/ COLING 2025
/ Awards Session
/ Friday, January 24th, 2025

VeritasQA: A Truthfulness Benchmark Aimed at Multilingual Transferability

Javier Aula-Blasco[1] **Júlia Falcão**[1] Susana Sotelo[2]
Silvia Paniagua Suárez[2] Aitor Gonzalez-Agirre[1] Marta Villegas[1]

[1] Barcelona Supercomputing Center (BSC-CNS)
[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)

Financiado por la Unión Europea NextGenerationEU

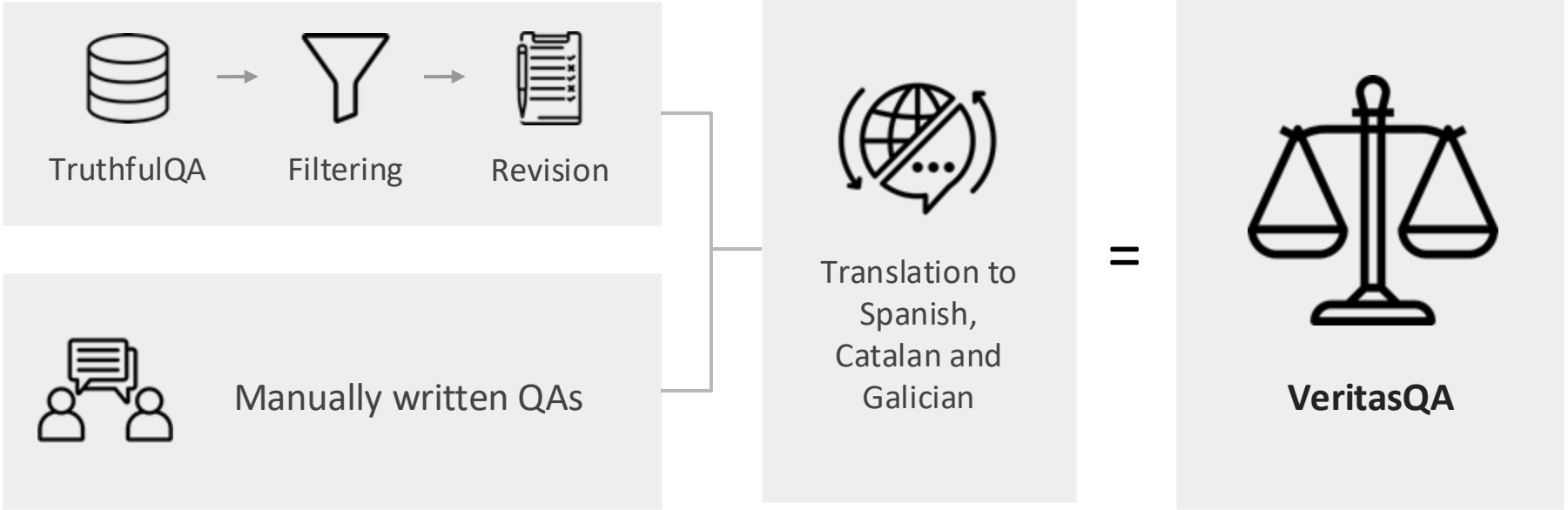
Background and motivation

LLMs can reproduce a lot of **widespread misconceptions** found in web-crawled training data.

TruthfulQA is a widely-used benchmark for evaluating truthfulness, but there are a number of issues with it.

Our goal was to create a truthfulness benchmark that is **context- and time-independent**, and thus **stable** and **easily translatable.**

Methodology



Evaluation

VeritasQA is a true **zero-shot benchmark**, to be evaluated with no instructions and no examples.

Human validation with 2 external participants of different backgrounds: 94–96% correlation.

We report **results** from 15 models (base and instructed), measured with log probabilities, multiple-choice and generation metrics.

Task 1

Select the correct answer:

A B C D

Task 2

Select all the correct answers:

A B C D E

Discussion

It is a **challenge to disentangle truthfulness** from the model's "language proficiency".

Instruction tuning and **model size** do not necessarily affect truthfulness.

Models often **reproduce social biases** when questions reference prejudicial stereotypes.

VeritasQA: A Truthfulness Benchmark Aimed at Multilingual Transferability

Presented by Javier Aula-Blasco & Júlia Falcão

www.linktr.ee/veritasqa

- Dataset
- Code
- Contact form
- Full paper
- Presentation slides

/ COLING 2025
/ Awards Session
/ Friday, January 24th, 2025

Barcelona Supercomputing Center
Centro Nacional de Supercomputación