Empowering Research with LLMs: Case Studies and Insights

Bernardo Leite
Artificial Intelligence and Computer Science Laboratory (LIACC)
Faculty of Engineering - University of Porto, Portugal (FEUP)
bernardo.leite@fe.up.pt | benjleite.com

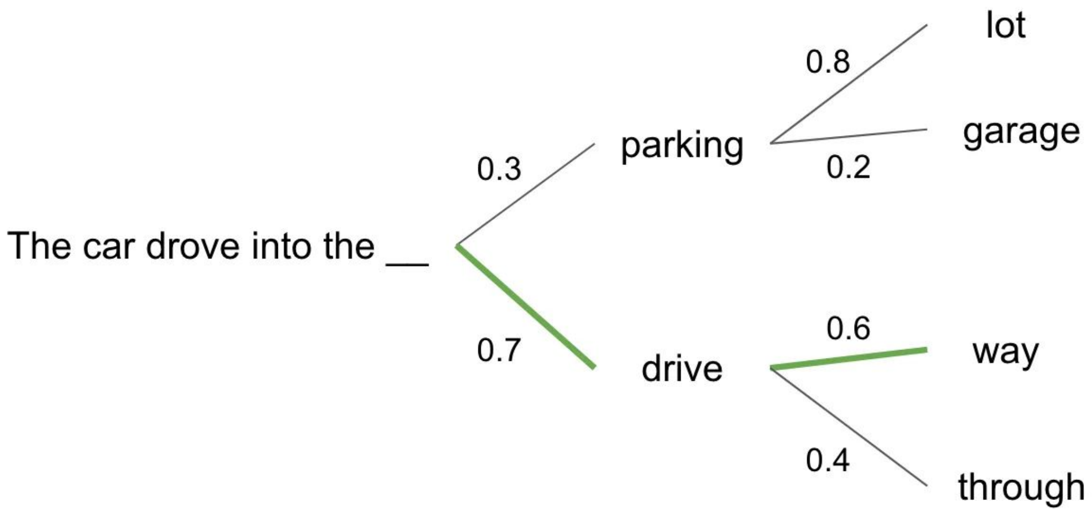
DeepLearn 2024
15 July, 2024, Porto - Maia, Portugal

https://www.aporia.com/learn/train-your-own-language-model/

Language Models

*Models that assign probabilities to upcoming words, or sequence of words, are called **language models** or **LMs**.*

Jurafsky, Dan, and James H. Martin. Speech and Language Processing. 3rd ed. draft, Feb 3, 2024.



https://towardsdatascience.com/despite-their-feats-large-language-models-still-havent-contributed-to-linguistics-657bea43a8a3

Language Models: Some Years Ago

GPT-1
2018

Improving Language Understanding by Generative Pre-Training

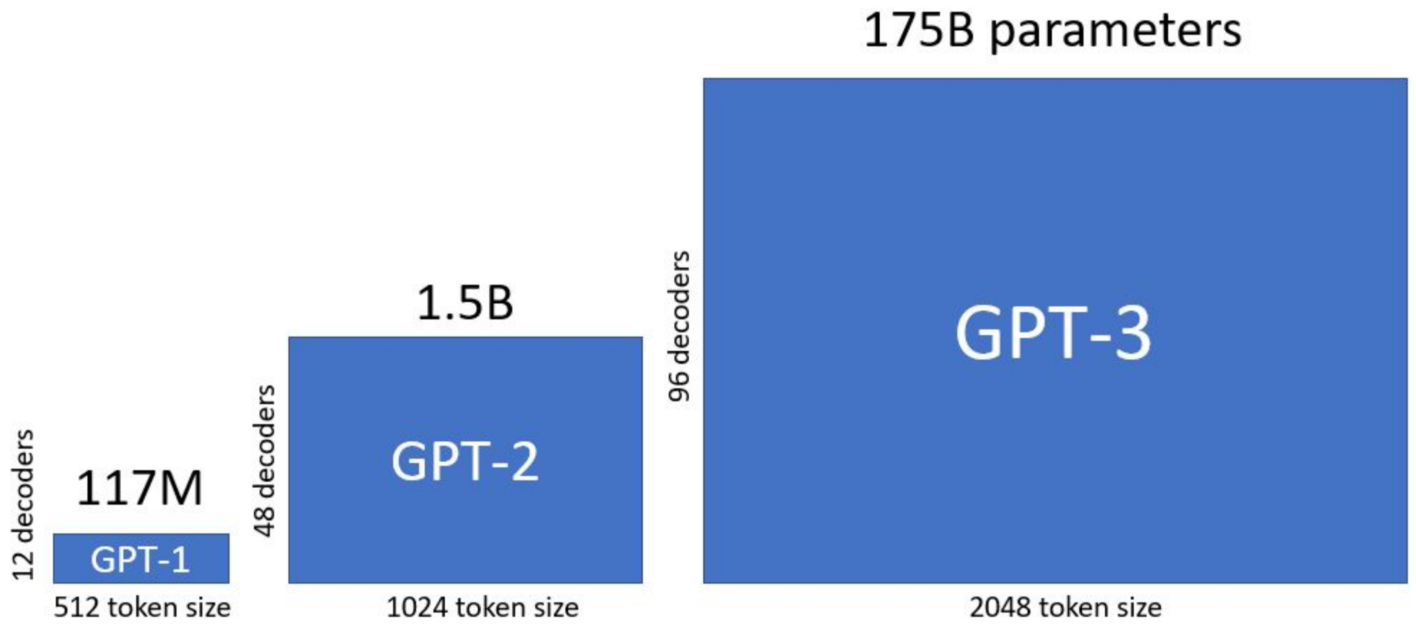
Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

GPT-2
2019

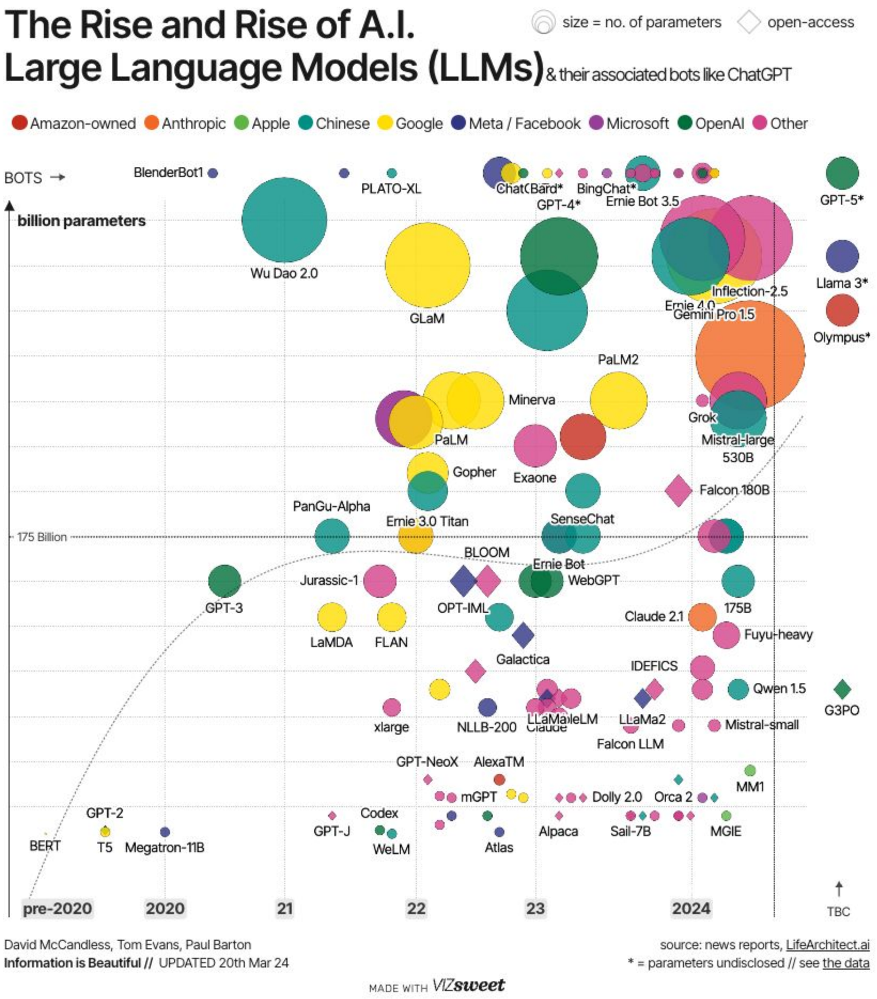
Language Models are Unsupervised Multitask Learners

Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

(*Large*) Language Models: How can we define *large*?



https://medium.com/@YanAIx/step-by-step-into-gpt-70bc4a5d8714



For this presentation:

- ***What is the impact of LLMs on our research?***

(Ongoing) Research Projects at LIACC

Albertina PT-*
(2023)

ADVANCING NEURAL ENCODING OF PORTUGUESE WITH TRANSFORMER ALBERTINA PT-*

João Rodrigues,[◊] Luís Gomes,[◊] João Silva,[◊] António Branco,[◊]
Rodrigo Santos,[◊] Henrique Lopes Cardoso,[♡] Tomás Osório[♡]
[◊]University of Lisbon
NLX – Natural Language and Speech Group, Dept of Informatics
Faculdade de Ciências (FCUL), Campo Grande, 1749-016 Lisboa, Portugal
[♡]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

(Main) Contributions:

- Foundation Encoder Model for **European** Portuguese
- Commendable Performance in Downstream Tasks
    e.g., similarity/inference tasks

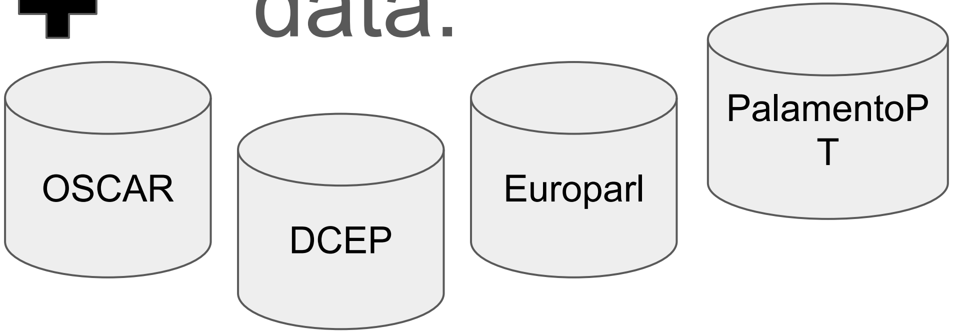
Albertina PT-* (2023)

ADVANCING NEURAL ENCODING OF PORTUGUESE WITH TRANSFORMER ALBERTINA PT-*

João Rodrigues,[◇] **Luís Gomes,**[◇] **João Silva,**[◇] **António Branco,**[◇]
Rodrigo Santos,[◇] **Henrique Lopes Cardoso,**[♡] **Tomás Osório**[♡]
[◇]University of Lisbon
NLX – Natural Language and Speech Group, Dept of Informatics
Faculdade de Ciências (FCUL), Campo Grande, 1749-016 Lisboa, Portugal
[♡]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

base model: <u>DeBERTa</u>

data:



Albertina PT-* (2023)

ADVANCING NEURAL ENCODING OF PORTUGUESE WITH TRANSFORMER ALBERTINA PT-*

João Rodrigues,[◇] Luís Gomes,[◇] João Silva,[◇] António Branco,[◇]
Rodrigo Santos,[◇] Henrique Lopes Cardoso,[♡] Tomás Osório[♡]
[◇]University of Lisbon
NLX – Natural Language and Speech Group, Dept of Informatics
Faculdade de Ciências (FCUL), Campo Grande, 1749-016 Lisboa, Portugal
[♡]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

(Some) Results:

	RTE	WNLI	MRPC	STS-B
Albertina PT-PT	**0.8339**	0.4225	**0.9171**	0.8801
Albertina PT-PT base	0.6787	0.4507	0.8829	0.8581
Albertina PT-BR	0.7942	0.4085	0.9048	**0.8847**
Albertina PT-BR base	0.6570	**0.5070**	0.8900	0.8516

Improved Albertina (2024)

Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family

Rodrigo Santos†, João Rodrigues†, Luís Gomes†, João Silva†, António Branco†, Henrique Lopes Cardoso‡, Tomás Freitas Osório‡, Bernardo Leite‡
†University of Lisbon
NLX - Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{rsdsantos, jarodrigues, luis.gomes, antonio.branco}@fc.ul.pt
‡University of Porto
Faculty of Engineering, Department of Informatics Engineering
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
hlc@fe.up.pt, tomas.s.osorio@gmail.com, bernardo.leite@fe.up.pt

(Main) Contributions:

- Improved Performance due to **model size increase + more data**
- **Fully Open** Encoder Models for Portuguese

Improved Albertina (2024)

Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family

Rodrigo Santos[†], João Rodrigues[†], Luís Gomes[†], João Silva[†], António Branco[†], Henrique Lopes Cardoso[‡], Tomás Freitas Osório[‡], Bernardo Leite[‡]
[†]University of Lisbon
NLX - Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{rsdsantos, jarodrigues, luis.gomes, antonio.branco}@fc.ul.pt
[‡]University of Porto
Faculty of Engineering, Department of Informatics Engineering
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
hlc@fe.up.pt, tomas.s.osorio@gmail.com, bernardo.leite@fe.up.pt

(Some) Results:

	GLUE				SuperGLUE			
model	RTE	WNLI	MRPC	STS-B	COPA	CB	MultiRC	BoolQ
Albertina 1.5B PTPT L	**0.8809**	0.4742	0.8457	**0.9034**	**0.8433**	**0.7840**	**0.7688**	**0.8602**
Albertina 1.5B PTPT S	0.8809	0.5493	0.8752	0.8795	0.8400	0.5832	0.6791	0.8496
Albertina 900M PTBR	0.8339	0.4225	**0.9171**	0.8801	0.7033	0.6018	0.6728	0.8224
Albertina 100M PTPT	0.6919	0.4742	0.8047	0.8590	n.a.	0.4529	0.6481	0.7578
DeBERTa 1.5B EN	0.8147	0.4554	0.8696	0.8557	0.5167	0.4901	0.6687	0.8347
DeBERTa 100M EN	0.6029	**0.5634**	0.7802	0.8320	n.a.	0.4698	0.6368	0.6829

PORTULAN EXTRAGLUE (2024)

PORTULAN EXTRAGLUE DATASETS AND MODELS: KICK-STARTING A BENCHMARK FOR THE NEURAL PROCESSING OF PORTUGUESE

Tomás Freitas Osório[†], Bernardo Leite[†], Henrique Lopes Cardoso[†], Luís Gomes[‡], João Rodrigues[‡], Rodrigo Santos[‡], António Branco[‡]

[†]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
Faculdade de Engenharia da Universidade do Porto
Rua Doutor Roberto Frias, s/n, Porto, Portugal
tomas.s.osorio@gmail.com, {bernardo.leite, hlc}@fe.up.pt

[‡]University of Lisbon
NLX—Natural Language and Speech Group, Dept of Informatics
Faculdade de Ciências (FCUL), Campo Grande, 1749-016 Lisboa, Portugal
{lmdgomes, jarodrigues, rsdsantos, antonio.branco}@fc.ul.pt

(Main) Contributions:

- Availability of a Portuguese Machine-Translated (Super) GLUE
- Benchmark on multiple tasks
  - e.g, question-answering

(Super) GLUE Tasks

Corpus	|Train|	|Dev|	|Test|	Task	Metrics	Text Sources
BoolQ	9427	3270	3245	QA	acc.	Google queries, Wikipedia
CB	250	57	250	NLI	acc./F1	various
COPA	400	100	500	QA	acc.	blogs, photography encyclopedia
MultiRC	5100	953	1800	QA	$F1_a$/EM	various
ReCoRD	101k	10k	10k	QA	F1/EM	news (CNN, Daily Mail)
RTE	2500	278	300	NLI	acc.	news, Wikipedia
WiC	6000	638	1400	WSD	acc.	WordNet, VerbNet, Wiktionary
WSC	554	104	146	coref.	acc.	fiction books

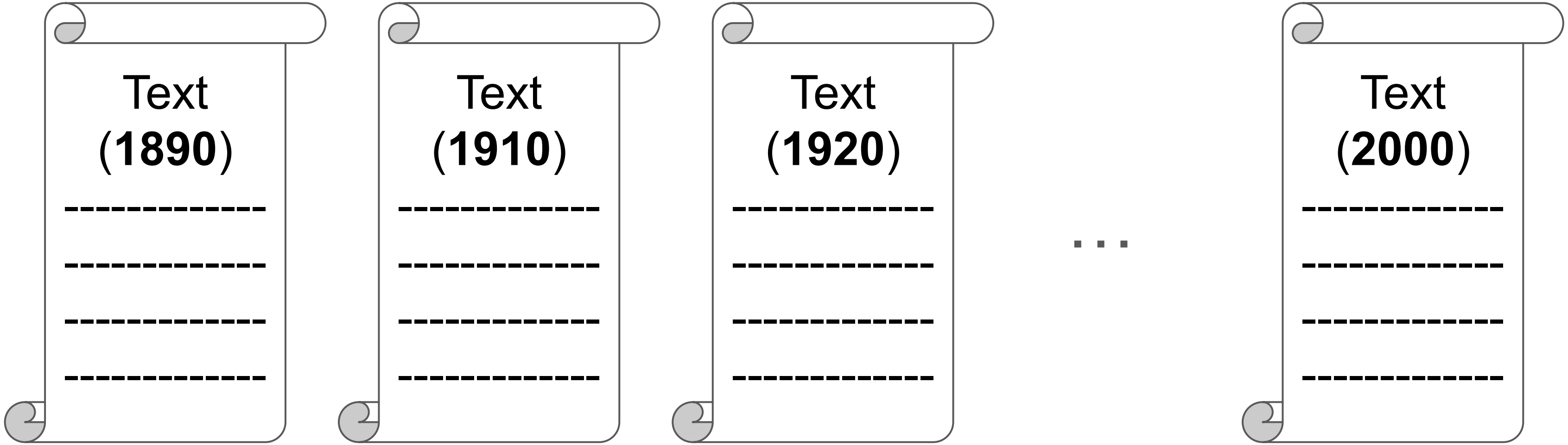
https://medium.com/nlplanet/two-minutes-nlp-superglue-tasks-and-2022-leaderboard-492d8c849ed

PORTULAN EXTRAGLUE (2024): Some Results

Fine-tuning using Low-Rank Adaptation (**LoRA**)

Task	Albertina 1.5B pt-PT	Albertina 1.5B pt-BR	>	XLM-RoBERTa-XL pt-PT	XLM-RoBERTa-XL pt-BR
Single sentence					
SST-2	0.9392	0.9450		0.9323	0.9392
Similarity					
MRPC	0.8969	0.9184		0.8696	0.8651
STS-B	0.8905	0.8940		0.8743	0.8734
Inference					
QNLI	0.9398	0.9361		0.9237	0.9237
RTE	0.7870	0.7978		0.6571	0.6606
WNLI	0.6197	0.6901		0.5634	0.5634
CB	0.8385	0.8554		0.6280	0.6160
QA					
BoolQ	0.7456	0.7807		0.6538	0.6587
MultiRC	0.7257	0.7169		0.6926	0.6925
Reasoning					
CoPA	0.8500	0.8200		0.5000	0.5600

LLMs for Portuguese Historical Data (***Tomás Osório***)



Machine Translation for Emakhuwa – A Bantu language spoken in Mozambique (***Felermino Ali***)

https://www.diplomaciabusiness.com/culturalmente-diverso-mocambique-celebra-hoje-25-sua-independencia/

Mozambique language panorama:

- ~ 32 million of people, 20 Bantu languages, 1 official language, bilingual-education
- Most spoken languages: **Emakhuwa (25%)**, Portuguese (official 10.8%), Xichangana/Tsonga (10.5%)
- Emakhuwa: 8 variants

Not much data 🙁

ChatGPT does not “speak” Emakhuwa 🙁

Machine Translation for Emakhuwa – A Bantu language spoken in Mozambique (***Felermino Ali***)

https://www.diplomaciabusiness.com/culturalmente-diverso-mocambique-celebra-hoje-25-sua-independencia/

(Some) Research Questions:

- *How does **different source of data** affect Portuguese-Emakhuwa machine-translation quality?*
- *How can we **create synthetic data** for Portuguese-Emakhuwa machine-translation?*
- *How to build **robust models** resilient to **spelling variations** in Emakhuwa?*

Machine Translation for Emakhuwa – A Bantu language spoken in Mozambique (***Felermino Ali***)

Train Data		Our Test set		Flores	
	# sent.	*pt→vmw*	*vmw→pt*	*pt→vmw*	*vmw→pt*
Ali et al. + News Trans $_{baseline}$	~63k	11.58 / 45.62	**22.90 / 46.65**	6.85 / 38.05	17.01 / 42.34
Ali et al. + News Trans + OCR-ed	~65k	29.65 / 66.05	23.14 / 46.47	24.79 / 62.40	**18.93 / 43.25**
Ali et al. + News Trans + BT	~78k	**40.90 / 74.04**	22.42 / 45.92	**39.23 / 73.88**	18.68 / 42.93
All	~80k	31.97 / 68.35	22.22 / 46.25	28.77 / 66.78	18.47 / 43.15

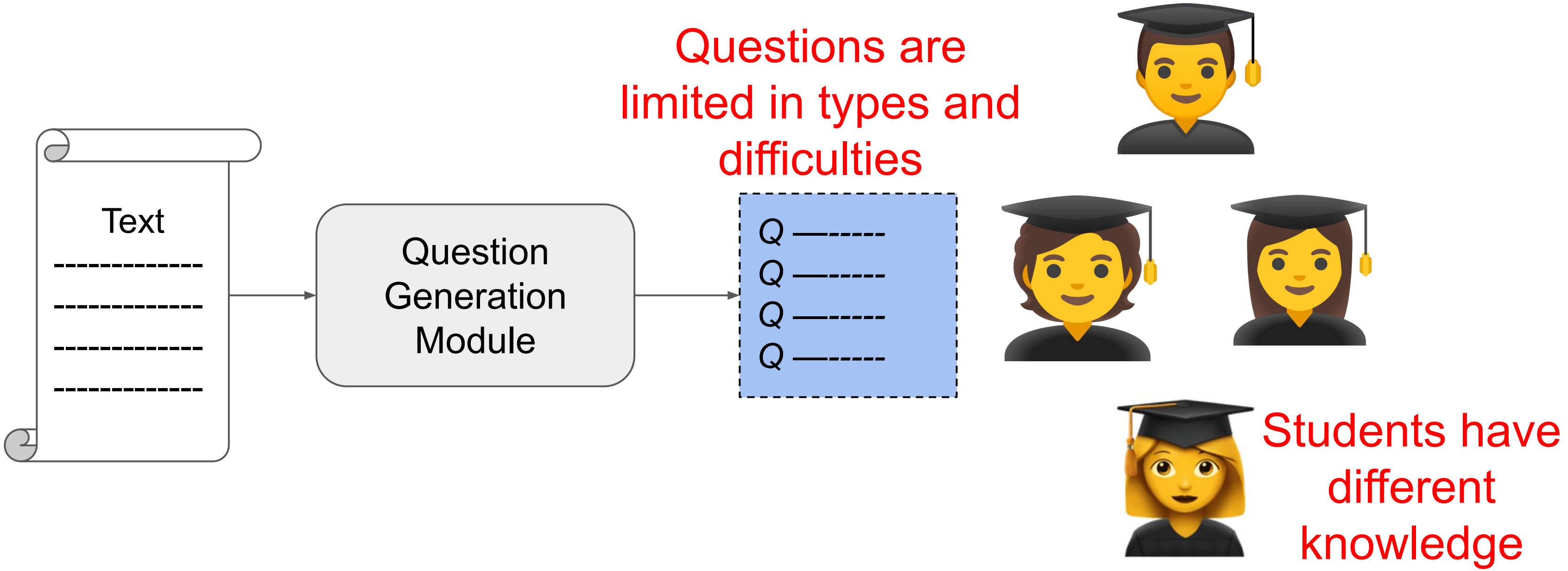
[Ali et al., 2024] (in revision)

Early Contributions

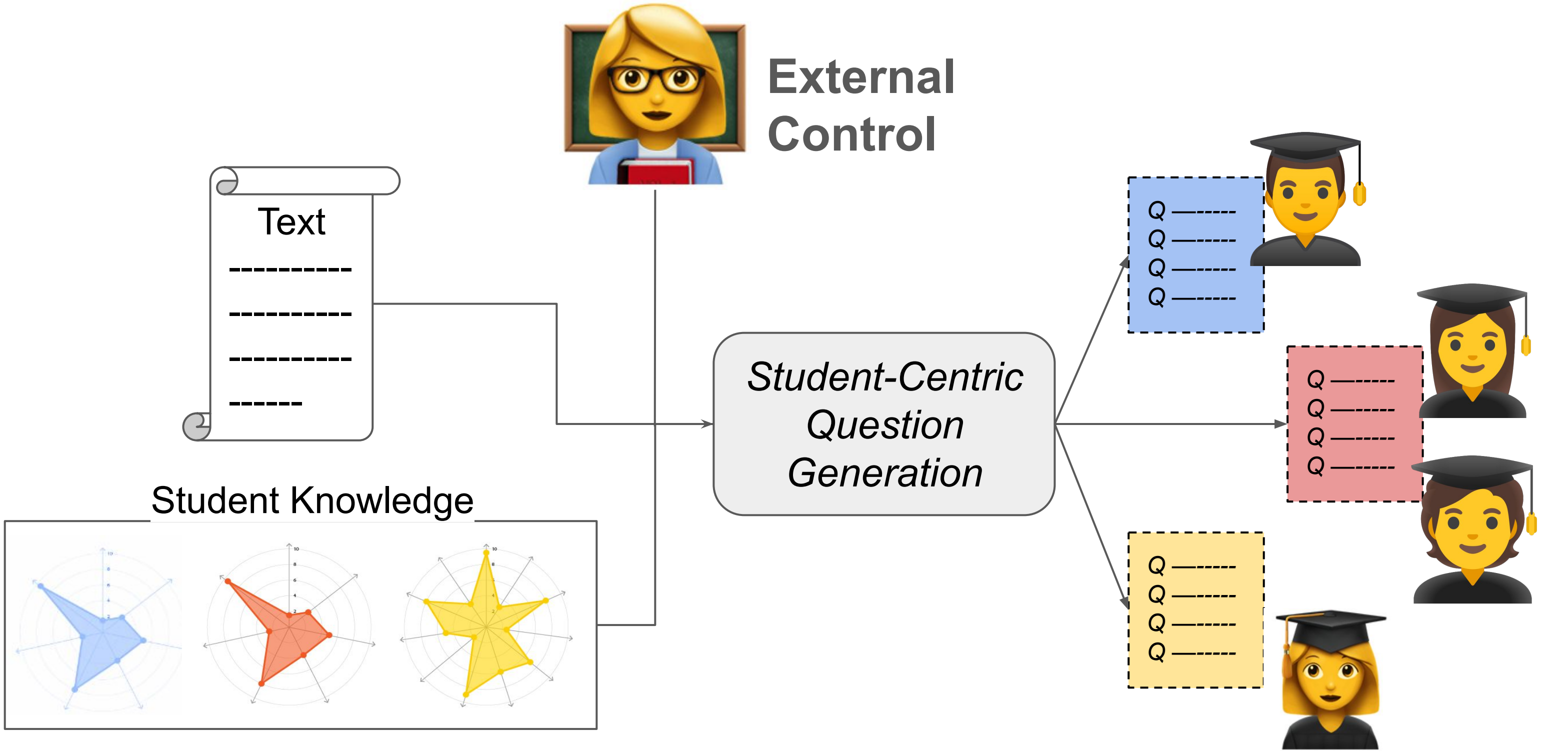
Model	*pt→vmw*	*vmw→pt*
Transformer-baseline	5.94 / 32.20	10.03 / 34.14
byt5	11.32 / 43.49	20.04 / 43.31
afri-byt5	10.13 / 41.13	19.86 / 43.01
mt5	4.47 / 32.48	9.83 / 40.07
mT0	5.99 / 33.61	14.61 / 36.50
afri-mt5	6.23 / 36.83	14.92 / 39.07
M2M-100	10.92 / 44.23	20.62 / 44.11
NLLB	**11.58 / 45.62**	**22.90 / 46.65**

[Ali et al., 2024] (in revision)

Student-Centric Question Generation (***Bernardo Leite***)



Student-Centric Question Generation (***Bernardo Leite***)



*How have LLMs **actually helped** me so far?*

😊

How have LLMs ***actually helped*** me so far?
➔ Quality of generated questions.

	QA Pairs Provenance		
Metric	**Real-Exam**	**GPT-4**	**QAPG**
Well-formedness	20/0	19/1	20/0
Relevance with Text	20/0	20/0	19/1
Answerability	20/0	20/0	14/6
Answer Alignment	18/2	20/0	8/12
Children Suitability	4,77	4,83	4,68

[Leite et al., 2024] (in publishing)

How have LLMs ***actually helped*** me so far?
➔ Additional (synthetic) data.

(...+3 prompt examples...)

Text: But the second son spoke most sensibly too, and said: ’Whatever I give to you I deprive myself of. Just go your own way, will you?’ Not long after his punishment overtook him, for no sooner had he struck a couple of blows on a tree with his axe, than he cut his leg so badly that he had to be carried home.
Question: *What happened to the second son?*
Answer: *He cut his leg so badly that he had to be carried home.*

[Leite and Cardoso, 2024]

*How have LLMs **<u>not helped</u>** me so far?*

🙁

How have LLMs ***not helped*** me so far?

- ➔ Accessibility and Cost
- ➔ Resource Intensive
- ➔ The division of tasks with LLMs:

Por que razão o urso disse aos coelhos que não tinha nenhum mel?

(a) Porque não queria ser incomodado.
(b) Porque não queria emprestar nada.
(c) Porque não queria ficar sem mel.

How have LLMs ***not helped*** me so far?

➔ Limited Customization **for Controlling Question Difficulty:**

Difficulty	Fine-tuning (LLaMA 2)	Few-shot (GPT-4)
-2.0	0.73	0.43
0.0	0.55	0.43
2.0	0.39	0.45

[Tomikawa and Uto, 2024] (In Publishing) (shared earlier with courtesy of the authors)

"...<u>Few-shot learning might be insufficient for controlling difficulty</u> and that fine-tuning with a substantial amount of data may be necessary."

Before finishing… 🤔❓

Should we prioritize…

fast, smaller, and more resource-efficient models?

Or opt for larger, smarter, but slower models?

Or maybe run large models faster?

Thank you!

Bernardo Leite
Artificial Intelligence and Computer Science Laboratory (LIACC)
Faculty of Engineering - University of Porto, Portugal (FEUP)
bernardo.leite@fe.up.pt | benjleite.com

DeepLearn 2024
15 July, 2024, Porto - Maia, Portugal