Beyond Art Masterpieces or Touristic Clichés: assessing LLMs for cultural alignment

Ciências ULisboa — Faculdade de Ciências da Universidade de Lisboa
UNIVERSIDADE BEIRA INTERIOR

António Branco, João Silva, Nuno Marques, Luis Gomes, Ricardo Campos, Raquel Sequeira, Sara Nerea, Rodrigo Silva, Miguel Marques, Rodrigo Duarte, Artur Putyato, Diogo Folques, Tiago Valente

Motivation

Mainstream approaches

- focus mostly on historical events or literate culture highlights
- echo stereotypical clichés, external to the social group at stake

As these belong also to universal or global culture: they fail to discriminate between models that may happen to be more or less culturally aligned

Our approach

- human annotators, native speakers, raised and living in that culture, fully immersed in and familiar with it

- questions posed naturally, as in casual conversations, relating to matters felt culturally specific

- without indicating that it pertains to Portugal or Portuguese culture, and avoiding proper names

Tuguesice-PT benchmark

Development

- 9 annotators
- 1054 QA pairs proposed
- 21 domains
- 2 adjudicators
- 500 QA pairs adjudicated

Guidelines

Linguistic constraints
L1 - Complexity of the question
L2 - Complexity of the answer
L3 - Factuality
L4 - Single answer
L5 - Correctness
L6 - Context independence
L7 - Not yes/no question

Endogenous point-of-view
E1 - Language
E2 - Scope
E3 - Knowledge level
E4 - Domains
E5 - Temporal horizon

Discriminative power
D1 - Pragmatic context
D2 - Bigtech chatbot failure
D3 - Portuguese-speaking cultures

Revealing intrinsic cultural alignment

Our approach is much better at revealing the intrinsic cultural alignment of models

Experiment with Tuguesice-PT

plain prompt
Asks the model to answer the question that follows

oracle prompt
Asks the model to answer the question that follows assuming the context of Portuguese culture

model	Tuguesice-PT plain	Tuguesice-PT oracle	Tuguesice-PT Δ	BLEnD-PT plain	BLEnD-PT oracle	BLEnD-PT Δ
gemini-2.5-flash	35.78	77.68	42	55.17	54.74	0
gervasio70b	39.76	60.86	21	51.72	53.45	2
llama70b	25.69	63.30	38	50.00	51.29	1
mistral24b	15.60	57.19	42	43.97	46.98	3
gervasio8b	11.31	38.53	27	41.81	42.67	1
llama8b	9.79	40.37	31	40.52	43.97	3
sabia7b	7.95	27.83	20	19.83	31.03	11

PRR Plano de Recuperação e Resiliência · República Portuguesa · Financiado pela União Europeia NextGenerationEU · 2030 Lisboa Programa Regional de Lisboa · Portugal 2030 · Cofinanciado pela União Europeia · fct Fundação para a Ciência e a Tecnologia