BertaQA: How Much Do Language Models Know About Local Culture?

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle and Mikel Artetxe
HiTZ Center, University of the Basque Country UPV/EHU

GitHub

https://github.com/juletx/BertaQA

1. Introduction

- LLMs have extensive knowledge about the world, but most evaluations have been limited to global or anglocentric subjects.
- How well do these models perform on topics relevant to other cultures, whose presence on the web is not that prominent?
- To address this gap, we introduce BertaQA, a multiple-choice trivia dataset

2. BertaQA

- First dataset with annotated questions related to the Basque and global cultures.
- 4756 parallel examples in Basque and English.
- Divided into 8 categories and 3 difficulties.
- Local subset with questions pertinent to the Basque culture
- Global subset with questions of broader interest.

	Local Questions	Global Questions
Basque and Literature	What does the "Karmel" magazine specialize in? a) Bertsolarism **b) Basque culture in the past and the present** c) The life of the Carmelites	In which of these novels does the sea not appear? **a) "The Adventures of Tom Sawyer"** b) "Moby Dick" c) "Treasure Island"
Geography and History	Where's Atxondo? **a) In Biscay** b) In Gipuzkoa c) In Navarre	Who was imprisoned in 1964? **a) Nelson Mandela** b) Mumia Abu Jamal c) Charles Ghankay

3. Main results in English

- Open and commercial models much worse in local than global.
- Bigger difference for open models.
- Performance on local and global correlated.
- Scaling differences open vs closed.
- Easier to improve on global questions.
- But this subset starts saturating for the strongest models.
- Resulting in bigger improvements on the local subset.

Model	Variant	Local	Global	Δ
Random	N/A	33.33	33.33	0.00
GPT	3.5 Turbo	55.08	82.40	27.32
	4	69.88	91.43	21.55
	4 Turbo	**72.17**	**91.68**	**19.51**
Claude 3	Haiku	58.71	84.16	25.45
	Sonnet	58.33	86.41	28.08
	Opus	**71.91**	**91.85**	**19.94**
Llama 2	7B	41.54	64.34	**22.80**
	13B	43.61	70.36	26.75
	70B	**49.15**	**77.68**	28.53
Llama 3	8B	50.38	76.63	26.25
	70B	**59.56**	**84.74**	**25.18**
Qwen 1.5	7B	42.51	71.45	**28.94**
	14B	44.67	75.92	31.25
	72B	**54.70**	**83.99**	29.29
Yi	6B	44.25	73.20	**28.95**
	9B	43.87	75.00	31.13
	34B	**54.06**	**83.61**	29.55
Mistral	7B	47.50	74.16	26.66
	47B	**57.40**	**82.78**	**25.38**
Gemma	7B	**45.69**	**76.42**	**30.73**
Average	N/A	53.25	79.91	26.66

4. Local knowledge transfer from Basque to English

- Local models trained with continued pretraining in Basque improve on local questions in English.
- Local models become worse on global questions.
- Bigger degradation and smaller improvement for the smallest model.
- Previous conclusions uncomplete, challenges curse of multilinguality.

Model	Local	Global	Δ
Llama 2 7B	41.54	**64.34**	22.80
+ *eu train*	**47.72**	53.26	**5.54**
Llama 2 13B	43.61	**70.36**	26.75
+ *eu train*	**56.60**	67.47	**10.87**
Llama 2 70B	49.15	**77.68**	28.53
+ *eu train*	**62.61**	73.62	**11.01**

5. Comparison of English and Basque

- Worse results in Basque for most models
- Local models better at answering local questions in Basque and global questions in English.
- Knowledge transfer is not perfect across languages.
- Local and global knowledge not transferred completely.

Model	Variant	Local	Global	Δ
Random	N/A	33.33	33.33	0.00
GPT	3.5 Turbo	47.25 (-7.83)	66.22 (-16.18)	**18.97**
	4	62.94 (-6.94)	85.91 (-5.52)	22.97
	4 Turbo	**69.46** (-2.71)	**89.21** (-2.47)	19.75
Claude 3	Haiku	58.21 (-0.50)	79.85 (-4.31)	21.64
	Sonnet	56.13 (-2.20)	83.24 (-3.17)	27.11
	Opus	**71.32** (-0.59)	**90.89** (-0.96)	**19.57**
Llama 2	7B	34.90 (-6.64)	37.08 (-27.26)	**2.18**
	13B	34.09 (-9.52)	43.77 (-26.59)	9.68
	70B	**37.39** (-11.76)	**54.22** (-23.46)	16.83
Llama 2 *+ eu train*	7B	49.45 (+1.73)	50.79 (-2.47)	**1.34**
	13B	60.24 (+3.64)	65.47 (-2.00)	5.23
	70B	**64.85** (+2.24)	**72.24** (-1.38)	7.39
Llama 3	8B	42.60 (-7.78)	63.09 (-13.54)	**20.49**
	70B	**57.40** (-2.16)	**82.15** (-2.59)	24.75
Qwen 1.5	7B	35.96 (-6.55)	46.15 (-25.30)	**10.19**
	14B	37.31 (-7.36)	53.39 (-22.53)	16.08
	72B	**42.77** (-11.93)	**63.25** (-20.74)	20.48
Yi	6B	37.94 (-10.32)	46.45 (-22.99)	**8.51**
	9B	38.20 (-13.79)	49.21 (-21.70)	11.01
	34B	**41.03** (-6.31)	**60.41** (-26.75)	19.38
Mistral	7B	37.18 (-5.67)	51.17 (-25.79)	**13.99**
	47B	**43.61** (-13.03)	**61.08** (-23.20)	17.47
Gemma	7B	**41.84** (-3.85)	**65.89** (-10.53)	**24.05**
Average	N/A	47.92 (-5.64)	63.53 (-14.41)	15.61

6. Translate-test and self-translate

- For Llama 2, translate-test improves results when compared to Basque. Self-translate does not provide big improvements.
- For local models, translate-test worsens results. Self-translate is better than translate-test, still does not reach Basque.
- For Gemma, translation improves global and harms local a bit.
- Overall, translation is better for global, does not work well on local.

Model	Size	Method	Local	Global
Llama 2	7B	Translate-test	**37.44** (+2.54)	**55.35** (+18.27)
		Self-translate	33.80 (-1.10)	38.71 (+1.63)
	13B	Translate-test	**37.69** (+3.60)	**62.50** (+18.73)
		Self-translate	34.81 (+0.72)	46.11 (+2.34)
	70B	Translate-test	**42.68** (+5.29)	**71.03** (+16.81)
		Self-translate	39.85 (+2.46)	55.23 (+1.01)
Llama 2 *+ eu train*	7B	Translate-test	35.79 (-13.66)	44.27 (-6.52)
		Self-translate	**44.37** (-5.08)	**50.04** (-0.75)
	13B	Translate-test	41.79 (-18.45)	59.36 (-6.11)
		Self-translate	**56.13** (-4.11)	**65.55** (+0.08)
	70B	Translate-test	46.28 (-18.57)	65.47 (-6.77)
		Self-translate	**60.15** (-4.70)	**70.48** (-1.76)
Gemma	7B	Translate-test	**41.67** (-0.17)	**69.19** (+3.30)
		Self-translate	**41.67** (-0.17)	67.68 (+1.79)

Thank you!

julen.etxaniz@ehu.eus

https://julenetxaniz.eus/en

@juletxara