Qeios

Peer Review

Review of: "Self-Pluralising Culture Alignment for Large Language Models"

Evgenii Arsentev[1]

1. Independent researcher

What the paper claims

CultureSPA tries to align one model to many cultures without importing external cultural data, by mining the model's own knowledge. From 260 World Values Survey Wave 7 seed questions across 13 topics, it generates 13,000 questions, drops 153 in filtering, then collects LLaMA-3-8B-Instruct answers twice: once with no cultural framing and once prompted to answer as a member of each of 18 countries (Section 5.1). Items where the two answers differ are kept, on the argument that a shift signals culture-specific knowledge rather than the model default, and 62,127 such pairs (77,086 with cross-culture thinking) are used for LoRA fine-tuning (Section 5.1). The cultural alignment score rises from 66.22 to 67.29 with the P1 prompt and to 69.11 when combined with P2+P3 (Section 5.3). Joint tuning beats culture-specific tuning at most data fractions (Table 2), and Table 3 reports general benchmarks at MMLU 67.61 to 67.97, GSM8K 79.30 down to 77.94, IFEval 67.84 up to 69.13, and down to 66.54 at 240K.

The selection rule in Section 4.3 is the interesting idea here, and the mechanism study in Section 6.1 that contrasts it against consistent and random sampling is exactly the right way to defend it. Figure 4, showing consistent sampling helping only the already-favored cultures while hurting Asia and Africa, is the most convincing evidence in the paper. My concerns are about what the reported gains can bear and about one possible leak.

Does the conclusion hold on the data?

1. The headline gain is about one point, and nothing in the paper bounds the noise. The main result moves the score from 66.22 to 67.29, which is 1.07 points, and the best combination reaches 69.11. No seeds, no repeated fine-tuning runs, no standard deviations, and no significance test appear anywhere. LoRA runs on the same data vary by a comparable margin from initialization and data order alone. The

 doi.org/10.32388/DBCPMZ

natural fix is cheap: three seeds per configuration, and a paired test across the 18 cultures, which is the right unit of analysis since each culture contributes its own score.

2. The same one-point question decides the claim about general abilities, and there the paper argues the other way. The abstract says alignment improves "without compromising general abilities," but Table 3 shows GSM8K falling from 79.30 to 77.94 at 60K and 78.39 at 240K, and IFEval falling to 66.54 at 240K. Those are drops of the same magnitude as the cultural gain being celebrated. Either differences around a point are meaningful, in which case the mathematics regression is a real cost that belongs in the abstract, or they are not, in which case the main result is not yet established. Variance estimates settle both at once.

3. The relationship between the 260 seed questions and the evaluation questions is never stated, and it decides how to read the whole paper. Training questions are generated from WVS seeds by the same model that is later evaluated on WVS items. If the evaluation questions are the seeds, or paraphrases close to them, the gain may be paraphrase-level memorization of the target answers rather than activated cultural knowledge. Please state explicitly whether the evaluation set is disjoint from the 260 seeds, report the textual overlap between generated questions and evaluation items with a similarity measure, and if any overlap exists, re-report the score on a clean held-out subset. This is the first thing a skeptical reader will ask.

4. A pluralism paper is scored against a majority answer. Section 3.2 builds the per-culture reference by aggregating participants with a majority rule, so a model that always emits the modal option of each country scores well while representing none of the internal disagreement that makes a culture plural. That is in tension with the stated goal. Add a distributional comparison between the model answer distribution and the full WVS response distribution, for instance, a Jensen-Shannon divergence per question, and report the entropy of the model answers before and after tuning. If tuning sharpens the model onto modal answers, the score improves while pluralism decreases, and readers should be able to see which of the two happened.

5. One model family. All results come from LLaMA-3-8B-Instruct, yet Section 6.1 states a general mechanism about biased versus adaptable knowledge. Since the selection rule depends on how a particular alignment procedure shaped the model's default answers, the finding could be a property of LLaMA3 rather than of LLMs. One additional family at a similar scale, tested only on the main configuration, would tell the difference.

6. Figure 5 carries a quantitative claim in a caption. The correlation between cross-cultural alignment from model outputs and from WVS is the evidence for the claim that the model recovers cultural relations, but the Pearson coefficients appear only in the figure. Put the values, with their intervals and the number of culture pairs behind them, in the text.

Reproducibility

Good on the structural side: the WVS wave, the 13 topics, the count of seed questions, the 18 countries, LLaMA-Factory with LoRA on a single A100, and the prompting templates in the appendix all make the pipeline legible.

What a replication still needs: the LoRA configuration, meaning rank, alpha, dropout, and target modules, plus learning rate, epochs, batch size, and maximum sequence length; the decoding parameters used to generate the 13,000 questions and the 19 answer sets, since temperature governs how much output shift the selection rule will see and therefore the size of the training set; the random seed; the rule that removed the 153 filtered questions; the exact WVS question identifiers used for seeding and for evaluation; and a release of the 62,127-example tuning set, or at minimum the generation code. Releasing the tuning data would let others test point 3 directly without re-running anything.

What to fix

1. Sections 5.3 and 5.4: run at least three seeds per configuration and report mean and standard deviation; add a paired test across the 18 cultures for the main comparison.
2. Abstract and Table 3: reconcile the claim of no compromise to general abilities with the GSM8K drop from 79.30 to 77.94 and the IFEval drop to 66.54, using the same variance estimates.
3. Sections 3.2, 4.1, and 5.1: state whether the evaluation questions are disjoint from the 260 seeds, quantify textual overlap between generated and evaluation questions, and re-report on a clean subset if needed.
4. Section 3.2: add a distributional metric against the full WVS response distribution and report model answer entropy before and after tuning, so that score gains can be distinguished from collapse onto modal answers.
5. Section 5.1: repeat the main configuration on a second model family to show the mechanism in Section 6.1 is not specific to LLaMA-3.

6. Section 6.2 and Figure 5: move the Pearson coefficients and the number of culture pairs into the text with intervals.
7. Appendix: give LoRA hyperparameters, optimization settings, generation decoding parameters, seeds, the 153-question filtering rule, WVS question identifiers, and release the tuning data or the generation code.
8. Table 1 is the main results table, but its values appear only through the narrative in Section 5.3; present the table itself with per-culture rows so the claim about gains for African cultures can be checked.
9. Typos: "CutureSPA" appears twice in Section 1; Section 6.2 has "generates LLM outputs that more accurately the cultural relationships," missing a verb; Section 6.3 has "Results in Table 3 shows" and "may enhances."

Recommendation

Major revisions, though the work is in good shape and the required additions are mostly analysis rather than new experiments. The idea of harvesting a model's own culture-conditioned shifts as a training signal is neat, the ablation against consistent and random sampling makes the case honestly, and the cross-cultural clustering analysis is a nice touch. What is missing is the statistical floor under a one-point effect, a clear statement that evaluation items were never seen as generation seeds, and an acknowledgment that a majority-matching metric is an odd yardstick for pluralism. Address those three, and the contribution stands on its own.

Evgenii Arsentev, PhD

Declarations

Potential competing interests: No potential competing interests to declare.