A Dataset for Sentiment and Sarcasm Classification for Varieties of English

Dr. Aditya Joshi (he/him) and Dipankar Srirag (he/him)
School of Computer Science Engineering,
UNSW Sydney

In collaboration with:

UNIVERSITY OF SURREY

Funded by:

Google Research

UNSW AI Symposium 2024
Organized by UNSW AI Institute

Current Benchmarks

In this case, benchmarks are datasets and evaluation results for language-related tasks.

Current Benchmarks

Model	URL	Score	CoLA	SST-2
Turing ULR v6	↗	91.3	73.3	97.5
Vega v1		91.3	73.8	97.9
Turing NLR v5	↗	91.2	72.6	97.6
DeBERTa + CLEVER		91.1	74.7	97.6
ERNIE	↗	91.1	75.5	97.8
StructBERT + CLEVER	↗	91.0	75.3	97.7
DeBERTa / TuringNLRv4	↗	90.8	71.5	97.5

In this case, benchmarks are datasets and evaluation results for language-related tasks.

Current Benchmarks

Model	URL	Score	CoLA	SST-2
Turing ULR v6	↗	91.3	73.3	97.5
Vega v1		91.3	73.8	97.9
Turing NLR v5	↗	91.2	72.6	97.6
DeBERTa + CLEVER		91.1	74.7	97.6
ERNIE	↗	91.1	75.5	97.8
StructBERT + CLEVER	↗	91.0	75.3	97.7
DeBERTa / TuringNLRv4	↗	90.8	71.5	97.5

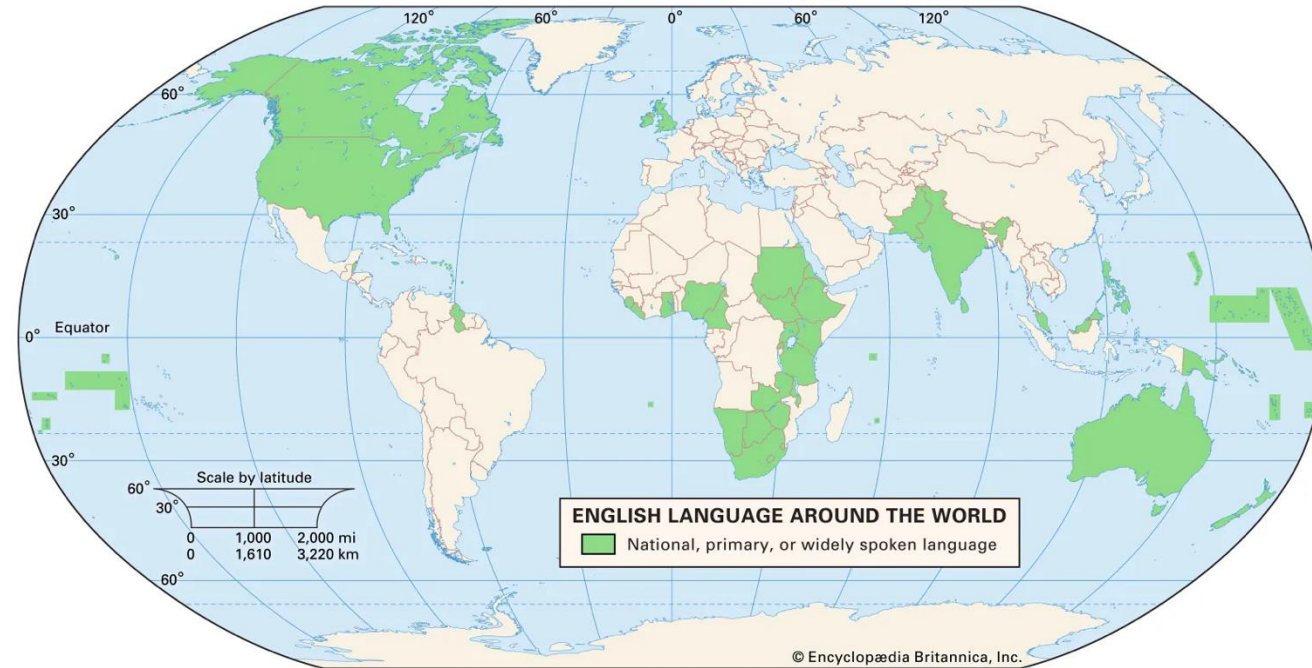In this case, benchmarks are datasets and evaluation results for language-related tasks.

What about World Englishes?

ENGLISH LANGUAGE AROUND THE WORLD
National, primary, or widely spoken language
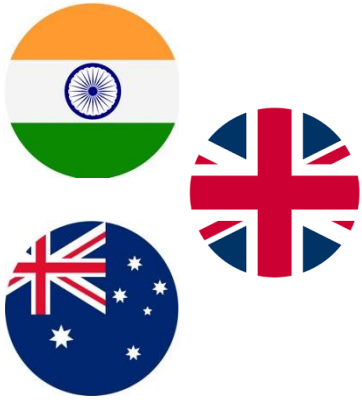
© Encyclopædia Britannica, Inc.

What about World Englishes?

We have created a benchmark for sentiment and sarcasm classification for three varieties of English.

How we created the dataset for this benchmark

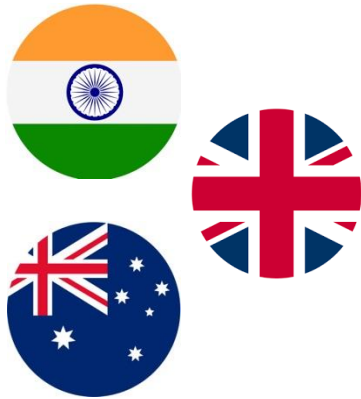# How we created the dataset for this benchmark

Language Varieties
- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)

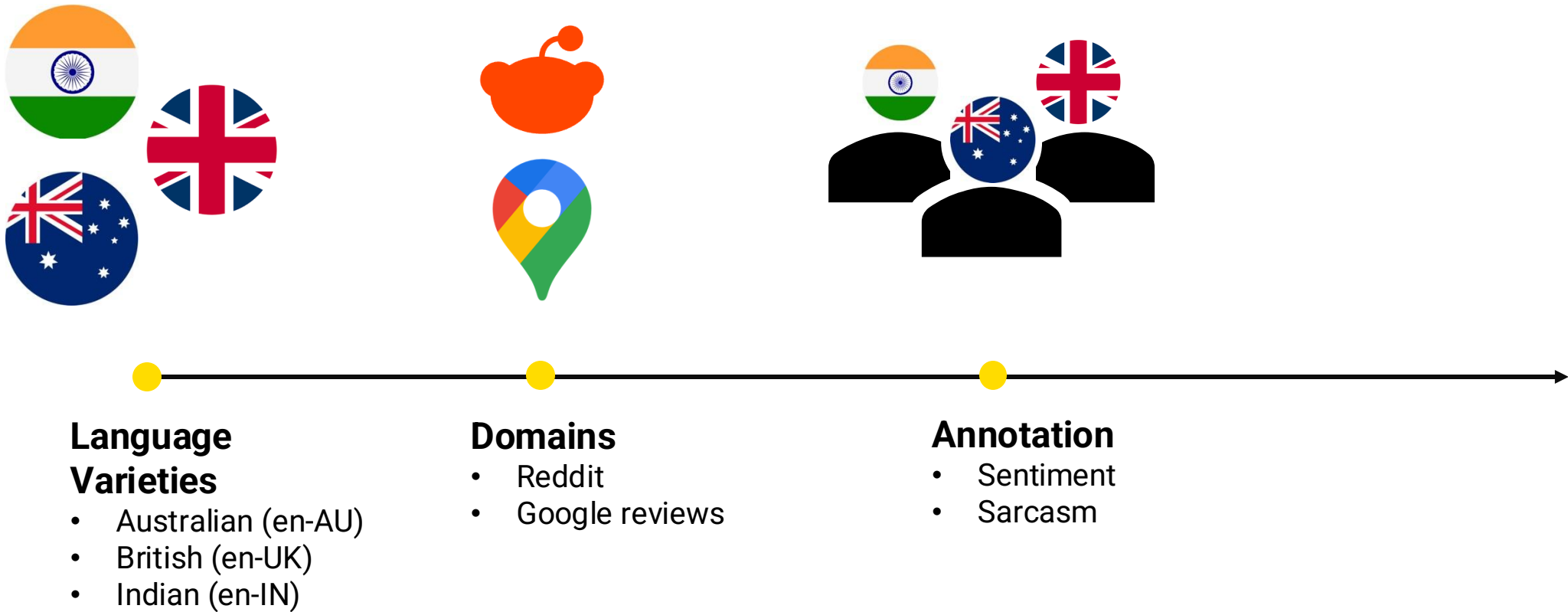# How we created the dataset for this benchmark

Language Varieties
- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)
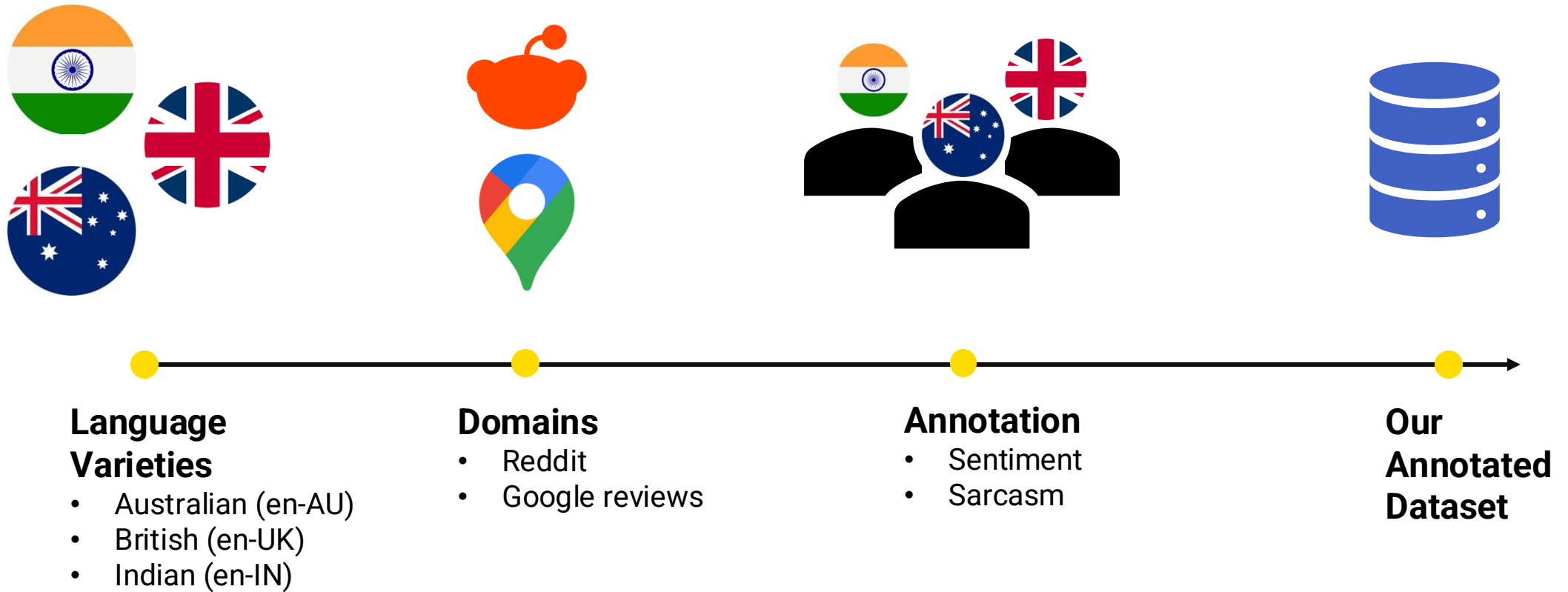
Domains
- Reddit
- Google reviews

How we created the dataset for this benchmark

Language Varieties
- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)

Domains
- Reddit
- Google reviews

Annotation
- Sentiment
- Sarcasm

How we created the dataset for this benchmark

Language Varieties
- Australian (en-AU)
- British (en-UK)
- Indian (en-IN)

Domains
- Reddit
- Google reviews

Annotation
- Sentiment
- Sarcasm

Our Annotated Dataset

Dataset Statistics

| Variety	Domain	Total Samples	%Samples with Positive Sentiment	%Samples with Positive Sarcasm	Avg. Language Probability	Avg. Number of Words
en-AU	Google	1346	73%	7%	0.99	64.0
	Reddit	2505	32%	42%	0.98	51.7
en-IN	Google	2342	75%	1%	0.99	44.3
	Reddit	2395	25%	13%	0.87	26.9
en-UK	Google	2582	75%	0%	0.99	72.2
	Reddit	1432	12%	22%	0.98	38.0

We annotate a combined total of **12,602** across all varieties and domains for both the tasks.
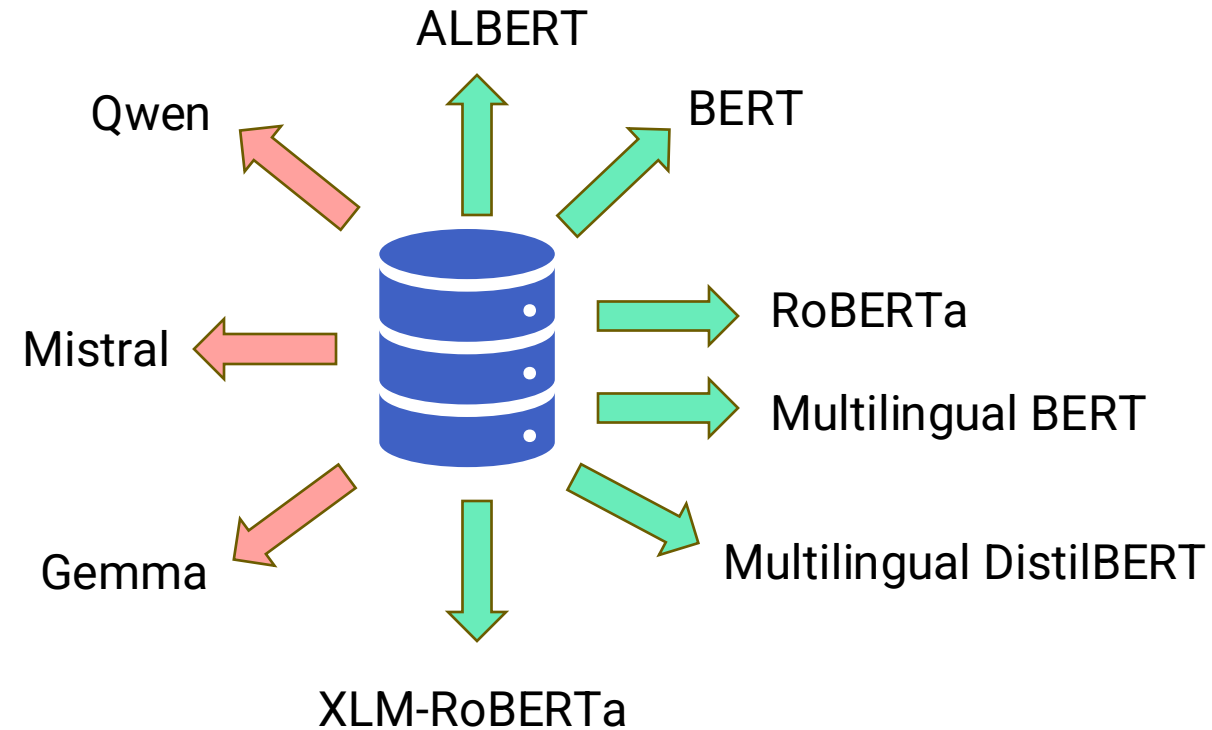
UNSW HREC Approval#: iRECS6514
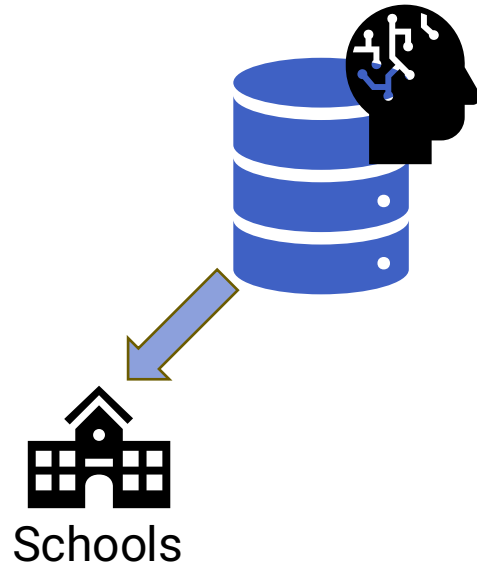
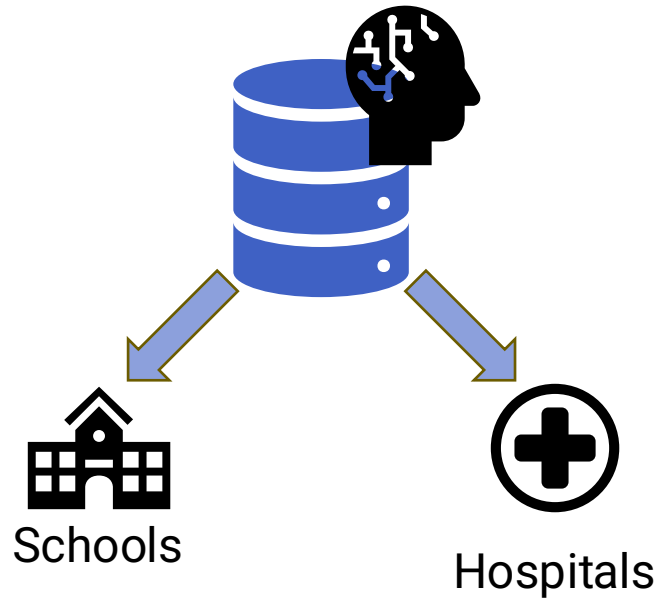# We evaluate NINE LLMs

We evaluate NINE LLMs
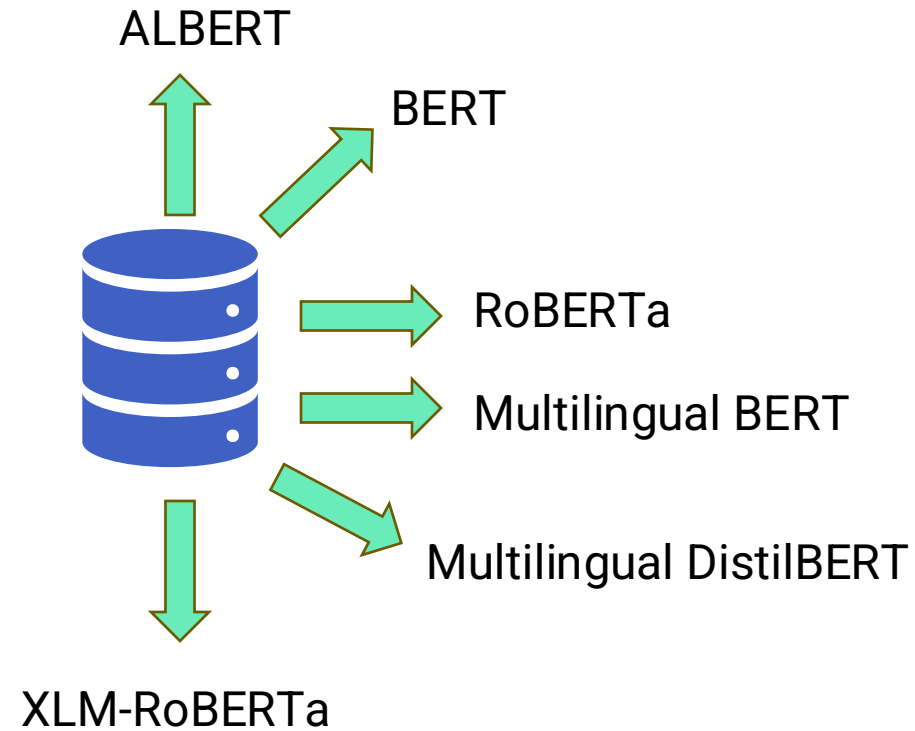
We evaluate NINE LLMs

Can we please connect if....

Can we please connect if….
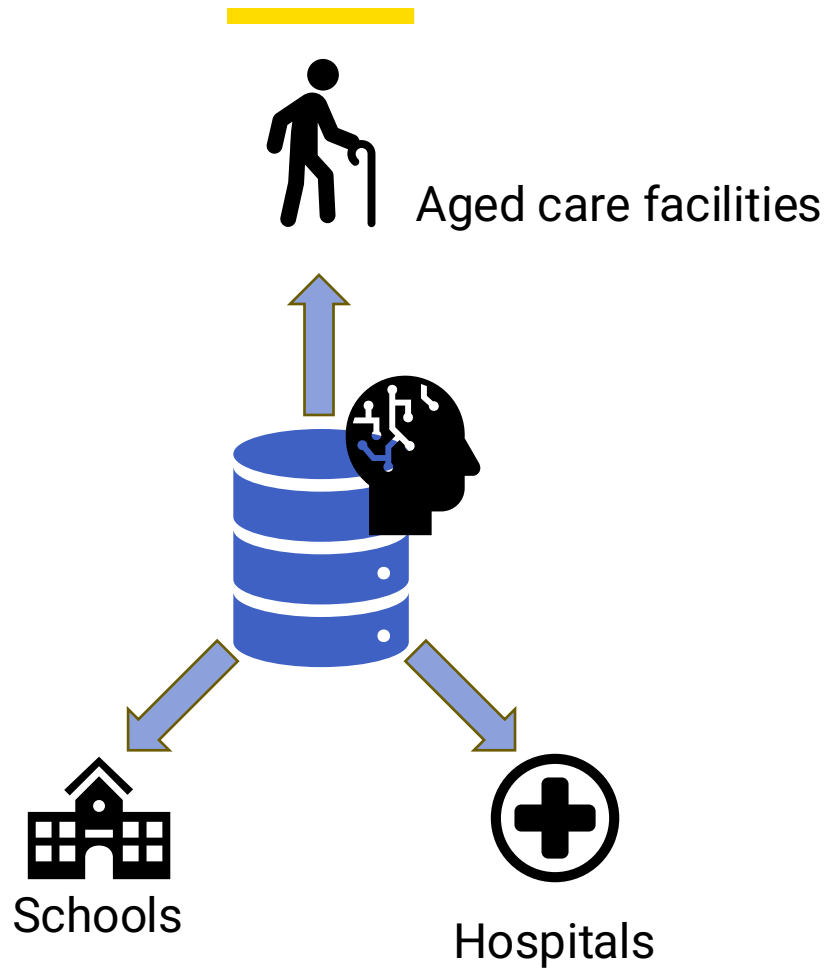
Schools

Can we please connect if....

Schools

Hospitals

Can we please connect if….

Aged care facilities

Schools

Hospitals

Can we please connect if….

Aged care facilities

…. many more.

Schools

Hospitals
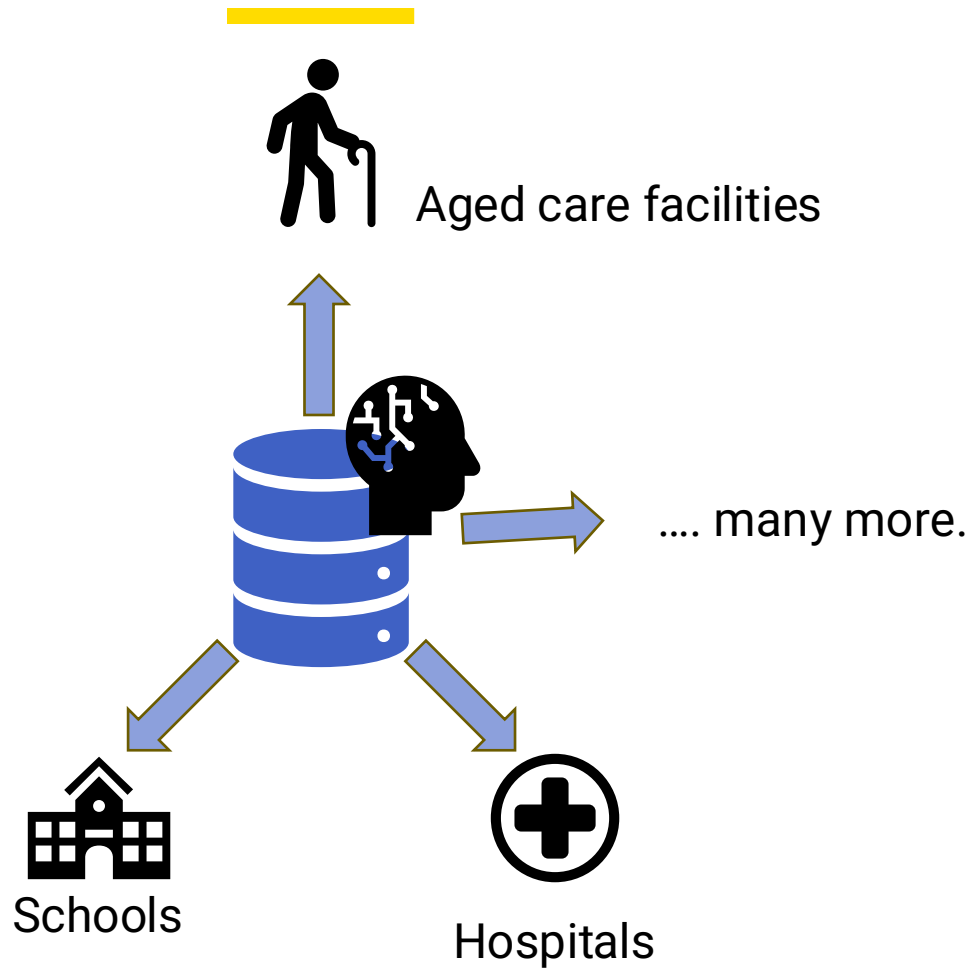
We would like to build LLM applications that serve ALL speakers of English.

We would like to build LLM applications that serve ALL speakers of English.

To collaborate:

Dr. Aditya Joshi: aditya.joshi@unsw.edu.au

Dipankar Srirag: d.srirag@unsw.edu.au