A Condition Difference

Since using LLMs to replicate experiments made with humans is still exploratory, it is important to understand how the change in context created by altering the order of the conditions may affect the results. This distribution can be seen in Figure 4. For reference, we include the distribution without condition distinction within labelled as 'overall'.

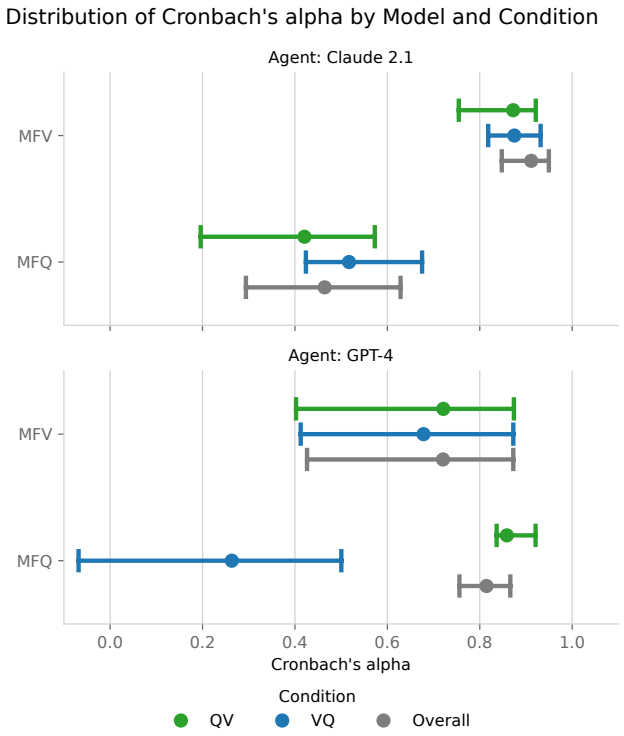

Figure 4: Distribution of Cronbach's alpha per AI agent and condition. Error bars cover all observations.

We ran t-tests to explore whether there were significant differences in α distributions between the instrument order conditions (***QV*** and ***VQ***) for each model. We found no significant difference for Claude 2.1 ($t(22) = -0.60, p = 0.55$). However, we found a significant difference for GPT-4 ($t(22) = 5.49, p < 0.001$), with a considerable effect size (Cohen's $d = 1.16$), as shown in the GPT-4 panel of Figure 4.

We understand this difference in itself may display consequences related to our concerns. The lower α values observed by GPT-4 in the ***VQ***condition regarding the MFQ indicate that the model cannot appropriately perform the task in every context, even though we submitted the same query and instructions.

This also relates to the Concept Mastery aspect raised in our discussion, GPT-4 does not display behaviour that would be compatible to the equivalent of someone's understanding of each concept under scrutiny. We opt not to elaborate on this in our results as it would be tangential to our proposed analysis.

To assess whether there was a difference in the endorsement of foundations in each condition we created a two way ANOVA (foundation $\times$ condition). We made separate models for each LLM and instrument, as MFQ and MFV used different scales. The results of all four analysis revealed a significant relation of foundation, order condition, and their interaction. The order condition effect size was substantially larger for GPT-4 in the MFQ and Claude 2.1 in the MFVs. The results of these analysis is available in our Supplementary Materials at OSF and GitHub. Figure 5 displays the answer distribution by Foundation for the MFQ, and Figure 6 for the MFVs.

Curiously, all foundations were more endorsed in the ***VQ***condition for both LLMs. All MFV foundations endorsement were significantly different for Claude 2.1, while GPT-4 presented a significant difference in 3 out of 5 foundation in the MFQ, and a single one in the MFV. The largest difference were in the MFV-Liberty pair for Claude 2.1 and MFQ-Purity for GPT-4. Although we will not elaborate on the substantive direction of alignment discussion, our data suggests the ordering effect was significant not only in the consistency but in the values outputted by models.

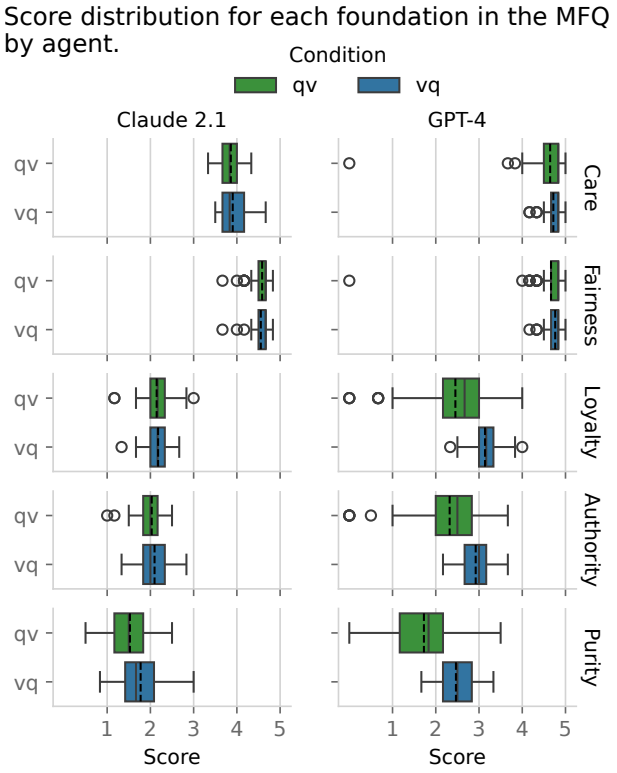

Figure 5: Distribution of Foundations scores in the MFQ by condition. Dashed lines represent mean

B Correct answer effect

Another relevant point regarding ours results is the occurrence or absence of the "correct answer effect": an LLM outputting the same single answer every time it is prompted with a question (*i.e.*, with no variability), even when it comes

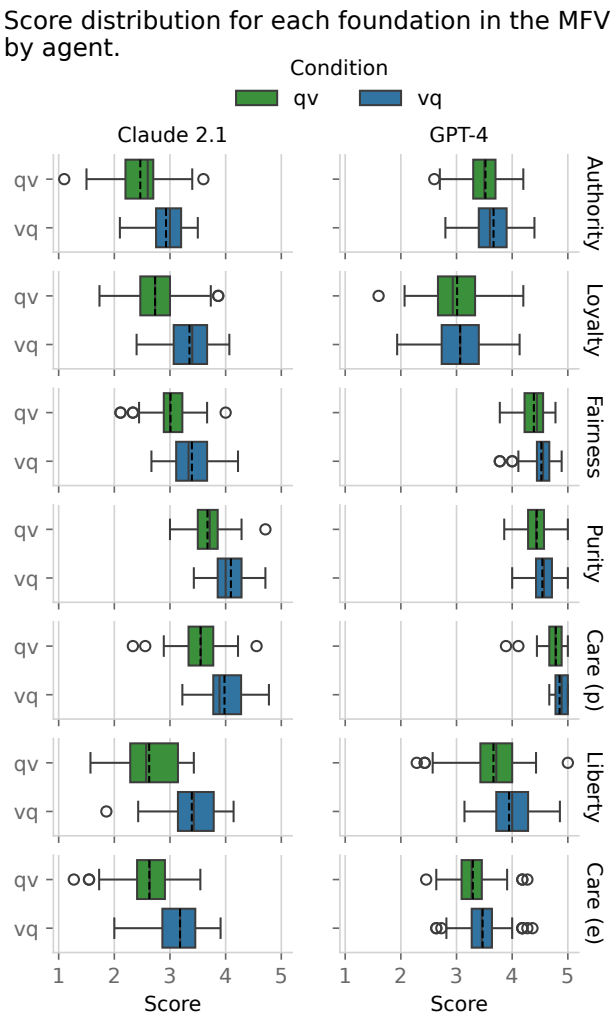

Figure 6: Distribution of Foundations scores in the MFVs by condition. Dashed lines represent mean

to complex moral issues. As indicated by prior research (Park, Schoenegger, and Zhu 2023; Almeida et al. 2024), this phenomenon seems to be characteristic of LLMs, where they always return the same answer for certain experiment questions. Although the "correct answer effect" is not the main focus of our investigation, the occurrence of this phenomenon should be kept in mind for further investigation on the alignment problem, especially in deciding whether displaying variability in responses is desirable or not. Furthermore, it would be an important caveat of our results.

To analyse this, we again looked at the condition order individually. In the MFVs, only GPT-4 displayed this behaviour, in 2 vignettes in the ***QV*** and 4 in the ***VQ*** condition (from which 1 was common to both conditions). With regard to the MFQ, the pattern was present in both models. GPT-4 presented a single answer to 5 items (plus one attention check) in ***VQ***, while only for the Math item in ***QV***.

Claude 2.1 displayed the correct answer effect for the MFQ but not for the MFVs. It occurred in both attention check items, both in the ***VQ*** and ***QV*** conditions. Furthermore, Claude generated a single answer to all 30 items in ***QV*** and to half of them in ***VQ***. This finding provides an explanation for the comparatively much lower α values observed in the MFQ for Claude when comparing it to other agent-instrument pairings.

This is in stark contrast to the MFQ experiment ran by Almeida et al. (2024) and Park, Schoenegger, and Zhu (2023). They only reported this effect in the political affiliation preliminary question, which we did not include, but not in any of the MFQ components. Although elaborating on the correct answer effect is not our focus, our results again show evidence of this phenomenon in the OpenAI family of models. It also provides evidence that it may occur with other models – in our case, Claude 2.1. Furthermore, this aspect seems to be highly impacted by context, as they differed across conditions in our study.