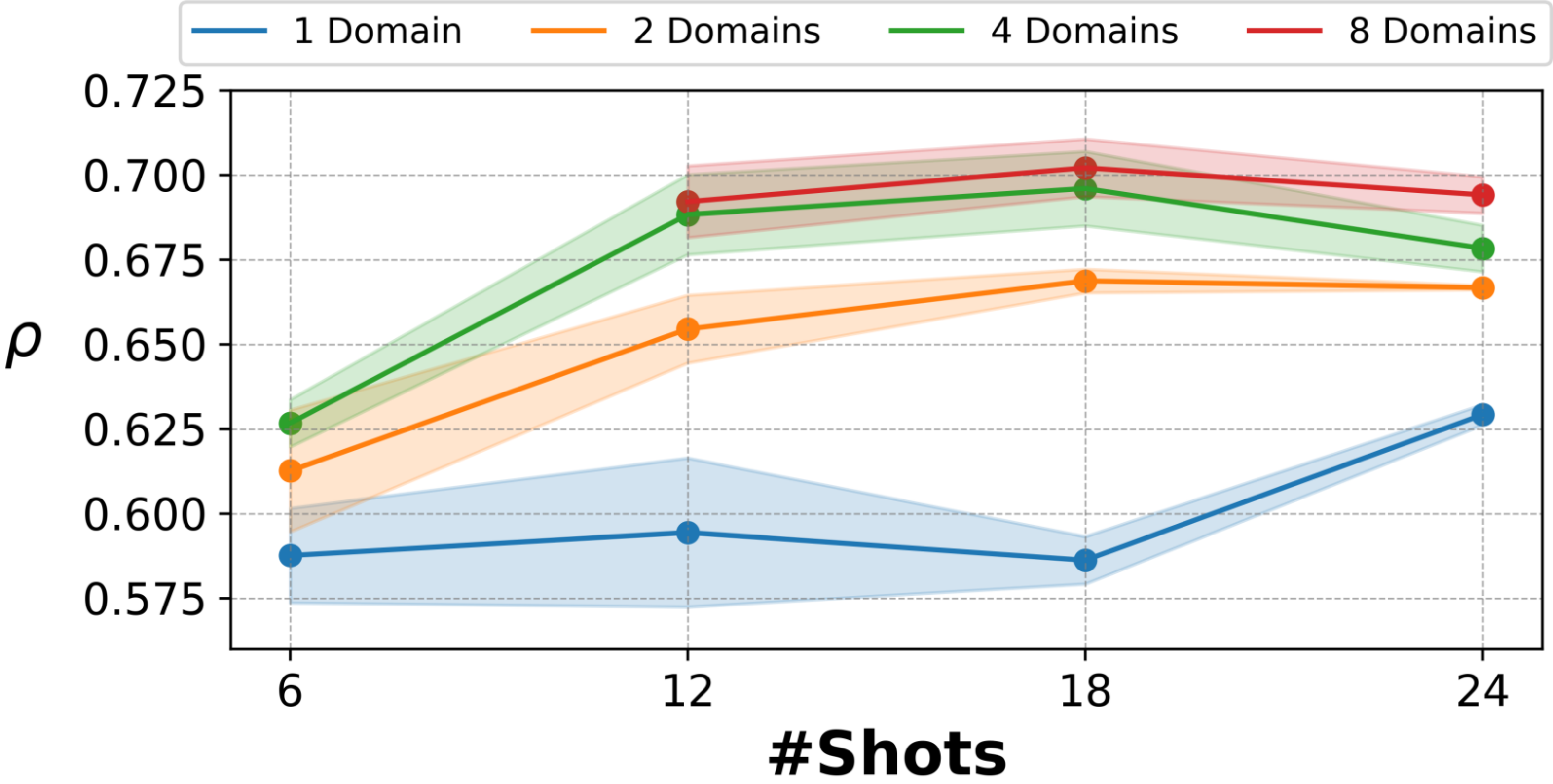
ReadMe++: Sentence Diversity

You've heard me say it ad nauseam: I don't know where it's written it says we can't be the manufacturing capital of the world

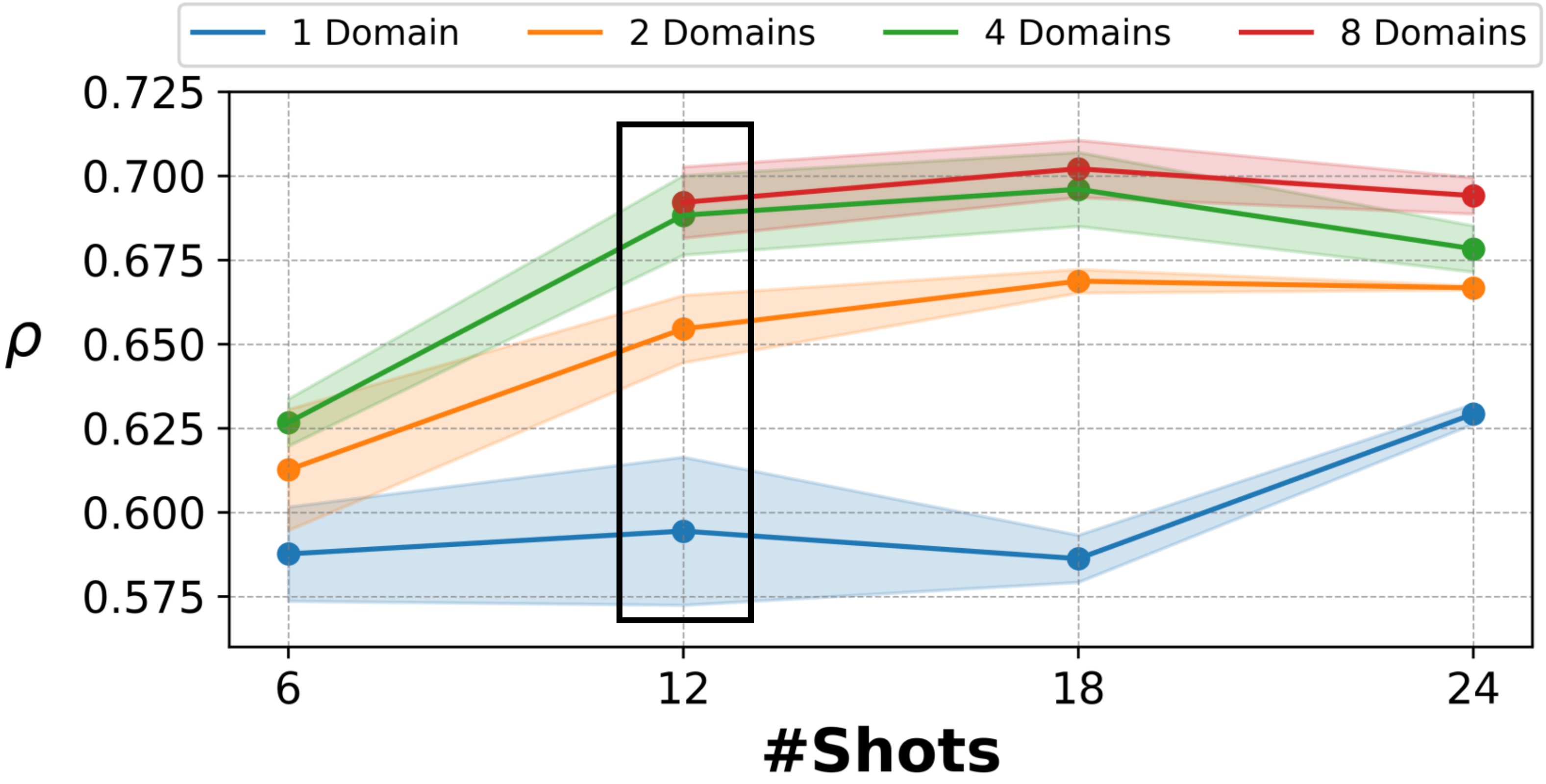


I had not thought death had undone so many

The Importance of Domain Diversity in Demonstrations

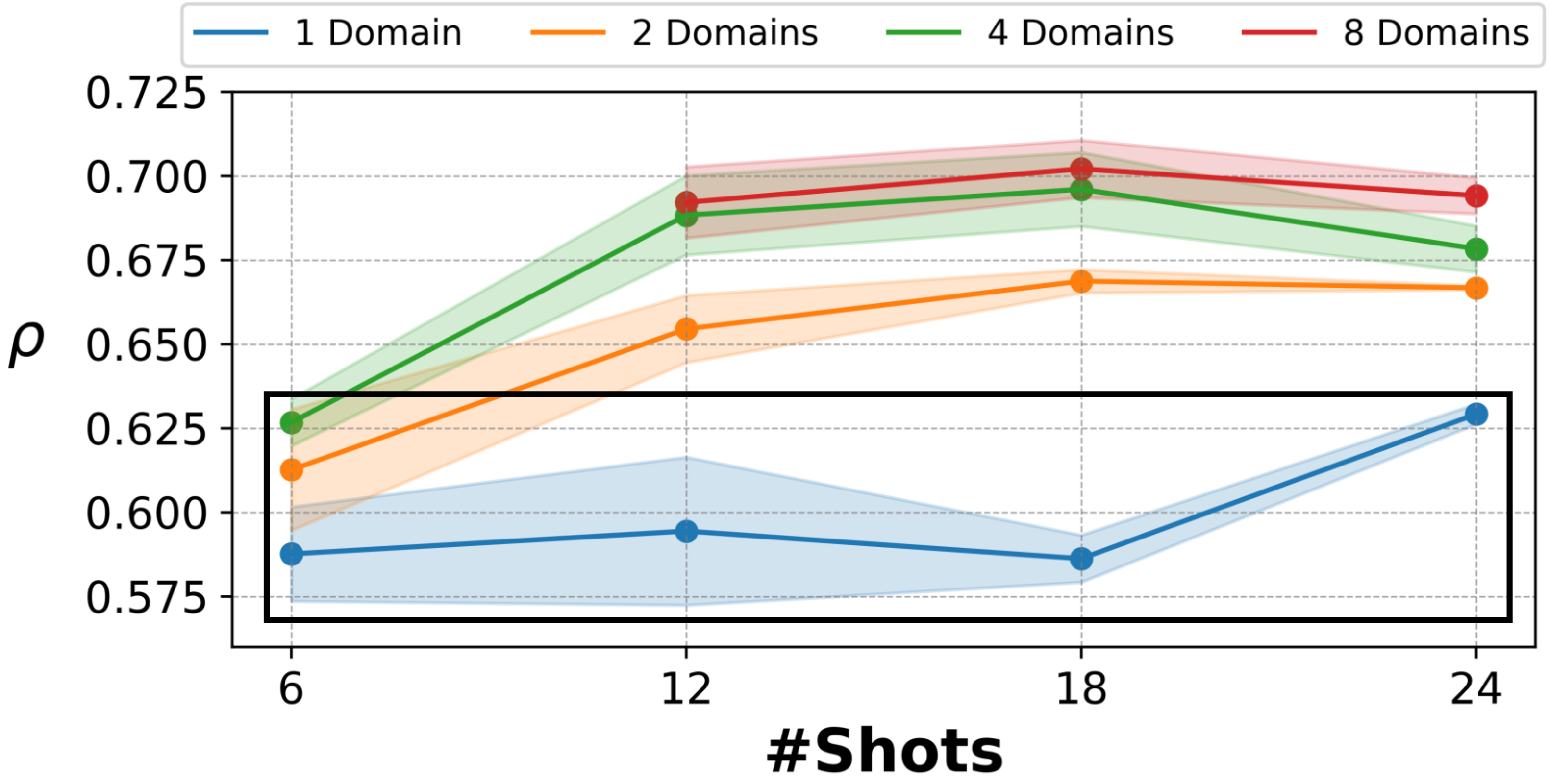


The Importance of Domain Diversity in Demonstrations



Prompting performance improves greatly as demonstrations are sampled from more domains

The Importance of Domain Diversity in Demonstrations



Performance improves with more shots, but domain diversity is more critical

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses

Martinc et al. "Supervised and unsupervised neural approaches to text readability" *Computational Linguistics,* 2021

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses

In the voluntary slowness of its gait, suppleness and agility were discernible

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses

$$\underbrace{\text{In}}_{1.1}\ \underbrace{\text{the}}_{0.3}\ \underbrace{\text{voluntary}}_{3.6}\ \underbrace{\text{slowness}}_{4.5}\ \underbrace{\text{of}}_{1}\ \underbrace{\text{its}}_{0.2}\ \underbrace{\text{gait}}_{11.2},\ \underbrace{\text{suppleness}}_{13.7}\ \underbrace{\text{and}}_{1.6}\ \underbrace{\text{agility}}_{8.6}\ \underbrace{\text{were}}_{2.1}\ \underbrace{\text{discernible}}_{10.1}$$

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses

$$\underbrace{\text{In}}_{1.1}\ \underbrace{\text{the}}_{0.3}\ \underbrace{\text{voluntary}}_{3.6}\ \underbrace{\text{slowness}}_{4.5}\ \underbrace{\text{of}}_{1}\ \underbrace{\text{its}}_{0.2}\ \underbrace{\text{gait}}_{11.2},\ \underbrace{\text{suppleness}}_{13.7}\ \underbrace{\text{and}}_{1.6}\ \underbrace{\text{agility}}_{8.6}\ \underbrace{\text{were}}_{2.1}\ \underbrace{\text{discernible}}_{10.1}$$

Rank losses from smallest to highest

0.2 0.3 1 1.6 2.1 3.6 4.5 8.6 10.1 11.2 13.7

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses

In the voluntary slowness of its gait, suppleness and agility were discernible

1.1 0.3 3.6 4.5 1 0.2 11.2 13.7 1.6 8.6 2.1 10.1

Rank losses from smallest to highest

0.2 0.3 1 1.6 2.1 3.6 4.5 8.6 10.1 11.2 13.7

Sum and assign higher weight to larger losses

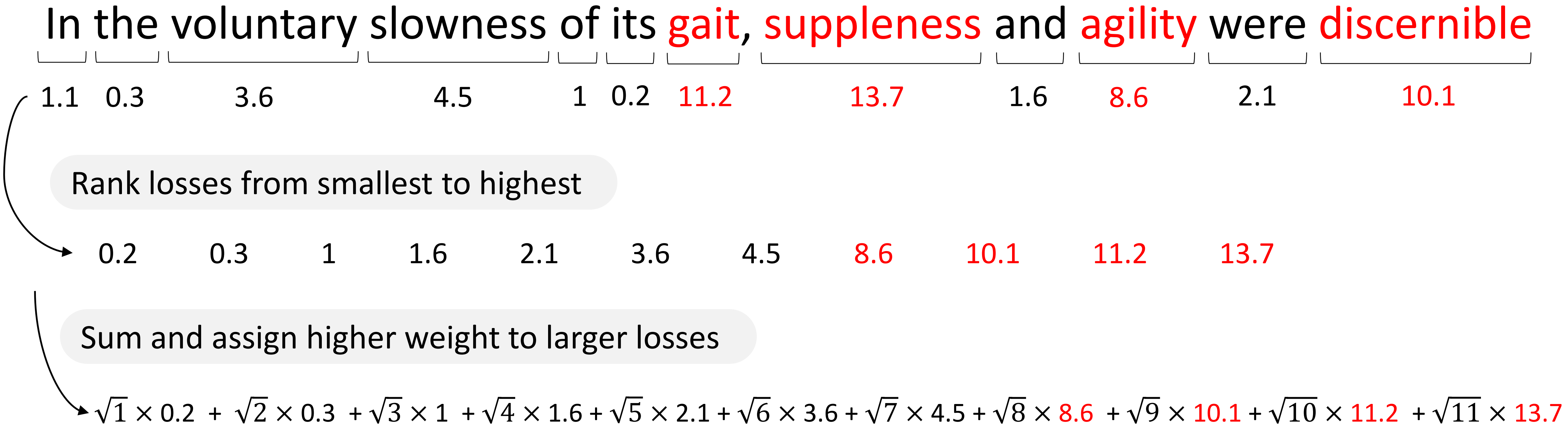
$$\sqrt{1} \times 0.2 + \sqrt{2} \times 0.3 + \sqrt{3} \times 1 + \sqrt{4} \times 1.6 + \sqrt{5} \times 2.1 + \sqrt{6} \times 3.6 + \sqrt{7} \times 4.5 + \sqrt{8} \times 8.6 + \sqrt{9} \times 10.1 + \sqrt{10} \times 11.2 + \sqrt{11} \times 13.7$$

LM-based Unsupervised Metric

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \times \mathrm{WNLL}(i)}{|S|}$$

Combines features (length) with LM statistics

Assumes difficult words have high losses



In the voluntary slowness of its gait, suppleness and agility were discernible

Rank losses from smallest to highest

Sum and assign higher weight to larger losses

$$\sqrt{1} \times 0.2 + \sqrt{2} \times 0.3 + \sqrt{3} \times 1 + \sqrt{4} \times 1.6 + \sqrt{5} \times 2.1 + \sqrt{6} \times 3.6 + \sqrt{7} \times 4.5 + \sqrt{8} \times 8.6 + \sqrt{9} \times 10.1 + \sqrt{10} \times 11.2 + \sqrt{11} \times 13.7$$

Performance in Cross-lingual Transfer

Train models on English portions of ReadMe++, CEFR-SP *(Wikipedia)* & CompDS *(News)*

Compare transfer performance to non-English languages

Arabic, **Hindi**, **French**, & **Russian** from ReadMe++, **Italian** *(Brunato et al. 2018)* and **German** *(Naderi et al. 2019)*

	Training Source Dataset		
Source → Target	*ReadMe++*	*CEFR-SP*	*CompDS*
English → Arabic	**0.606**	0.071	0.322
English → Hindi	**0.702**	0.267	0.381
English → French	**0.768**	-0.026	0.335
English → Russian	**0.760**	0.173	0.412
English → Italian	**0.239**	-0.043	0.099
English → German	**0.701**	-0.092	0.408

LMs trained on ReadMe++ perform better cross-lingual transfer compared with past datasets