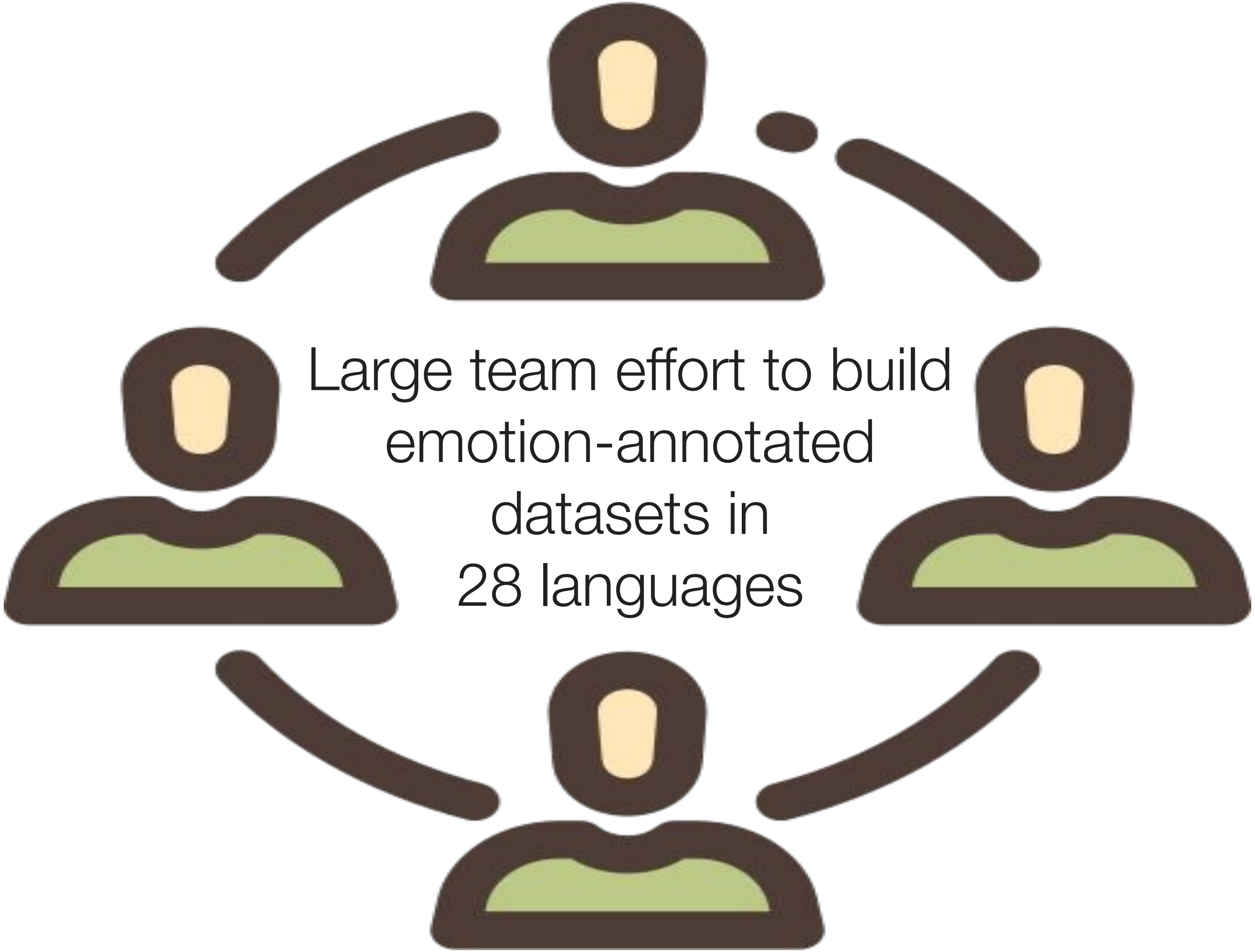
ACL 2025
VIENNA
JULY 27 - AUGUST 1

✨ BRIGHTER ✨

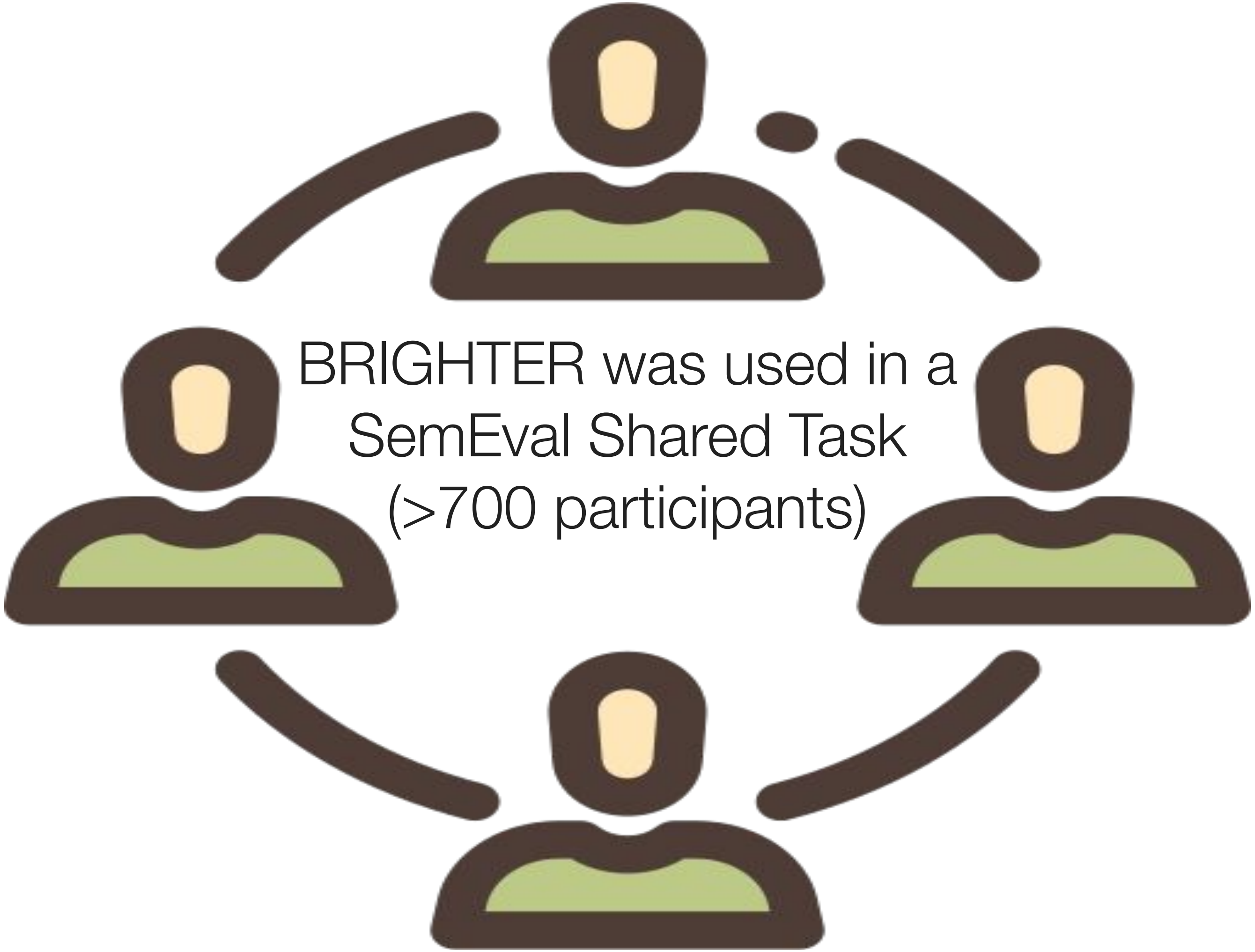
BRIdging the Gap in Human-Annotated Textual Emotion Recognition Datasets for 28 Languages

S. H. Muhammad*, **N. Ousidhoum***, I. Abdulmumin, J. P. Wahle, T. Ruas, M. Beloucif, C. de Kock, N. Surange, D. Teodorescu, I. S. Ahmad, D. I. Adelani, A. F. Aji, F. D. M. A. Ali, I. Alimova, V. Araujo, N. Babakov, N. Baes, A.-M. Bucur, A. Bukula, G. Cao, R. Tufiño, R. Chevi, C. I. Chukwuneke, A. Ciobotaru, D. Dementieva, M. S. Gadanya, R. Geislinger, B. Gipp, O. Hourrane, O. Ignat, F. I. Lawan, R. Mabuya, R. Mahendra, V. Marivate, A. Piper, A. Panchenko, C. H. Porto Ferreira, V. Protasov, S. Rutunda, M. Shrivastava, A. C. Udrea, L. D. A. Wanzare, S. Wu, F. V. Wunderlich, H. M. Zhafran, T. Zhang, Y. Zhou, S. M. Mohammad.

https://brighter-dataset.github.io



Large team effort to build
emotion-annotated
datasets in
28 languages

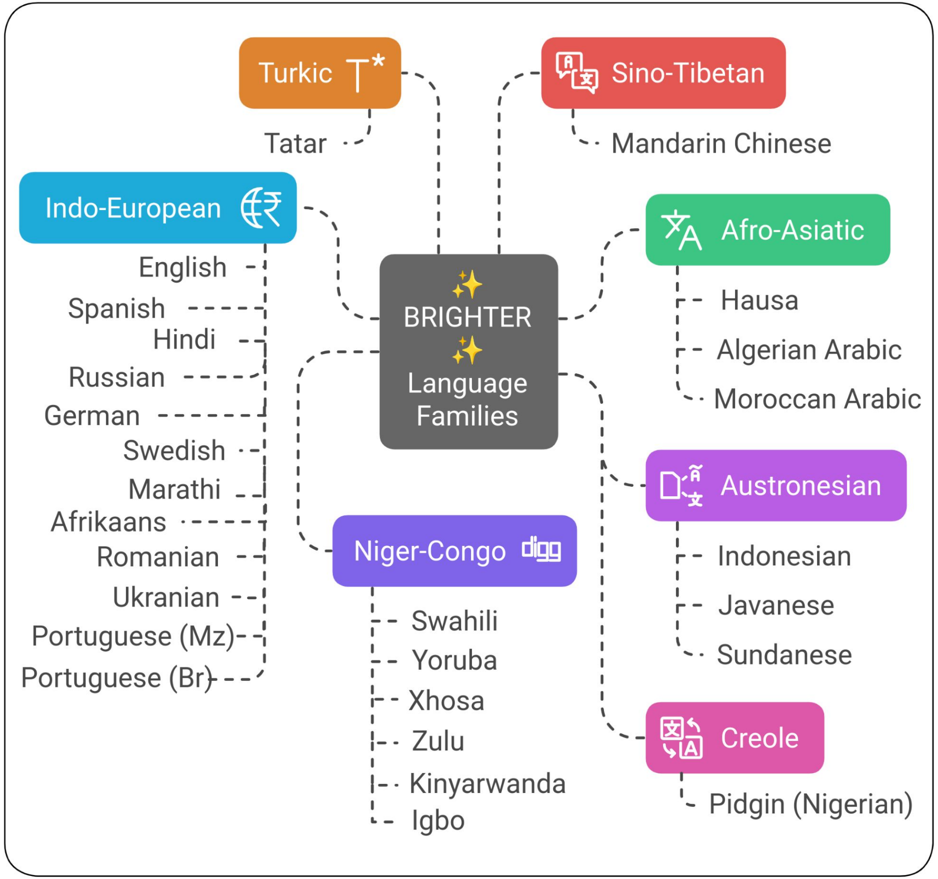


BRIGHTER was used in a
SemEval Shared Task
(>700 participants)

BRIGHTER: Coverage

BRIGHTER primarily covers low-resource languages from Africa, Asia, Eastern Europe, Latin America

BRIGHTER: Coverage of 28 languages



✨ BRIGHTER ✨ Language Families

Turkic
- Tatar

Sino-Tibetan
- Mandarin Chinese

Indo-European
- English
- Spanish
- Hindi
- Russian
- German
- Swedish
- Marathi
- Afrikaans
- Romanian
- Ukranian
- Portuguese (Mz)
- Portuguese (Br)

Afro-Asiatic
- Hausa
- Algerian Arabic
- Moroccan Arabic

Austronesian
- Indonesian
- Javanese
- Sundanese

Niger-Congo
- Swahili
- Yoruba
- Xhosa
- Zulu
- Kinyarwanda
- Igbo

Creole
- Pidgin (Nigerian)

BRIGHTER: Emotion Recognition Datasets

- BRIGHTER focuses on **perceived emotions**
  - I.e., emotion(s) most people think the speaker might have felt given a text snippet uttered by them
- The datasets are multi-labeled

Dataset Construction

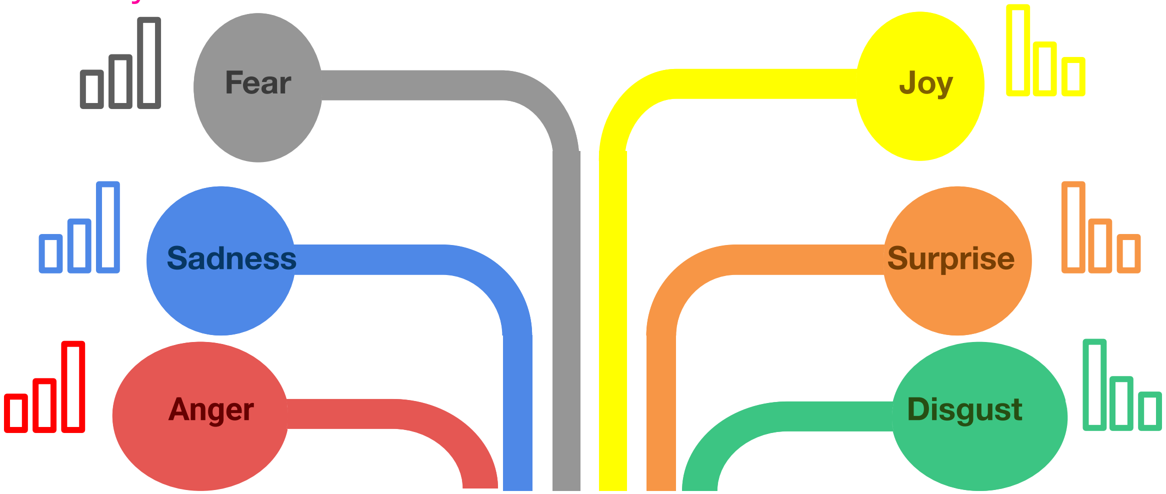
Data Collection

- We targeted emotionally rich text (e.g., personal narratives)
- Eventually, we used various sources depending on the availability text data
  - Social media (e.g., Reddit in English, German, Romanian, others)
  - Speeches (e.g., in Afrikaans)
  - A translated novel (e.g., in Algerian Arabic)
  - News data and combined sources when text sources are scarce

Dataset Construction

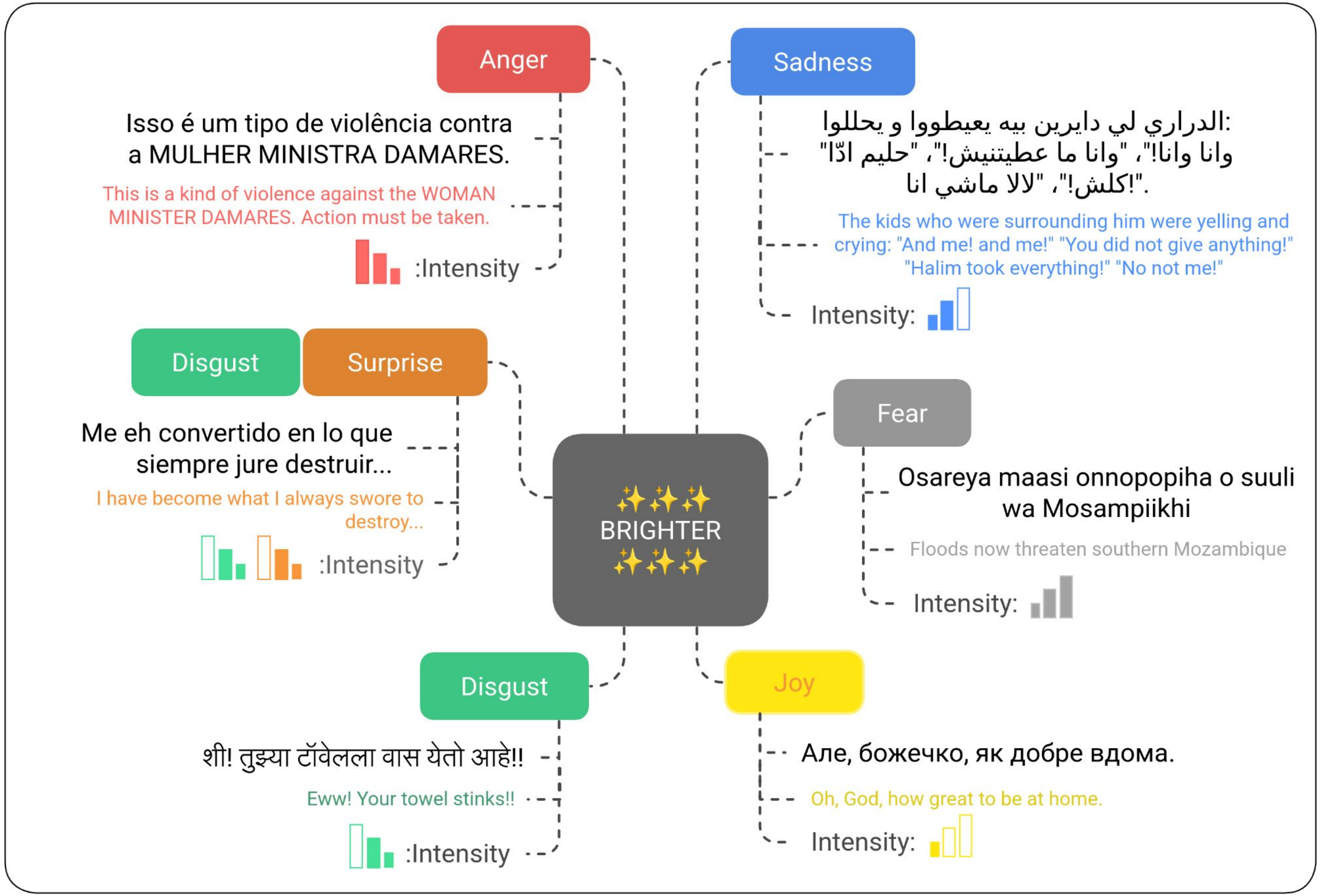
Data Annotation

Given a text snippet, we asked the annotators to select all the emotions that apply on a 4-level intensity scale



If no emotion is selected then the text is considered neutral

BRIGHTER: Multi-labeled Datasets



✨✨✨
BRIGHTER
✨✨✨

Anger

Isso é um tipo de violência contra
a MULHER MINISTRA DAMARES.

This is a kind of violence against the WOMAN
MINISTER DAMARES. Action must be taken.

:Intensity

Sadness

الدراري لي دايرين بيه يعيطووا و يحللوا:
وانا وانا!"، "وانا ما عطيتنيش!"، "حليم ادّا"
كلش!"، "لالا ماشي انا!".

The kids who were surrounding him were yelling and
crying: "And me! and me!" "You did not give anything!"
"Halim took everything!" "No not me!"

Intensity:

Disgust Surprise

Me eh convertido en lo que
siempre jure destruir...

I have become what I always swore to
destroy...

:Intensity

Fear

Osareya maasi onnopopiha o suuli
wa Mosampiikhi

Floods now threaten southern Mozambique

Intensity:

Disgust

शी! तुझ्या टॉवेलला वास येतो आहे!!

Eww! Your towel stinks!!

:Intensity

Joy

Але, божечко, як добре вдома.

Oh, God, how great to be at home.

Intensity:

Dataset Construction

Quality Control

Intensity scores

Intensity scores are kept for datasets >=5 annotators per instance (i.e., 10 languages)

Label determination

Final labels are chosen based on agreement and intensity score threshold

Pre-processing

Text is processed by native speakers.

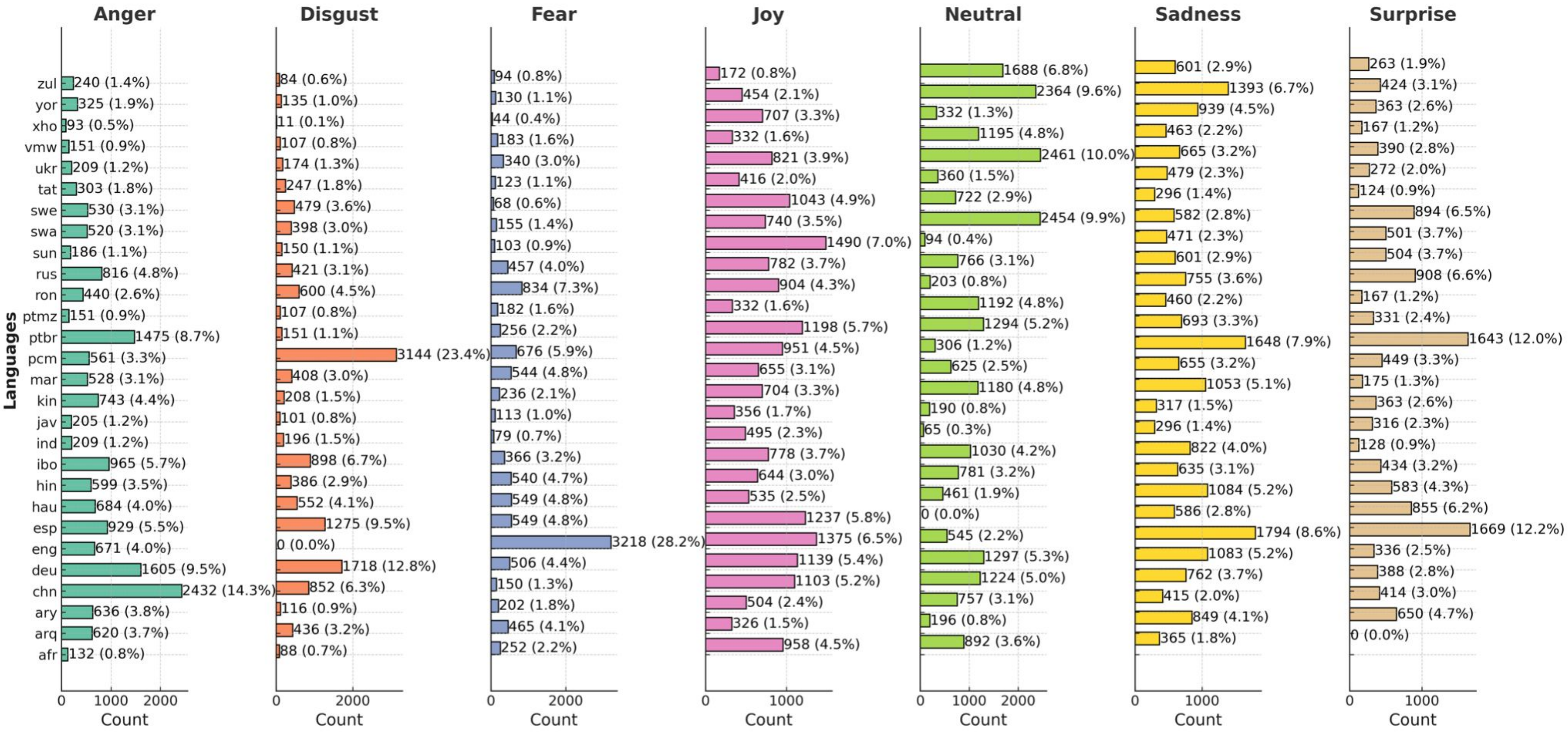
Annotation

Annotators are native speakers, >=3 annotators per instance

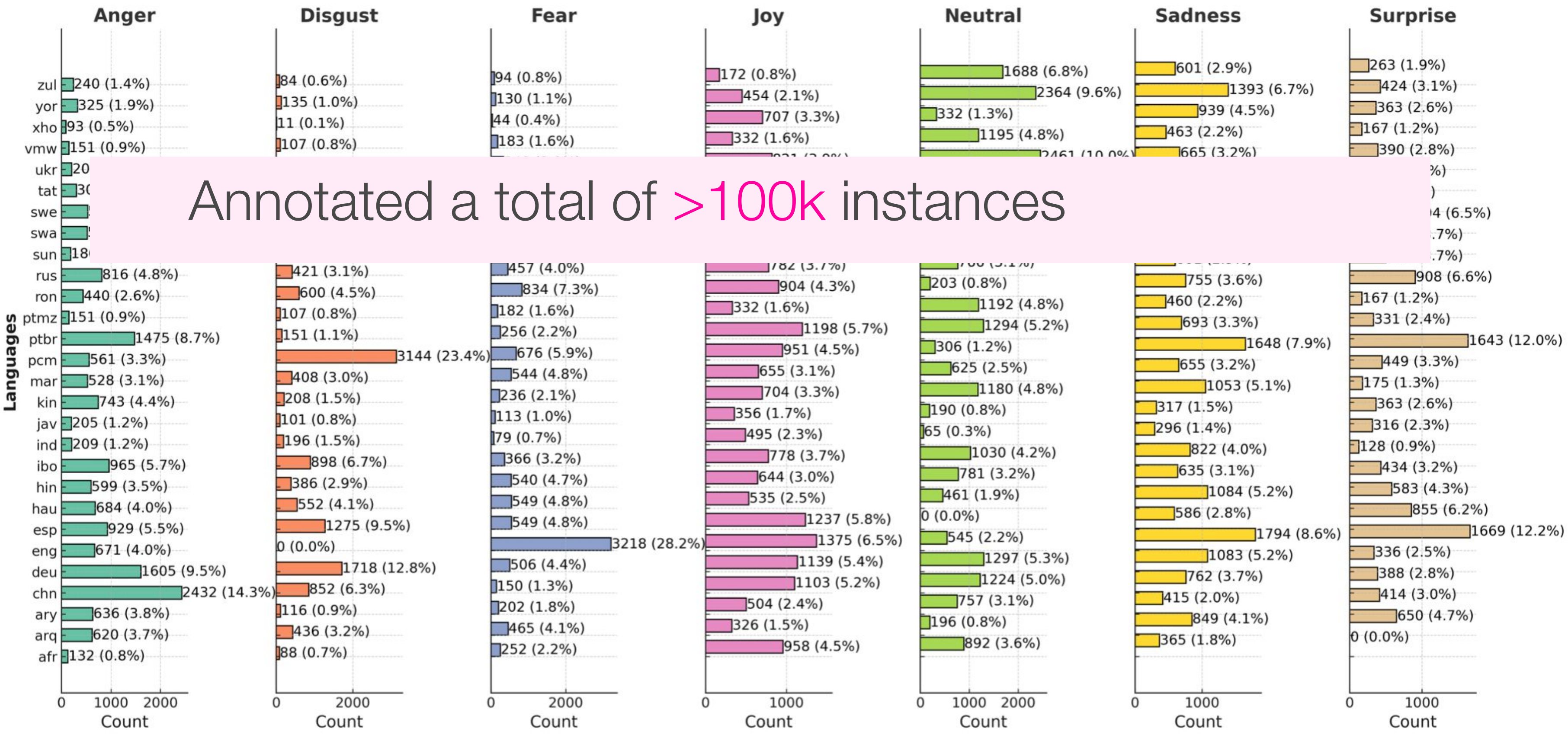
Reliability scores of the final datasets

Reliability Scores >62%

BRIGHTER: Final Datasets

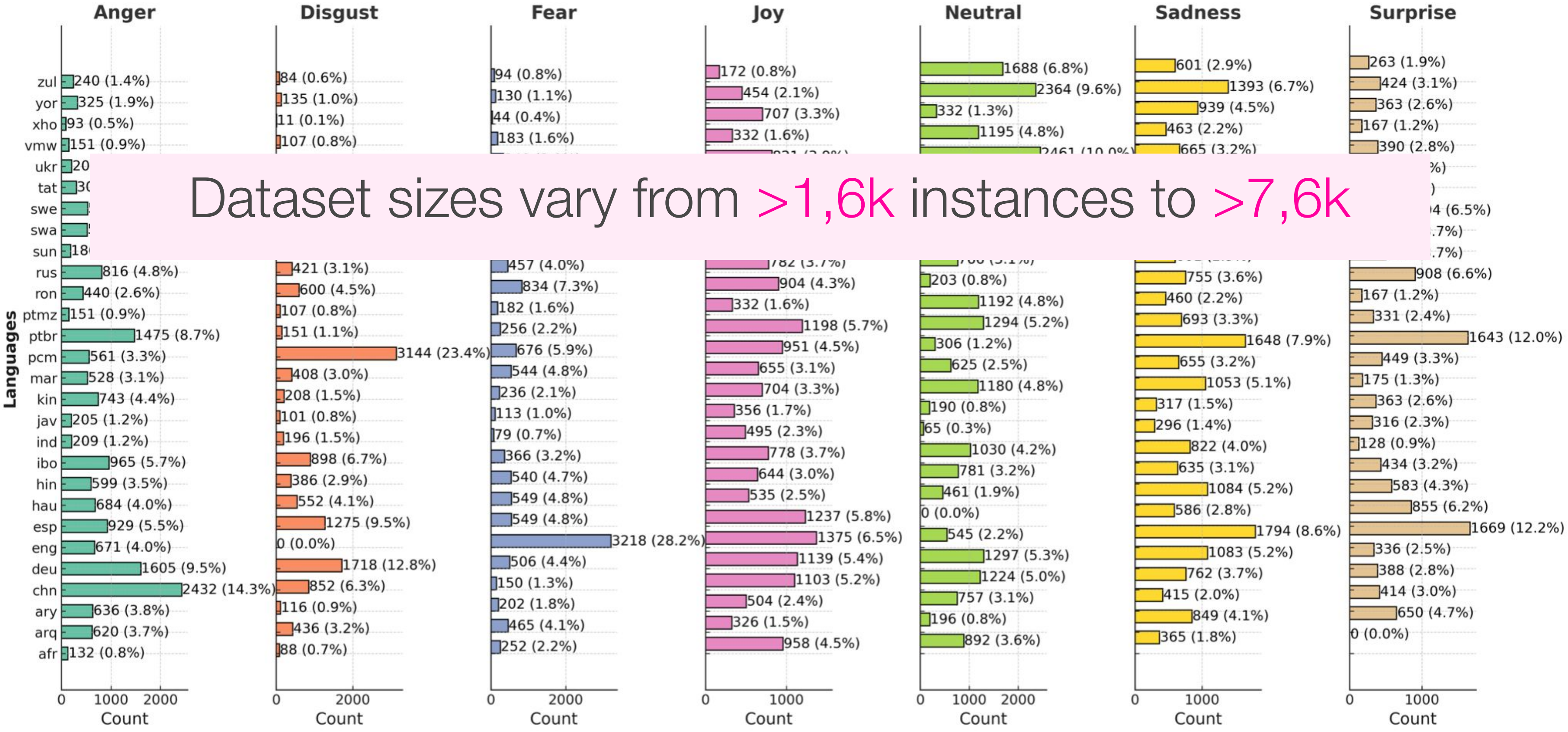


BRIGHTER: Final Datasets



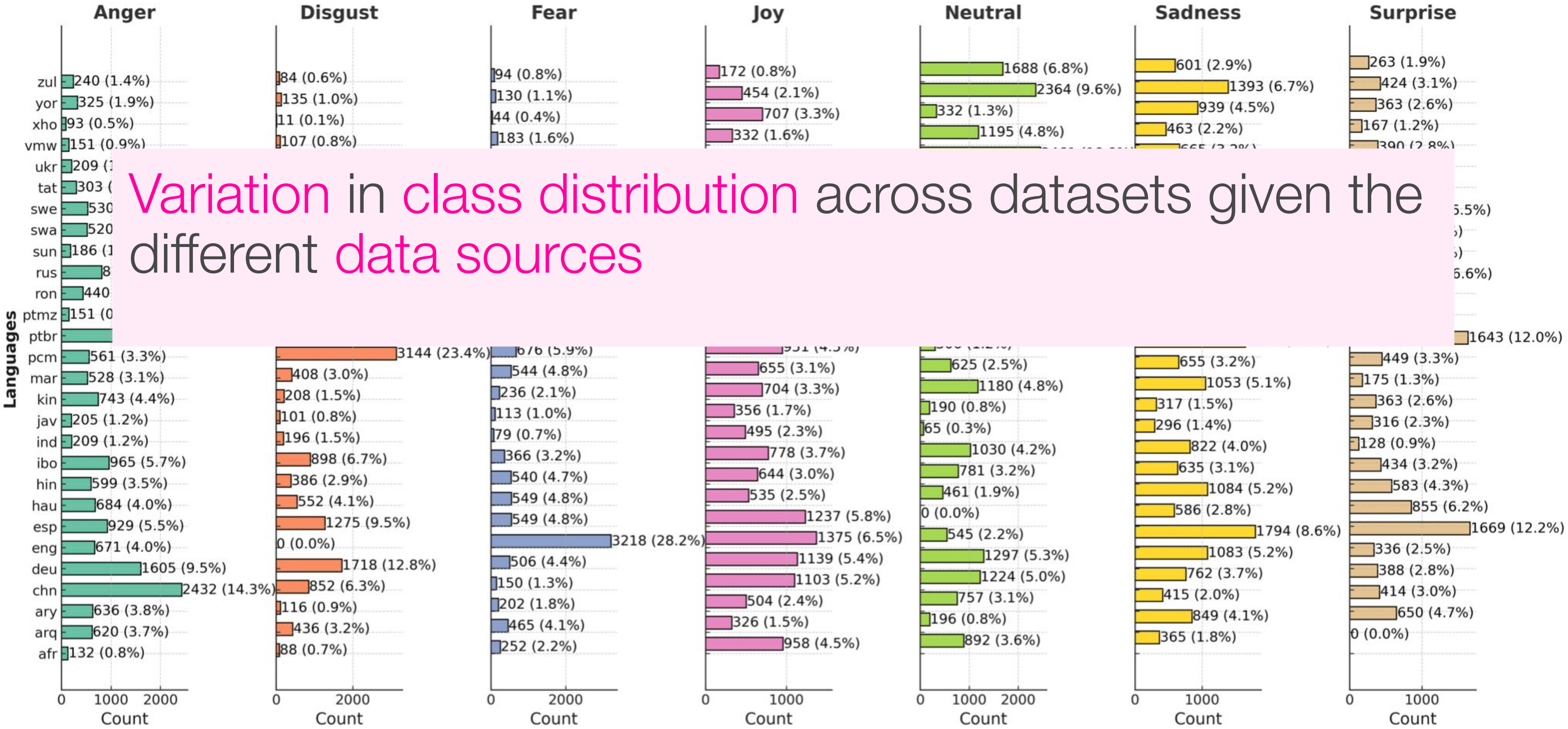
Annotated a total of >100k instances

BRIGHTER: Final Datasets



Dataset sizes vary from >1,6k instances to >7,6k

BRIGHTER: Final Datasets



Variation in class distribution across datasets given the different data sources

Experiments

Multi-label Emotion Classification

The results were highly language-dependent

	afr	arq	ary	chn	deu	eng	esp	hau	hin	ibo	ind	jav	kin	mar	pcm	ptbr	ptmz	ron	rus	sun	swa	swe
Qwen	60.18	37.78	52.76	55.23	59.17	55.72	72.33	43.79	79.73	37.4	57.29	50.47	31.96	74.58	38.66	51.6	40.44	68.18	73.08	42.67	27.36	48.89
Dolly	23.58	38.59	24.27	27.52	26.86	42.6	36.41	29.43	27.59	24.31	36.61	36.18	19.73	25.69	34.41	25.9	16.7	43.58	29.72	32.2	17.63	21.79
Llama	61.28	55.75	44.96	53.36	56.99	**65.58**	61.27	50.91	60.59	33.18	39.2	41.88	**34.36**	67.4	48.67	45.03	34.06	71.28	62.61	46.33	29.47	50.26
Mixtral	53.69	45.29	35.07	44.91	51.2	58.12	65.72	40.4	62.19	31.9	54.37	48.37	26.35	50.36	45.61	41.64	36.52	68.51	61.72	42.1	26.51	48.61
Deep Seek	43.66	50.87	47.21	53.45	54.26	56.99	**73.29**	51.91	76.91	32.85	49.51	43.05	32.52	76.68	45	51.49	39.58	65.02	76.97	44.61	33.27	44.6

Experiments

Multi-label Emotion Classification

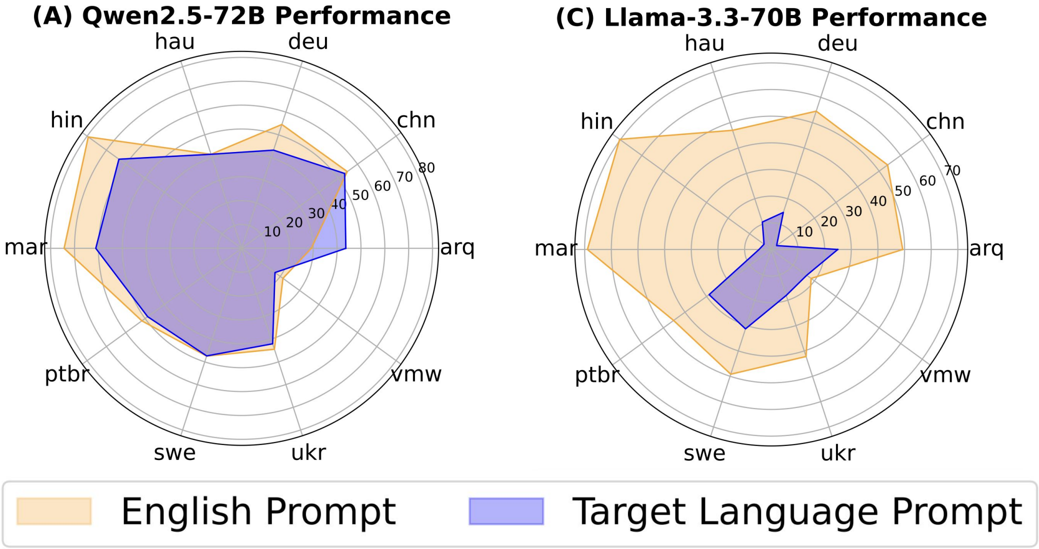
Qwen2.5-72B performed the best on average

	afr	arq	ary	chn	deu	eng	esp	hau	hin	ibo	ind	jav	kin	mar	pcm	ptbr	ptmz	ron	rus	sun	swa	swe
Qwen	**60.18**	**37.78**	**52.76**	**55.23**	**59.17**	**55.72**	**72.33**	**43.79**	**79.73**	**37.4**	**57.29**	**50.47**	**31.96**	**74.58**	**38.66**	**51.6**	**40.44**	**68.18**	**73.08**	**42.67**	**27.36**	**48.89**
Dolly	23.58	38.59	24.27	27.52	26.86	42.6	36.41	29.43	27.59	24.31	36.61	36.18	19.73	25.69	34.41	25.9	16.7	43.58	29.72	32.2	17.63	21.79
Llama	61.28	55.75	44.96	53.36	56.99	65.58	61.27	50.91	60.59	33.18	39.2	41.88	34.36	67.4	48.67	45.03	34.06	71.28	62.61	46.33	29.47	50.26
Mixtral	53.69	45.29	35.07	44.91	51.2	58.12	65.72	40.4	62.19	31.9	54.37	48.37	26.35	50.36	45.61	41.64	36.52	68.51	61.72	42.1	26.51	48.61
Deep Seek	43.66	50.87	47.21	53.45	54.26	56.99	73.29	51.91	76.91	32.85	49.51	43.05	32.52	76.68	45	51.49	39.58	65.02	76.97	44.61	33.27	44.6

Experiments

Sensitivity to the Language of the Prompt

LLMs generally perform better when prompted in English



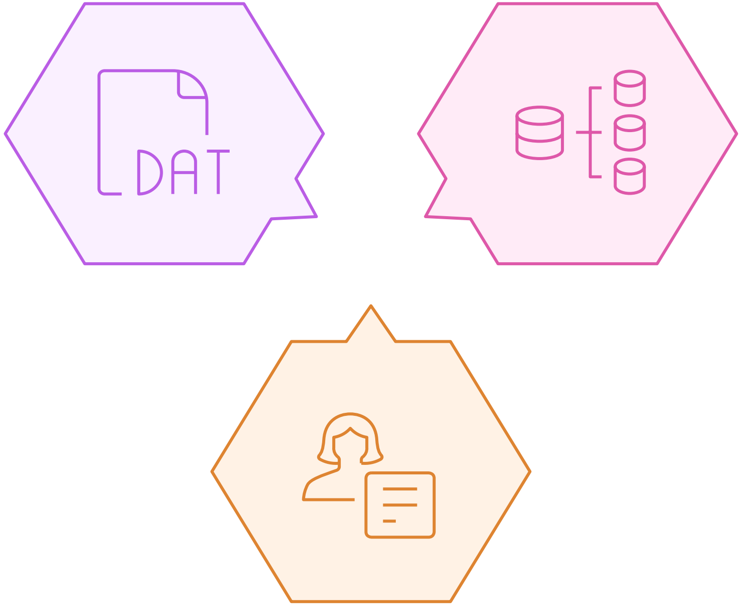
(A) Qwen2.5-72B Performance

(C) Llama-3.3-70B Performance

Takeaways from Additional Experiments

- LLMs still struggle with emotion recognition
- We observe large performance gaps across languages
- Performance still depends on prompt wording, number of shots and language

BRIGHTER Public Release



Annotations guidelines

Datasets

Individual labels

https://brighter-dataset.github.io

Thank you!

Any questions?

✨✨✨
BRIGHTER
✨✨✨