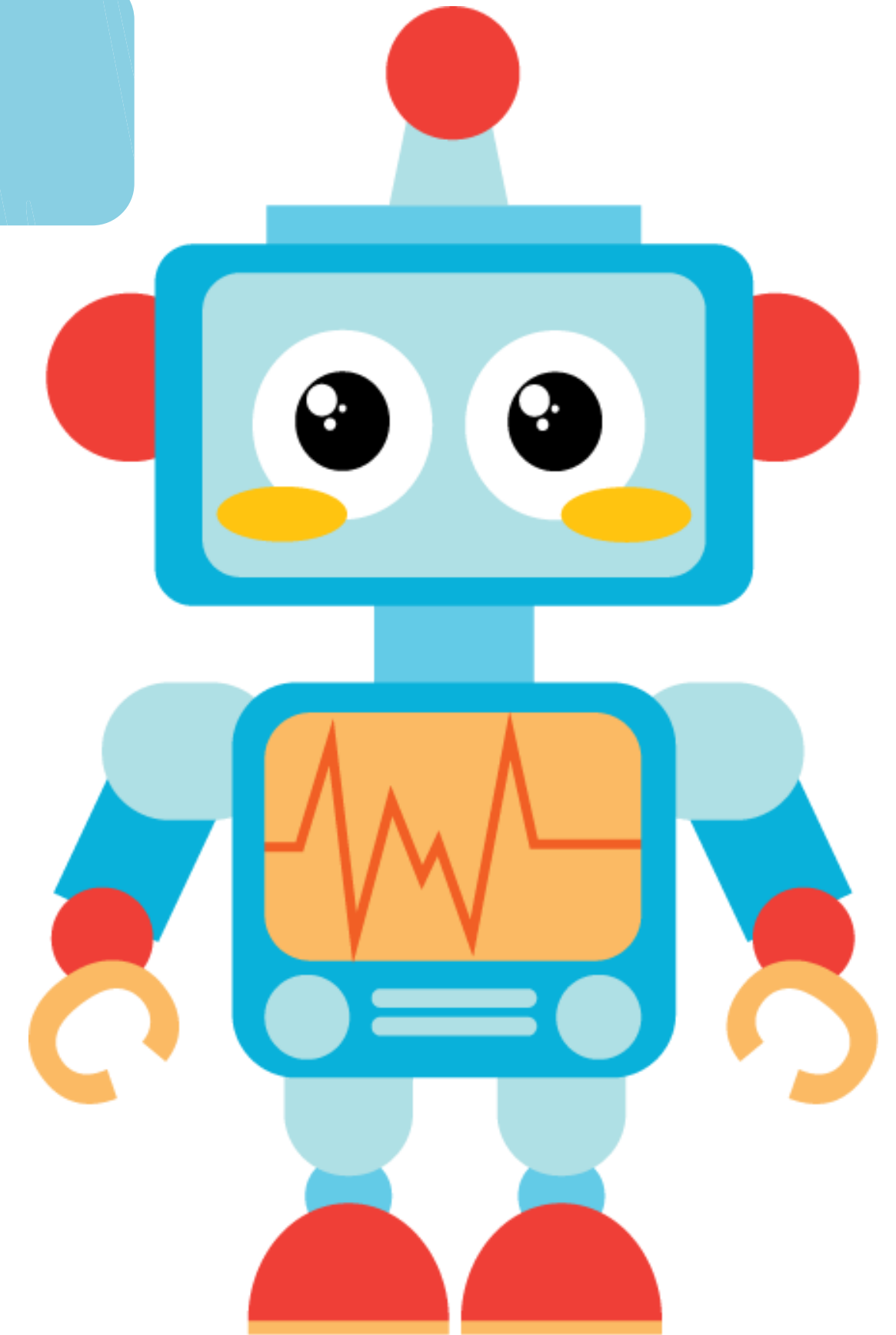
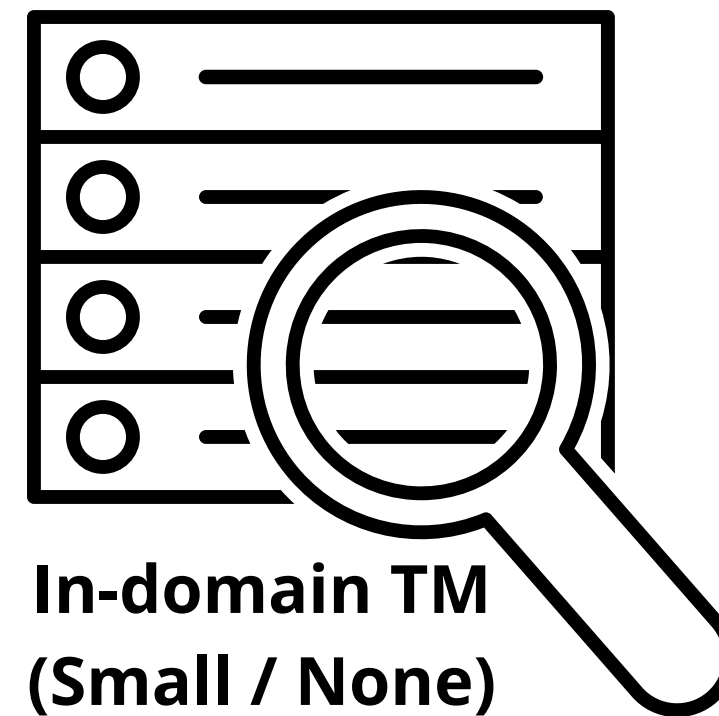
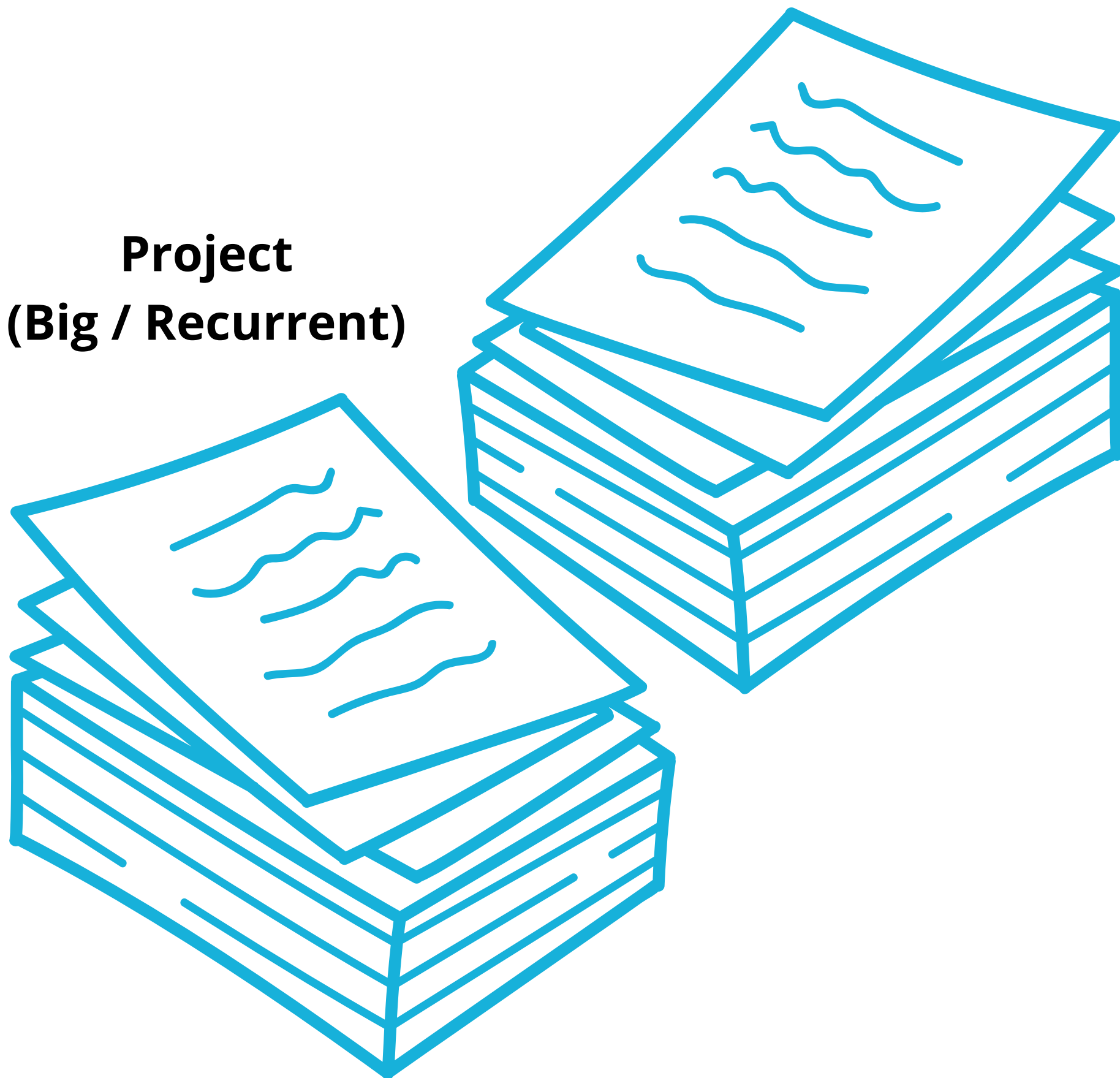


Domain-Specific Text Generation for Machine Translation

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, Andy Way

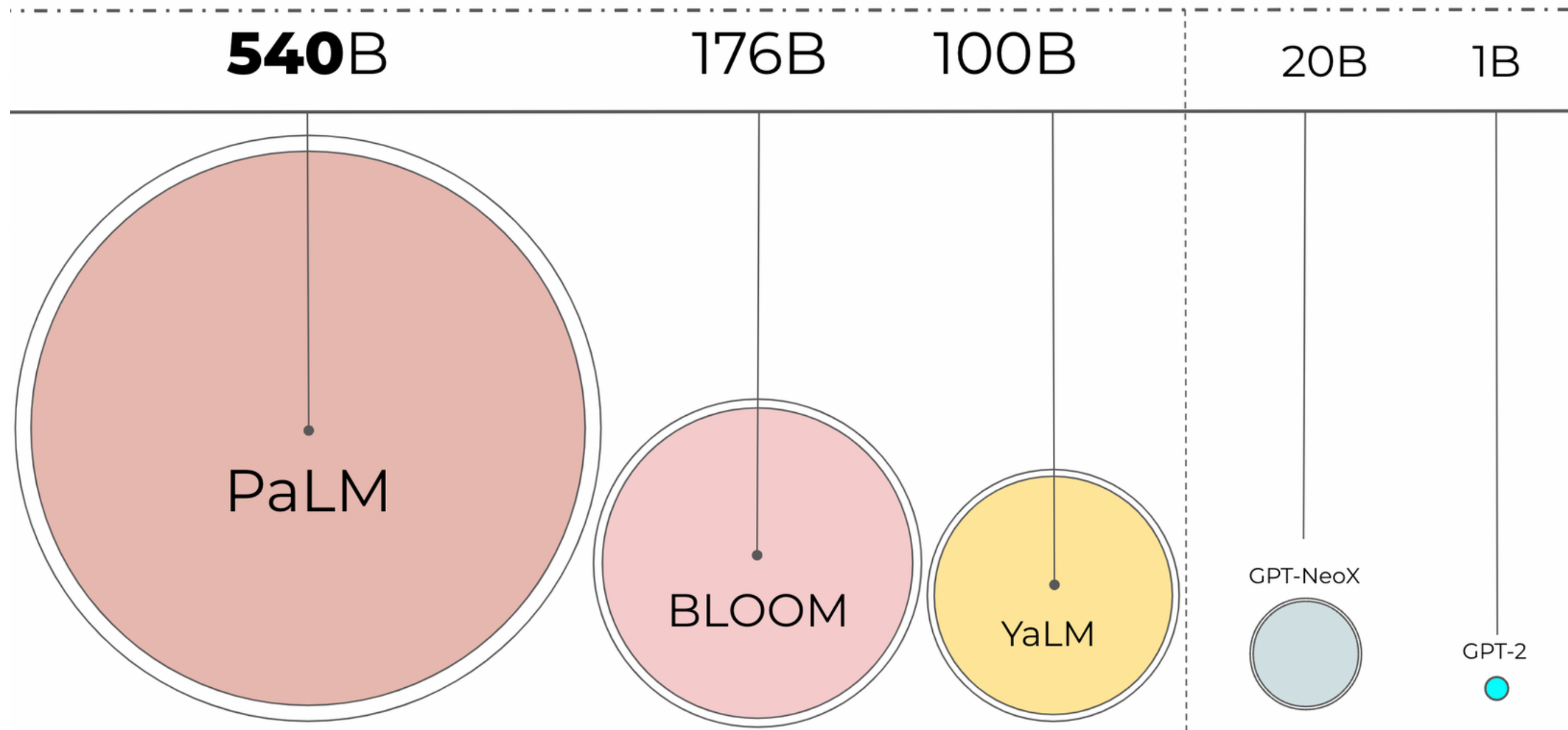


**Project
(Big / Recurrent)**

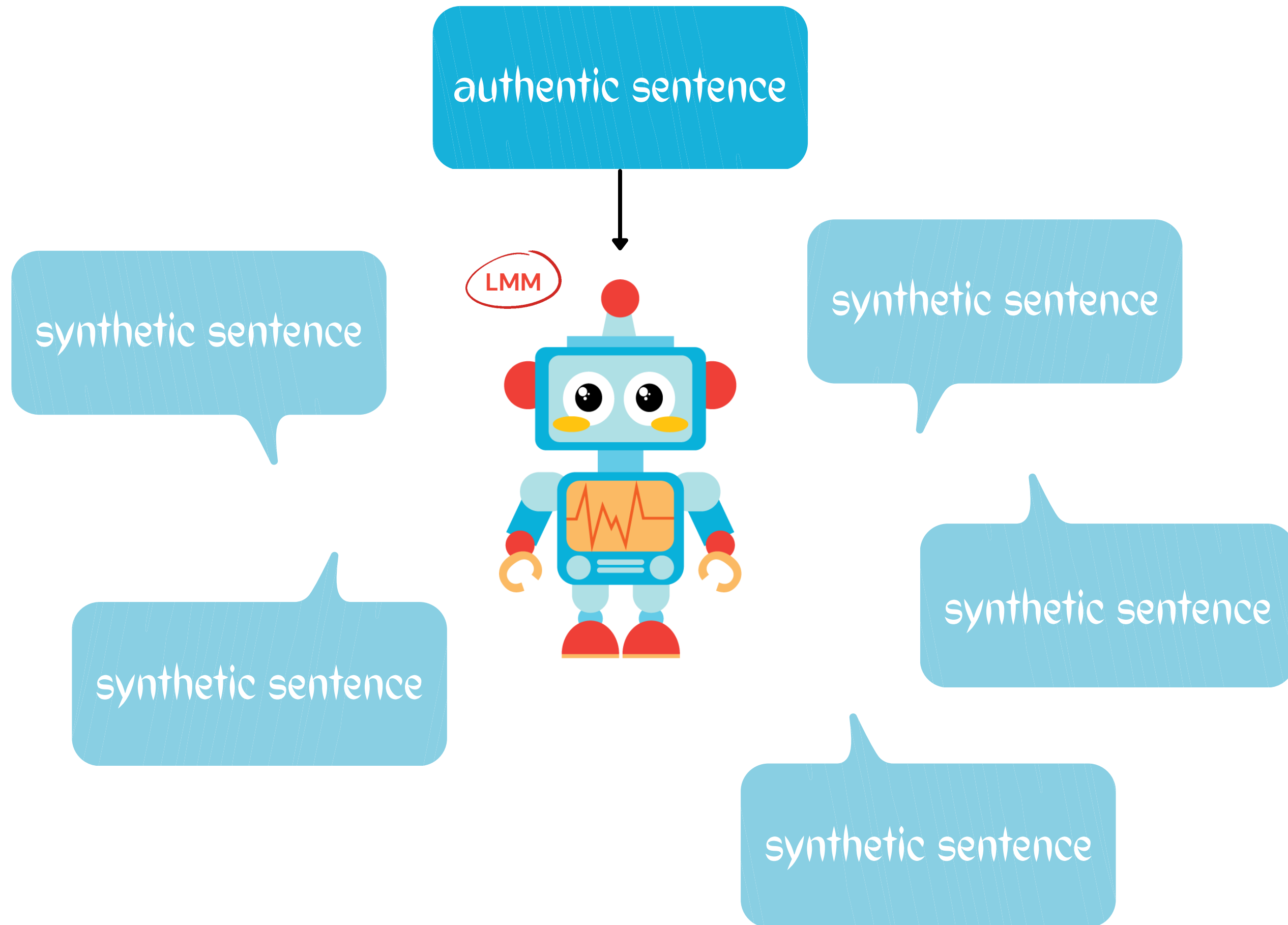


**In-domain TM
(Small / None)**

Large Language Models - sorted by billion parameters



Source: Hugging Face blog

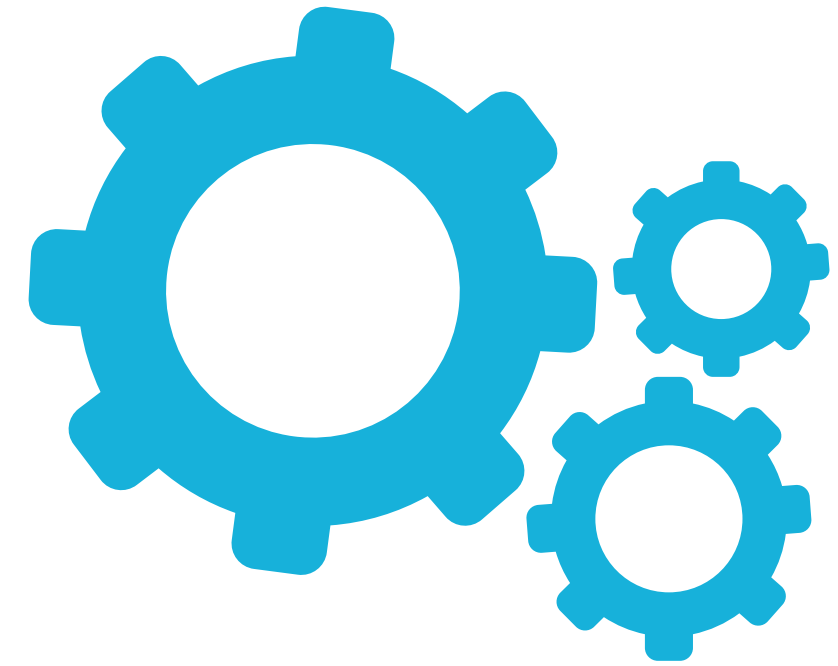


Setup #1: Small in-domain TM

1. Text Generation
2. Back-translation
3. Mixed Fine-tuning

Setup #2: No in-domain TM

0. Forward-translation
- ... then apply 1, 2, & 3



```
from transformers import GPTJForCausalLM, AutoTokenizer
import torch

tokenizer = AutoTokenizer.from_pretrained("EleutherAI/gpt-j-6B")

▼ model = GPTJForCausalLM.from_pretrained("EleutherAI/gpt-j-6B",
                                          revision="float16",
                                          torch_dtype=torch.float16,
                                          low_cpu_mem_usage=True,
                                          cache_dir = "models_cache/",
                                          pad_token_id=tokenizer.eos_token_id)


model = model.half()
model = model.to("cuda")
```

```
input_ids = tokenizer(tm_segment, return_tensors="pt").input_ids.to("cuda")  
▼ sample_outputs = model.generate(input_ids,  
                                do_sample=True,  
                                max_length=300,  
                                top_k=50,  
                                top_p=0.95,  
                                num_return_sequences=5,  
                                early_stopping=True)  
  
generated_text = tokenizer.batch_decode(sample_outputs, skip_special_tokens=True)
```


Semantically correct; possibly a fact



LM



In March 2020, India ordered the countrywide shut down of all non-essential economic activities due to the spreading COVID-19 pandemic.



While the overall worldwide economic impact of COVID-19 will only be realized through the end of 2020 and the recovery phase in 2021, it is clear that certain parts of the world have been severely impacted.

Semantically correct; possibly *fictitious*

LM

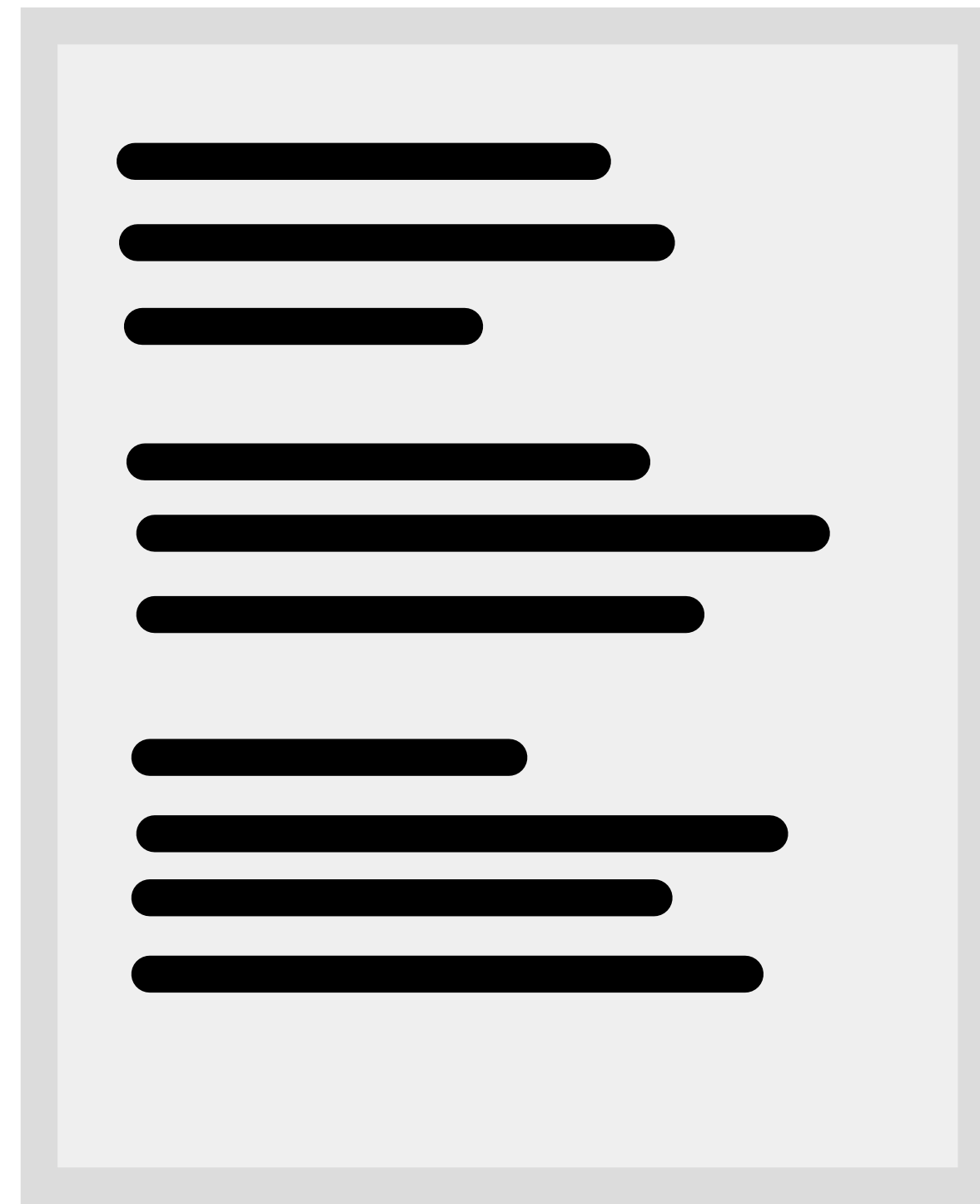
Antiviral drugs are approved for pregnant women and should be considered for children younger than **XX** years, although some are still being investigated.

Scientists have found some species of **unicorn** in Amazon rainforests.

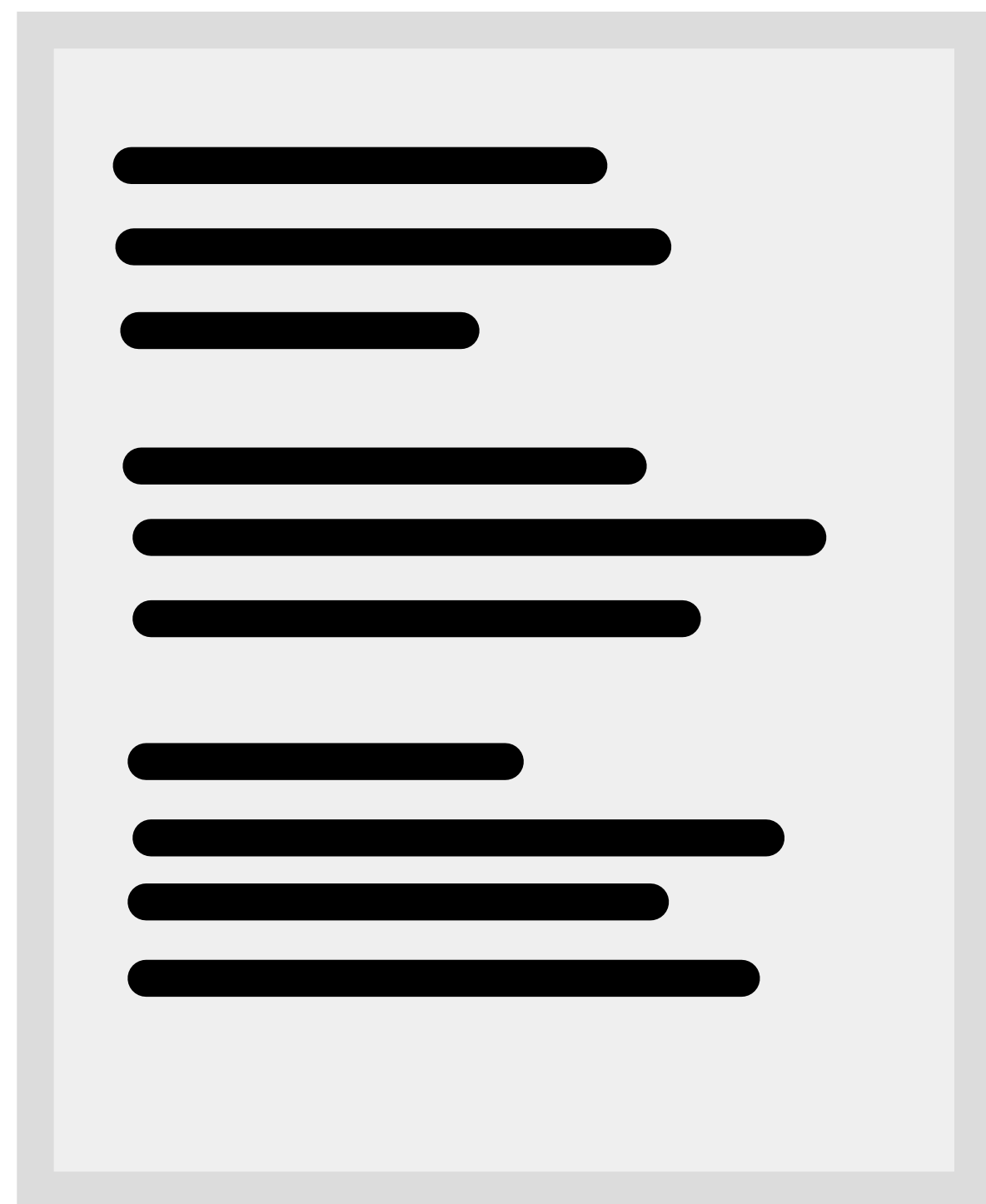




Source

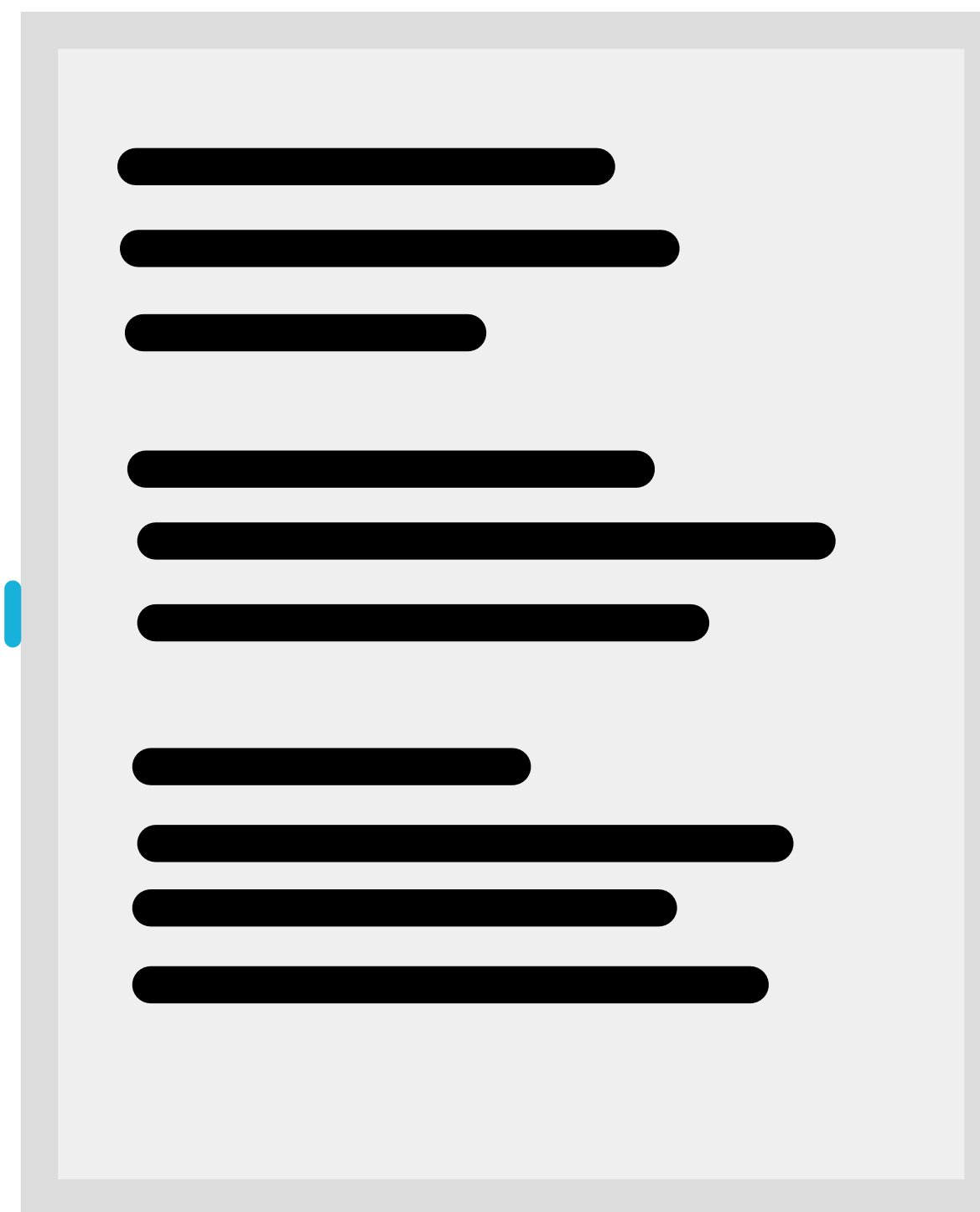


Target

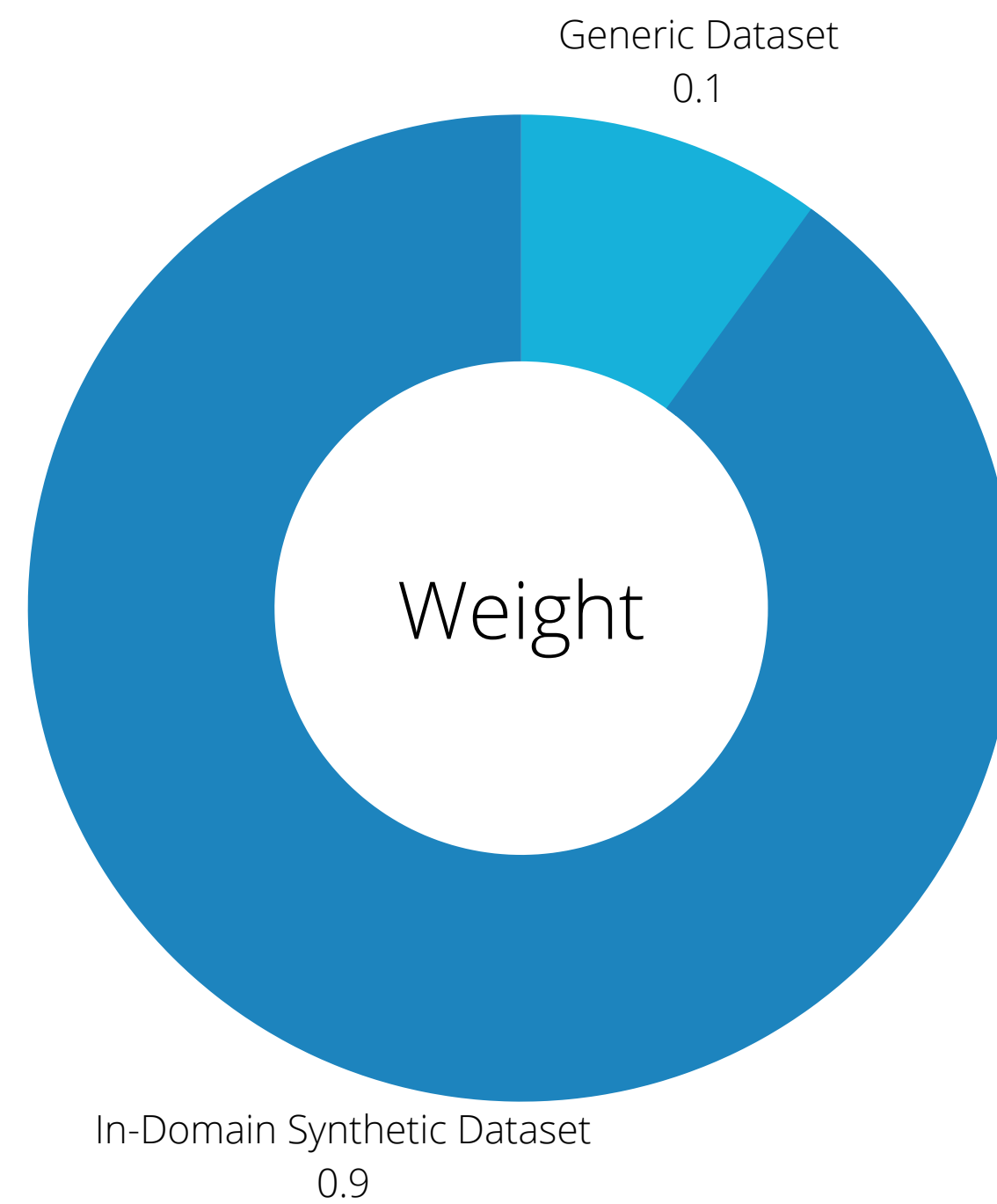
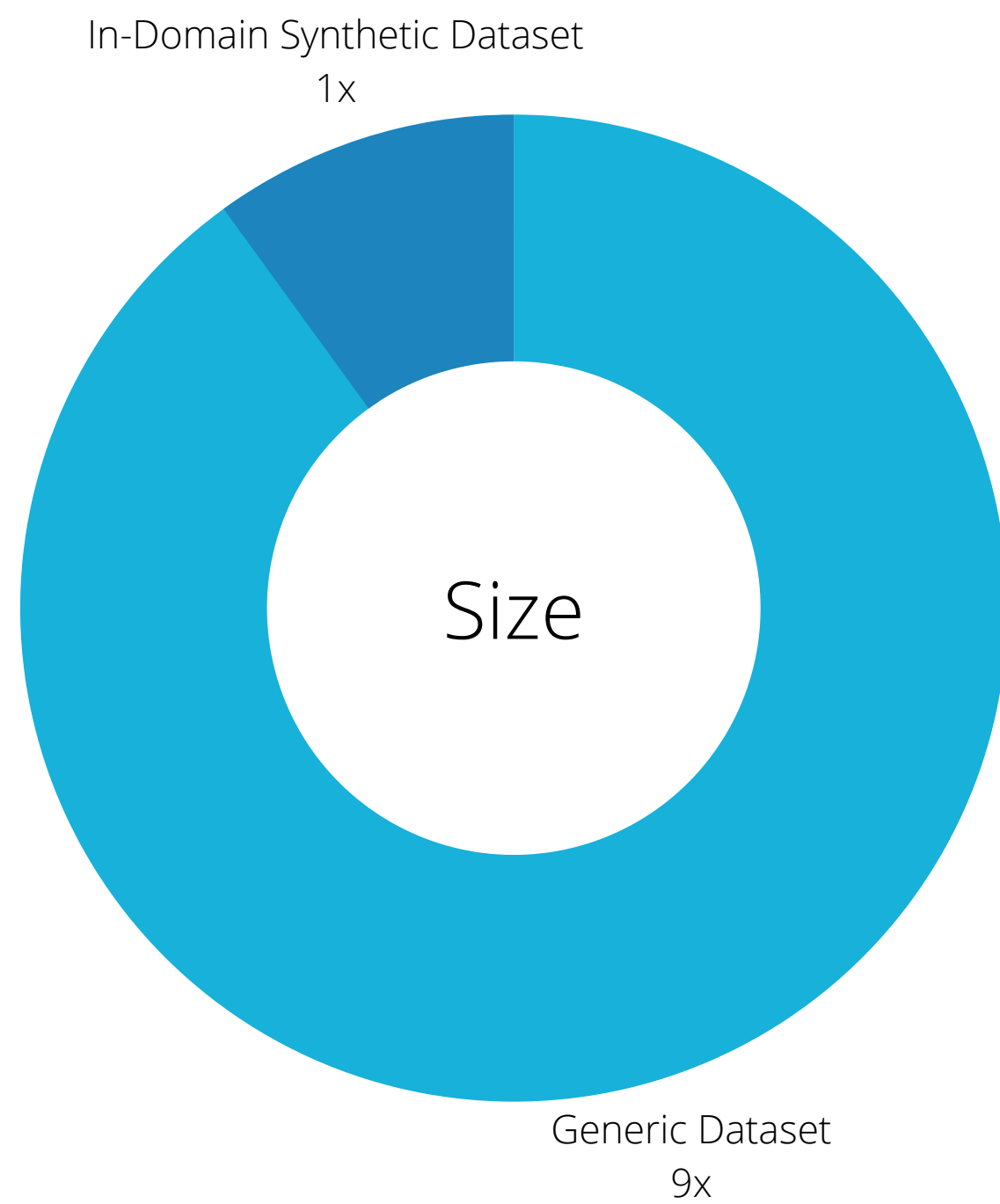


Source

Back-Translation



Target



Mixed Fine-tuning (Chu et al., 2017)

Language	Model	spBLEU ↑	chrF++ ↑	TER ↓	COMET ↑	Human ↑
AR-EN	Baseline	44.57	66.68	46.67	65.78	87.0
	Setup 1 Mixed Fine-Tuning	49.79	70.54	43.32	71.89	93.5
	Setup 2 Mixed Fine-Tuning	47.22	69.38	45.38	70.08	94.5
EN-AR	Baseline	36.15	58.3	58.29	57.5	87.0
	Setup 1 Mixed Fine-Tuning	42.38	62.52	53.99	67.48	90.0
	Setup 2 Mixed Fine-Tuning	37.91	59.42	55.95	59.47	88.5

Evaluation results on the in-domain test set, TICO-19

ymoslem/MT-LM

Scripts and config files for the paper, Domain-Specific Text Generation for Machine Translation



1

Contributor



0

Issues



3

Stars



0

Forks



ymoslem/MT-LM: Scripts and config files for the paper, Domain-Specific Text Generation for Machine Translation

Scripts and config files for the paper, Domain-Specific Text Generation for Machine Translation - GitHub - ymoslem/MT-LM: Scripts and config files for the paper, Domain-Specific Text Generation for...



Questions?

yasmin.moslem@adaptcentre.ie



Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

