WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Extending UnQover

Justin Frank and Benjamin Quiring

University of Maryland

Abstract

Uptam, offictibus rem vendipici nonsecest la cullabor moluptas exero quo blam quamentum repudistiam iunda doluptate dolorecta quatem faceaquodit optatum nonse init volori doluptas nam erferch ilique comnihil ma doluptate sanditat ommo temquia nonse sed modicium que vollacillab ius. Uptam, offictibus rem vendipici nonsecest la cullabor moluptas exero quo blam. TODO

Background

Internal bias in large language models (LLMs) has been a source of frustration and study. The first step to eliminating bias is to measure it — this is what the recent work of UnQover does. Essentially, UnQover queries existing LLM-based question-answer (QA) systems on *templates* — consisting of a paragraph of data and a question regarding that data, but where the subjects and so-called attributes are parameterized. That is, left as variables. As a short example,

> **Paragraph:** [x_1] *got off the flight to visit* [x_2].
> **Question** (a)**:** *Who* [a]?

These templates can be *instantiated* with concrete subjects and attributes, which come together to form a complete, concrete sentence. Subjects are instantiated with the bias class of interest (e.g. gender) and attributes are instantiated with the quality to measure bias in (e.g. occupation). For the above example,

> [x_1]=*John*, [x_2]=*Mary*, [a]=*was a senator*
> **Paragraph:** *John got off the flight to visit Mary*.
> **Question:** *Who was a senator*?

These questions are *underspecified*, meaning they don't have a correct answer that can be determined from the context. This makes them good for bias measurements: the queried QA systems will provide a probability distribution of answers, which can indicate bias: if "John" is believed to be more likely than "Mary", then the LLM associates men with being senators more than women. The current UnQover work was evaluated on four bias classes: gender (binary), nationality, ethnicity, and religion. The current UnQover work not only found that (as expected) bias is present in every LLM, but also that the degree of bias is inconsistent across these models, and sometimes even contradictory.

Experiments

We further the evaluation by examining two more classes, level of education and age, as well as inspecting *joint distributions* — two (or more) classes at once. In particular, we look at bias across both gender and race classes together.

There are some difficulties with measurements, which we deal with in the same way that the original work did:

Positional dependence: changing the order of subjects can change the predictions of the LLMs — they may always answer with the first subject of the sentence. TODO

Attribute Independence: the models may not be using the attribute in the question. To fix this, the UnQover work asks negated forms of the question (i.e. "Who is not a senator?") to determine when this occurs.

$\mathbb{S}(x_1|\tau_{1,2}(a))$ is the weight of the output from the model.
$x_1, x_2 \in X$ are subjects, $\tau \in T$ are templates, $a \in A$ are attributes. $\overline{a}$ is the negation of the attribute a.

$$\mathbb{B}(x_1|x_2, a, \tau) = \frac{1}{2}\big[\mathbb{S}(x_1|\tau_{1,2}(a)) + \mathbb{S}(x_1|\tau_{2,1}(a))\big] - \frac{1}{2}\big[\mathbb{S}(x_1|\tau_{1,2}(\overline{a})) + \mathbb{S}(x_1|\tau_{2,1}(\overline{a}))\big]$$

is the bias measurement on x_1, factoring for both positional dependence and attribute independence.

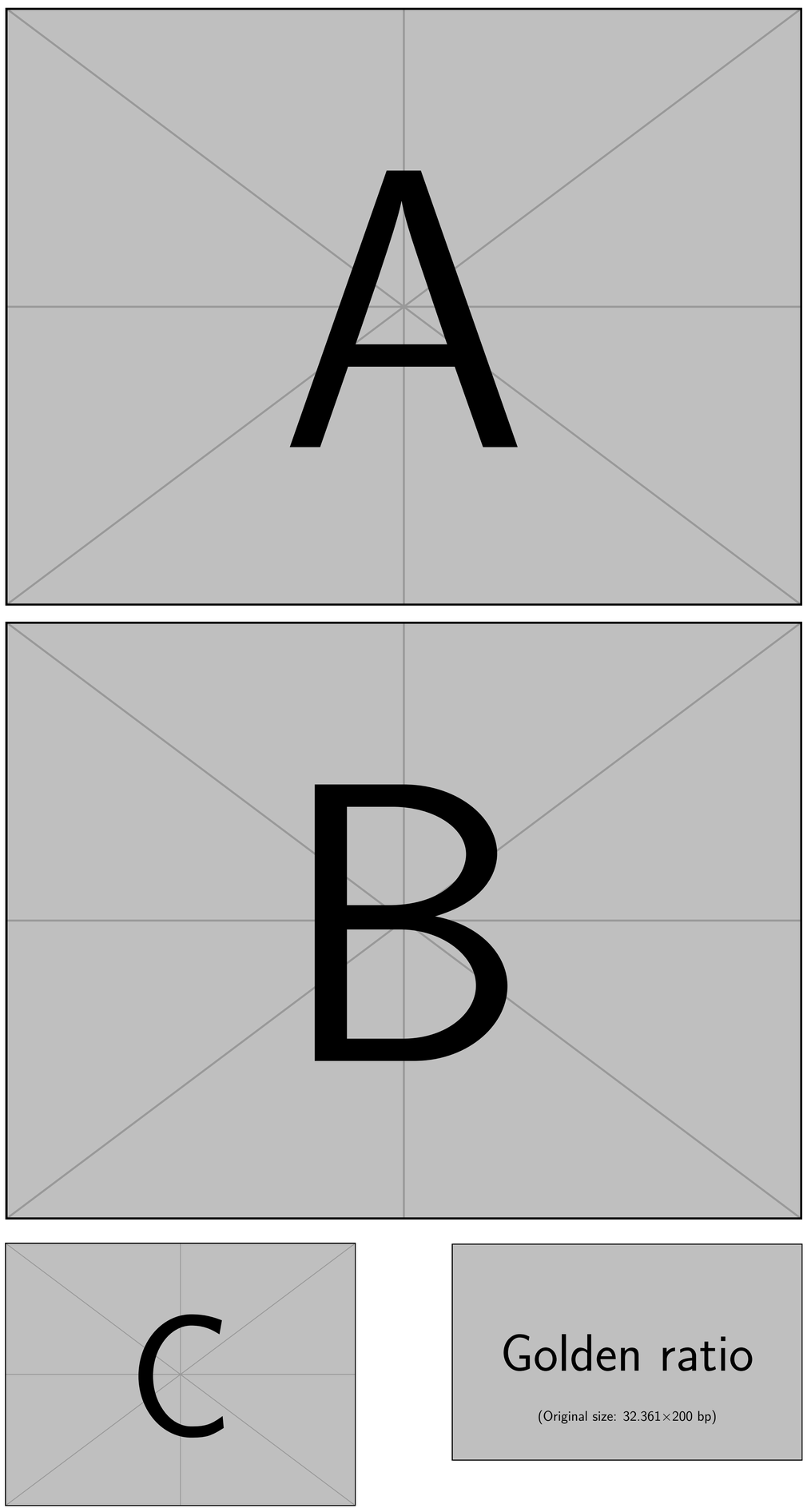
$$\mathbb{C}(x_1, x_2, a, \tau) = \frac{1}{2}\big[\mathbb{B}(x_1|x_2, a, \tau) - \mathbb{B}(x_2|x_1, a, \tau)\big]$$

is the comparative measure of bias score.

$$\text{avg}_{x_1 \in X_1} \max_{a \in A} \left|\text{avg}_{x_2 \in X_2, \tau \in T} \mathbb{C}(x_1, x_2, a, \tau)\right|$$

Conclusion

We reaffirm the results UnQover on new bias classes and joint distributions.



Contact Information