Bridging the Bosphorus: Advancing Turkish Large Language Models through Strategies for Low-Resource Language Adaptation and Benchmarking

Website

Emre Can Acikgoz[1,2], Mete Erdoğan[1,2], Deniz Yuret[1,2]

[1] Koç University, KUIS AI Center, [2] Koç University, Department of Computer Engineering

Contributions

This work explores the challenges faced by low-resource languages, with a **special focus on Turkish**. Our contributions are as follows:

- Release the **Hamza LLM series**, encompassing models from 124M to 1.3B parameters. Notably, Hamza-xl with 1.3B trained on 300B tokens.
- Our analysis explores two distinct methodologies for developing Turkish LLMs: (i) **extending pretrained models** (Mistral-7b and GPT2-xl) with Turkish-only data, and (ii) **constructing a model from scratch**, similar to the GPT2 approach.
- We have established a meticulously cleaned and novel **Turkish LLM evaluation benchmark**.

Pretraining Dataset

We trained using the Turkish subset of CulturaX, comprising 128 documents totaling 180GB and 130B unique tokens determined by using the GPT-2 tokenizer.

Corpus	Documents	Ratio	# of Tokens
mC4	75,859,899	80.52%	104.3 B
OSCAR-2019	5,867,831	6.23%	8.1 B
OSCAR-2109	6,614,512	7.02%	9.1 B
OSCAR-2201	2,580,896	2.74%	3.5 B
OSCAR-2301	3,284,322	3.49%	4.5 B
CulturaX (total)	94,207,460	100.0%	129.5 B

Table: **Statistics of the pretraining dataset.**

Method 1: Further Training a Base Model

We aim to further training state-of-the-art models on Turkish data, which was initially unfamiliar with Turkish (i.e., not trained on Turkish data):

- **Selecting the Base-Models:** We have selected Mistral and we opted for GPT2-xlarge since our Hamza models are trained following GPT2 architecture.
- **Dataset:** To integrate Turkish knowledge into Mistral and GPT-2, we initially pretrained on 100MB of Turkish data and gradually increased it up to 5GB.
- **Training:** We used LoRA adapters with $r = 32$, $\alpha = 32$, 0.05 dropout, AdamW optimizer, 0.0001 learning rate with batch size of 1.

Method 2: Pretraining from Scratch

- **Dataset:** Our pretraining data contains 128 parquet files each 1.4GB, totaling almost 179.2GB with 129,486,207,634 (130B) training tokens.
- **Architecture:** Our approach led to the creation of four variants of Hamza, following GPT-2: hamza-small, hamza-medium, hamza-large, and our largest model, hamza-xlarge.
- **Architecture:** We trained using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$), a cosine learning rate schedule to reduce to 10% maximum, a 0.1 weight decay, and limited gradient norms to 1.0 to avoid overfitting. Our initial setup included 2,000 warm-up steps with global batch sizes of 491,520.
- **Training:** All models were trained on 300B tokens with a uniform 500K batch size and a 1024-token context window. We fine-tuned the learning rate for each variant, used half-precision (fp16), and employed tensor and data parallelism on eight A100 GPUs with 80GB each, without any dropout.

Model	Parameters	Layers	Heads	d$_{model}$	Learning Rate	Batch Size	Tokens
hamza-small	124M	12	12	768	6.0e^{-4}	0.5M	300B
hamza-medium	354M	24	16	1024	3.0e^{-4}	0.5M	300B
hamza-large	772M	36	20	1280	3.0e^{-4}	0.5M	300B
hamza-xlarge	1.3B	24	16	2048	2.0e^{-4}	0.5M	300B

Table: **Architecture and optimization hyperparameters for the four Hamza model sizes that trained from scratch.**

Turkish LLMs Leaderboard

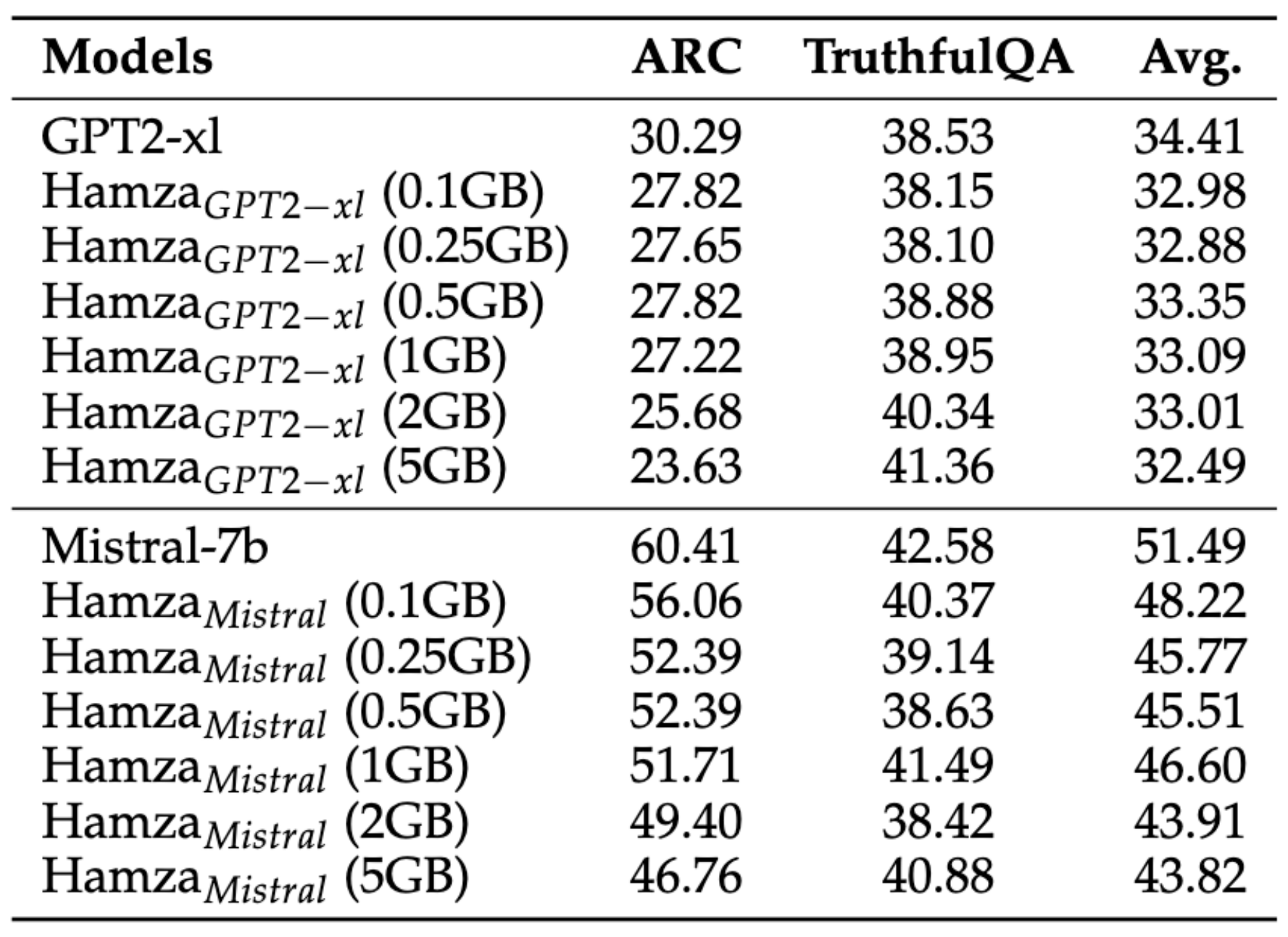
We compare models across different metrics: zero and few-shot accuracies on ARC-TR and TruthfulQA-TR, and Bits-Per-Character (BPC) on TRNews-64 corpus.

Type	Models	Accuracy (↑)		BPC (↓)
		ARC-TR	TruthfulQA-TR	trnews-64
Base & SFT Models	LLaMA2 7b	25.94	41.18	1.374
	LLaMA3 8b	43.09	44.77	0.929
	Mistral 7b	32.68	41.16	1.260
	Gemma 7B	**46.16**	42.35	0.989
	GPT2-xl	24.91	40.97	2.533
	LLaMA2 7b-chat	25.00	40.07	1.374
	Mistral 7b-chat-v2	35.24	48.34	1.428
Multi-lingual Models	XGLM-7.5B	29.01	39.09	0.880
	XGLM-4.5B	25.94	40.18	0.949
	XGLM-564M	23.55	42.59	1.125
	mGPT	26.54	42.37	1.306
Huggingface Turkish Models	Kanarya-2b	29.78	41.43	**0.724**
	Kanarya-750m	28.16	41.50	0.767
	Turkcell-LLM-7b-v1	43.09	44.91	1.208
	ytu-gpt2-large	27.13	43.09	0.805
	Trendyol-7b-chat	35.58	44.35	0.820
	Trendyol-7b-dpo	39.93	**50.11**	0.859
	Commencis-LLM	33.28	44.50	1.306
	Sambalingo-tr	44.37	46.61	0.894
	Thestral-tr-chat	34.00	41.90	1.314
	Mistral-7b-chat-v2-tr	33.96	45.71	1.411
Our Models	Hamza-xl	28.24	42.33	0.754
	Hamza$_{GPT2-xl}$	24.74	44.95	1.152
	Hamza$_{Mistral}$	39.85	46.40	0.816

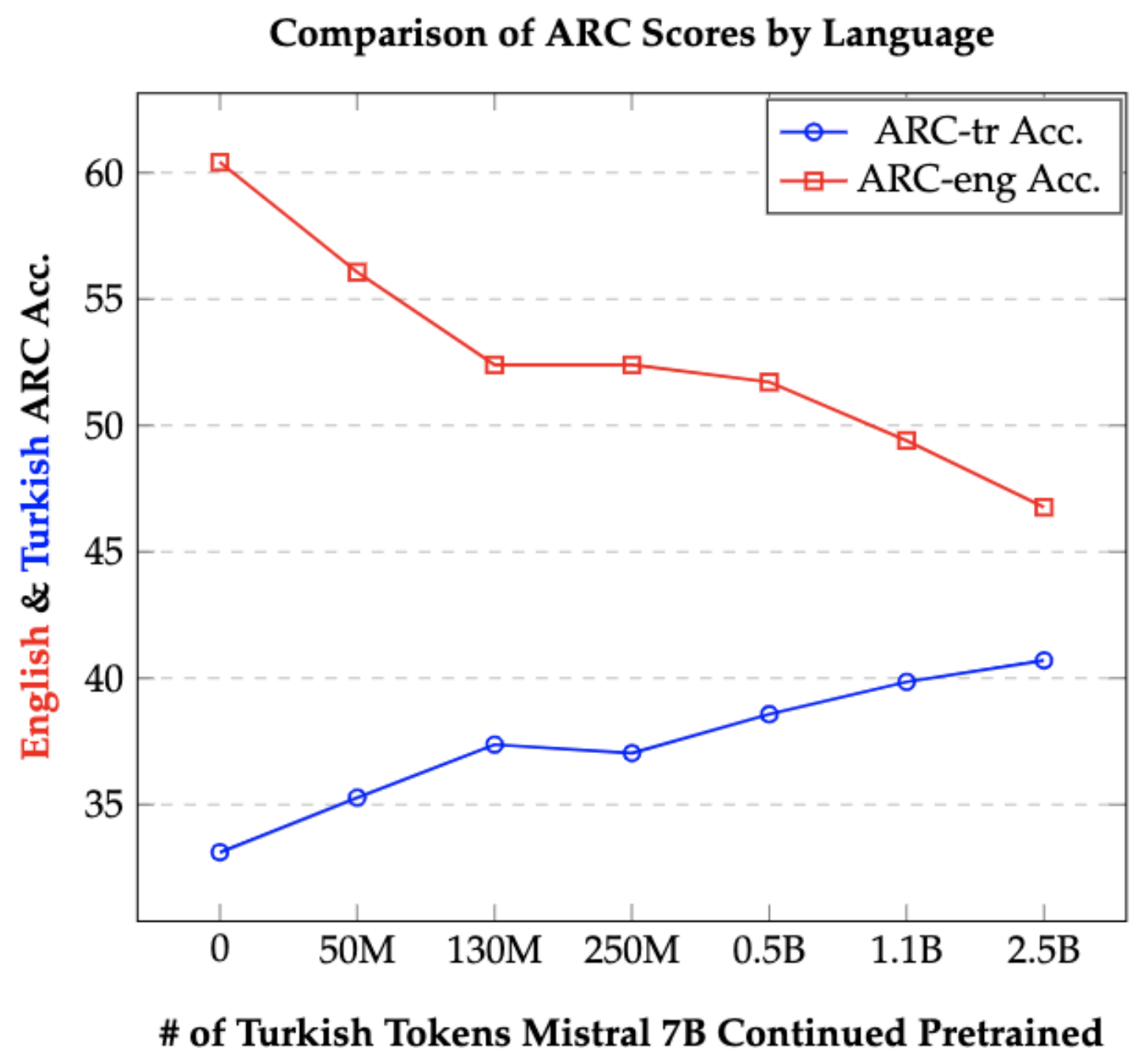
Table: **Performance comparison on various Turkish tasks.**

Retention after Fine-Tuning: Will Models Forget English-Learned Skills in Another Language?

Continued pretraining of models like Mistral on Turkish reduces their accuracy in the original language, showing **catastrophic forgetting**.



Models	ARC	TruthfulQA	Avg.
GPT2-xl	30.29	38.53	34.41
Hamza$_{GPT2-xl}$ (0.1GB)	27.82	38.15	32.98
Hamza$_{GPT2-xl}$ (0.25GB)	27.65	38.10	32.88
Hamza$_{GPT2-xl}$ (0.5GB)	27.82	38.88	33.35
Hamza$_{GPT2-xl}$ (1GB)	27.22	38.95	33.09
Hamza$_{GPT2-xl}$ (2GB)	25.68	40.34	33.01
Hamza$_{GPT2-xl}$ (5GB)	23.63	41.36	32.49
Mistral-7b	60.41	42.58	51.49
Hamza$_{Mistral}$ (0.1GB)	56.06	40.37	48.22
Hamza$_{Mistral}$ (0.25GB)	52.39	39.14	45.77
Hamza$_{Mistral}$ (0.5GB)	52.39	38.63	45.51
Hamza$_{Mistral}$ (1GB)	51.71	41.49	46.60
Hamza$_{Mistral}$ (2GB)	49.40	38.42	43.91
Hamza$_{Mistral}$ (5GB)	46.76	40.88	43.82



Accuracy comparison of Continued Pretrained models on English (Left, Right) and Turkish (Right) question answering tasks.

Hardware Details

Model	Trained Parameters	GPU Type	GPU Count	Training Hours
Hamza-small	124M	A100 (80GB)	8	72
Hamza-medium	354M	A100 (80GB)	8	201
Hamza-large	772M	A100 (80GB)	8	378
Hamza-xlarge	1.3B	A100 (80GB)	8	460
Hamza$_{GPT2-xl}$	17M	A40 (48GB)	1	334
Hamza$_{Mistral}$	57M	A40 (48GB)	1	501

Table: **Device Overview of hamza Model Configurations.**