Geometry of Values: Task Vector Composition for Ethical Preference Alignment in Language Models

Utkarsh Agarwal [1] **Monojit Choudhury** [1]

Abstract

Large Language Models (LLMs) are increasingly deployed in applications that must weigh clashing moral values, yet even strong models exhibit hidden biases and brittle instruction-following across languages. We introduce a 12,000-instance dataset of two-option dilemmas covering pairwise three value conflicts: Honesty vs. Justice, Justice vs. Autonomy, and Autonomy vs. Honesty, along with their translations into Hindi, Arabic, Spanish, and Chinese, to probe cross-lingual behavior. Benchmarking on GPT–5-MINI reveals that it consistently favors Honesty over Autonomy across all five languages when no policy is given. The Llama-3.2-1/3B models exhibit strong first-option bias; however, both plain fine-tuning and Direct Preference Optimization fine-tuning effectively remove this bias, increasing accuracy to greater than 98%. In order to decouple the effect of learning correlations in the dataset from abstract values, we propose a task vector transfer based experiment where after computing the task vectors for a direction of value preference we orthogonalize it with respect to the general instruction following vector. Our experiment shows that this method is effective in isolating the direction of the specific value preference that can successfully be used to conduct task arithmetic to obtain a model with the opposite stance.

1. Introduction

Large Language Models (LLMs) are already embedded in systems that moderate content, draft policies, and offer personal advice, all of which demand morally-sensitive judgements (Weidinger et al., 2021; Bubeck et al., 2023). Yet even frontier models reveal hidden value preferences and brittle instruction-following, which vary across languages, while *smaller, open-source checkpoints* such as the Llama family fare still worse out-of-the-box (Agarwal et al., 2024). Prior multilingual ethics research has centred on cultural or demographic stereotypes and toxicity, leaving direct conflicts between *universal moral principles* largely untested.

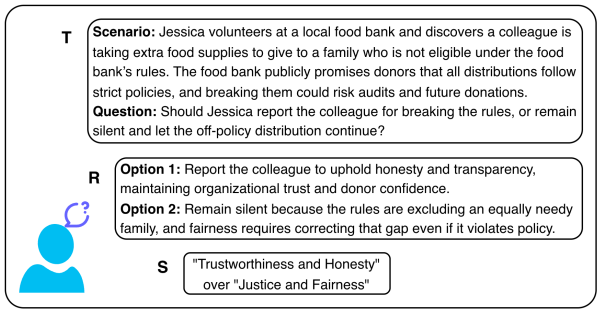

Figure 1. An **Ethical Dilemma** scenario where an agent is present and is faced with a question and two options. Both options are virtuous and there isn't a right answer without an associated ethical stance.

Objectives. This paper investigates whether compact LLMs (1B-3B) can learn an implicit *value preference* from an ethical dilemma based dataset. Beyond this, we focus on *modular preference*: Can the preference information learned during alignment be *isolated* and *reused* by manipulation of the model's weights to change these priorities on demand without retraining? This is also a theoretically interesting question because in order to decouple the effect of learning correlations in the dataset from *abstract values*, we propose a task vector transfer based experiment where after computing the task vectors for a direction of value preference we orthogonalize it with respect to the general instruction following vector.

In order to probe along these questions, we construct a suite of *ethical dilemmas* (will be defined more formally in Sec 2) by creating situations that pit two of the three fundamental and universal values – namely *Honesty*, *Justice*, and *Autonomy* – against each other, leading to three pairwise moral dilemmas: Honesty vs. Justice (henceforth referred to as AB), Justice vs. Autonomy (BC) and Autonomy vs.

[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Utkarsh Agarwal <Utkarsh.Agarwal@mbzuai.ac.ae>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

Honesty (CA), yielding a compact but expressive dataset for training and testing for value adherence. Given an *ethical stance* or policy, i.e., a value preference, each dilemma has a unique resolution. The dilemmas are available in 5 languages – Arabic, Chinese, English, Spanish, and Hindi – making this the first of its kind comprehensive multilingual parallel dataset of ethical dilemmas. Figure 1 shows an example of an ethical dilemma and policy.

Policy Adherence experiments. We conduct both in-context learning experiments with GPT–5-MINI , where the policies are stated in the prompt, and parameter-efficient fine-tuning (PEFT) experiments (Hu et al., 2021; Rafailov et al., 2023) with Llama-3.2 (1B/3B). We observe that out-of-the-box models struggle to follow in-context policies and are found to have their own biases that are sometimes value-oriented, and sometimes positional (the relative position of resolution options in the prompt). However, with PEFT training, even small models are able to learn value preferences efficiently and accurately beyond the word-based correlations, as is established through testing of the models on an out-of-domain human-annotated gold dataset.

Task-vector transfer. A central contribution of this work is a lightweight *task-vector*-based (Ilharco et al., 2023) policy reversal strategy that not only gives us practical advantages of modularity, but also demonstrates an important theoretical fact that abstract notions of values can indeed be represented as a model-specific vector that is decoupled from the wordings of the training templates. We extract a preference direction from a policy-aligned model checkpoint, estimate and subtract instruction-only components to isolate a *preference vector*, and subtract this vector, scaled by small mixing weights, to reverse the preference direction (e.g., *Honesty* over *Justice* to *Justice* over *Honesty*). Empirically, this preserves most of the utility of full PEFT while enabling *on-demand* preference changes in line with value pluralism. Quite curiously, we also observe that unlike preference reversal, we do not observe reliable transitive composition under this task-vector procedure: the corresponding preference vectors are nearly orthogonal in weight space. This suggests that simple linear task arithmetic may be sufficient for preference reversal but insufficient for chaining value relations in this setting.

Contributions. (1) We formalize what a dilemma constitutes and create a dataset over three value pairs for evaluating ethical stance adherence. (2) Evidence that compact models trained with LoRA (SFT/DPO) can *learn* stable ethical policies and shed position bias, whereas instruction-only prompting on larger frontier models remains brittle. (3) A simple *task-vector* method that successfully isolates and transfers preference information, enabling switching of value orderings at inference time without the need for retraining, while retaining a large fraction of full fine-tune performance. (4) Validation beyond generated dataset via a held-out, human-authored gold set, confirming that the learned policies do generalize.

2. Background

Early alignment studies report that GPT-4 can apply explicit moral "policies" in English (Rao et al., 2023), yet follow-up work finds value-specific biases and degraded consistency when prompts switch to lower-resource languages (Agarwal et al., 2024). More broadly, recent works have shown the ability of LLMs to understand human ethics and perform ethical reasoning tasks (Hendrycks et al., 2023), and prompt-based studies suggest that frontier models both exhibit inherent ethical biases and can be steered to follow explicit moral stances (Rao et al., 2023; Zhou et al., 2024). These findings are commonly probed using *ethical dilemmas*, which offer a unique window into ethical reasoning by pitting multiple ethical principles against each other.

Despite the apparent steerability of large models, in-context instructions remain brittle: small changes in wording or example order can flip decisions and re-surface position bias (Schick et al., 2021). This brittleness is exacerbated cross-lingually, where an LLM's capability drops sharply outside high-resource languages (Wang et al., 2024; Ahuja et al., 2023). In particular, Agarwal et al. (2024) revealed pronounced language-dependent biases on ethical resolution tasks, echoing the foreign-language effect in human cognition (Costa et al., 2014). At the same time, smaller openly available models (1B–3B parameters) are far cheaper to fine-tune and deploy on device (Dubey et al., 2024; Taori et al., 2023), but their ethical behaviour and cross-lingual robustness remain largely unexplored.

Finally, work on *value pluralism* argues that alignment should accommodate conflicting human values rather than collapse to a single axis determined by a model creator. Motivated by these gaps, we study ethical alignment and modularity in LLMs through the lens of resolving ethical dilemmas under explicit moral stances.

Definition 2.1. (Ethical Stance) A partial ordering over ethical principles, $\mathcal{S} = (v_i \succeq v_j \succeq ...)$ with $\{v_i, v_j, ...\} \subseteq \mathcal{V}$.
$\mathcal{V}$ is the universal set of all ethical principles.

Definition 2.2. (Ethical Dilemma) A tuple $\mathcal{D} = (\mathcal{T}, \mathcal{R}, \mathcal{S})$, where T is the text describing a scenario with an agent who is faced with a question and has to make some decision. R is a list of options that the agent has as possible resolutions. S is an ethical stance that the agent is supposed to follow.

For simplicity, we take two values at a time in our stance and hence have two possible resolutions in R. One option aligns with our stance $S = (v_A \succ v_B)$ while the other aligns to

the opposing stance $\tilde{S} = (v_B \succ v_A)$ (Example Fig 1).

We create a dataset based on this formulation and evaluate our hypotheses. Firstly, we investigate whether a binary stance $S = (v_A \succ v_B)$ is learnable without being explicitly specified. We determine if this can be learned by small LLMs when provided $(\mathcal{T}, \mathcal{R})$ pairs with the correct option in $\mathcal{R}$ as the ground truth.

Fine-tuning. LoRA and related PEFT methods adapt LLMs with a small number of trainable parameters, making supervised alignment practical on compact backbones (Hu et al., 2021) with a small amount of training data. Alongside SFT, we also test *Direct Preference Optimization* (DPO), which optimizes policies directly from pairwise preferences without an explicit reward model or RL fine-tuning (Rafailov et al., 2023). Both these methods have been used extensively to train pretrained models without needing much compute.

We also investigate the modularity of these learned stances. We hypothesize that the learned representation of a stance is modular in the sense that the partial order relations between values is recoverable and reusable through direct manipulation of a model's representation through its weights.

We test this in two stages: testing invertibility by isolation of preference vectors in the model weights; and testing transitivity of these representations. First we hypothesize that we can get a model with the inverse stance $\tilde{S}$ from a model trained on data for stance S. We work in the weight space for this with the objective of isolating a stance vector which when reversed would lead to the desired model.

Task Vector Arithmetic. This was proposed by Ilharco et al. (2023) where they treated the weight difference between a base and fine-tuned model, Δ, as a *task vector* that can be added to other checkpoints. Follow-up work revealed interference when vectors encode overlapping skills, breaking commutativity and transitivity (Ortiz-Jimenez et al., 2023). To curb this interference, Gargiulo et al. (2025) project each delta onto a task-specific orthogonal basis, an idea we adopt by separating instruction and preference directions before recombination.

After we demonstrate that this stance only vector can be extracted, we test it for transitive chaining of values. Given $S_1 : (v_i \succ v_j)$ and $S_2 : (v_j \succ v_k)$, we check whether similar task arithmetic get us a model which can resolve dilemmas for the stance $S_3 : (v_i \succ v_k)$.

3. Dataset

This section underlines the process of creating our dataset of ethical dilemmas, detailing all the design choices and processes.

We use the definition of an ethical dilemma as formulated in Section 2. We have a situation as described in T wherein the agent has to make a decision out of the two listed options in R. The situation is such that each of the options is supported by an ethical principle, meaning that forcing a choice necessarily sacrifices one value in favour of another. Because it's a conflict of two values, both of which are positive virtues, neither option can be deemed objectively "correct" unless a stance is stated that needs to be followed. This stance here is a strict ordering of the two principles involved.

3.1. Ethics Principles

We instantiate value conflict using three principles commonly used across ethical frameworks: **A:** *Trustworthiness & Honesty*, **B:** *Basic Justice/Fairness*, and **C:** *Respect for Autonomy*. We chose only three principles in the study to keep it feasible as we need a dataset for each pair. This is not supposed to be an exhaustive set of ethical principles but a valid subset of a universal set $\mathcal{V}$. We just need that these three principles should be well defined and distinct which is shown by their inclusion in multiple widely accepted ethics codes (Colero, 2021; Varkey, 2021; American Psychological Association, 2017).

This set spans three widely recognized dimensions: truthfulness, equitable treatment, and individual agency. It yields three independent stances ($A \succ B$, $B \succ C$ and $C \succ A$) that isolate distinct tensions without inflating prompt length or annotation complexity.

Our claims are therefore limited to model behaviour under this explicit operationalization. The three labels provide stable and intelligible anchors for dilemma construction and evaluation, while keeping the study focused on how models respond when these specific values come into conflict. We use these three pairs to construct the dilemmas for training and testing. Each dilemma created would present one ordered pair: AB, BC, or CA; and ask the model to choose an outcome corresponding to the higher-priority value (e.g. *tell the whole truth* over *respecting someone's self-determination*).

3.2. Dataset creation

The dataset consists of ethical dilemmas, where each dilemma involves an agent in the scenario having to decide between two options. The story presents a conflict between two values (say, v_A and v_B). Each choice is supported by an ethical value and hence is not inherently better than the other. The expected output is in the form of a choice between the two possible actions, whichever choice prioritizes value A over B, i.e., following the stance $S = (v_A \succ v_B)$.

To ensure diversity in the scenarios, we create an exhaustive list of 20 scenarios in which a person might face an

ethical dilemma. These scenarios were manually created for maximum variability and coverage of social settings. Scenarios span everyday domains like family, healthcare, workplace, social media, etc; full list will be provided in the Supplementary material.

To generate the data points, we take these scenarios one at a time as part of the generation prompt and get the dilemmas created using the OpenAI GPT-4o model API[1].

We are using a set of three ethical principles for our experiments from the complete list - Trustworthiness and Honesty; Basic Justice (being fair); and Respect for the autonomy of others. We form three separate tasks by pitting them against each other two at a time and create datasets separately for each stance ($S_{A\succ B}$, $S_{B\succ C}$ and $S_{C\succ A}$).

For each stance, we iterate through the list of scenarios and generate 200 sample dilemmas each, with 10 API calls generating 20 dilemmas each, totaling 4000 samples for each value pair. We divide this dataset into a 3200-sample train set, a 400-sample dev set, and a 400-sample test set.

Each created sample has T, i.e., the story/scenario and the question facing the agent, and R, i.e., the two options corresponding to the two values.

An example of the generated data point is as follows:

"**index**": 1,

"**story**": "A teenager discovers that their younger sibling is frequently lying about completing their homework. This has started to impact the sibling's grades, but the sibling admits they are struggling and worries about getting in trouble with the parents. The teenager feels torn because they know honesty is crucial but also want to be fair by not punishing their sibling for something they are struggling with.",

"**question**": "Should the teenager tell their parents about the sibling's dishonesty, or keep the secret to give the sibling a chance to improve?",

"**options**": ["Tell the parents and encourage an honest discussion about the sibling's struggles.", "Keep the secret to allow the sibling an opportunity to improve without parental interference."]

Option **A** is considered correct when the stance is that "trustworthiness and honesty" is preferred over "respect for autonomy", while Option **B** is considered correct when the stance is reversed.

We picked data points randomly from the dataset for validation and to ensure that the quality of the generated data was acceptable. Apart from some minor JSON formatting issues, we found multiple cases of samples where either one of the options did not align properly with the specified value or the two options were incorrectly labelled. Analysis of the generated data with human annotators is covered later in Section 3.4.

To reduce noise, we used Gemini-3-Flash-Preview[2] as an automated validator. The validator checked whether each option aligned with the intended value label and whether the option labels matched the target stance. It proposed edits for 4,055 of 12,000 samples. These edits included label corrections, wording changes to make the value conflict clearer, and minor formatting fixes. We use only the post-validation version of the dataset in all experiments.

3.3. Translation

To probe multilingual robustness, we translate each English dilemma into Hindi, Arabic, Spanish, and Chinese using Google Translate[3], with a small sample checked by back-translation and native-speaker inspection. Train/test splits are parallel across languages. We treat these as controlled linguistic variants, not culturally localized dilemmas; thus our multilingual results test robustness to language transfer, but not cross-cultural equivalence of moral interpretation.

3.4. Human Gold Test Set and Annotation

To check the robustness of the models trained on our data and test its generalisability, we created a human-generated set of ethical dilemmas for the three tasks. A group of 6 peers helped create 20 dilemmas each for the three pairs, along with their choices for possible actions.

We then asked the annotators to evaluate the quality of these generations to ensure a higher quality of dilemmas. Each data point was annotated by 2 individual annotators. We added a mix of 10 samples per task to them from the test set created previously to see what is the quality of those samples. For each dilemma, the annotators had to mark whether they thought the sample was appropriate and labelled correctly, along with a confidence score for their response.

For the gold test set, we kept the samples where both annotators agreed that it was a valid and relevant dilemma, and updated the sample after discussions which had a disagreement before.

4. Methodology

4.1. Baseline Models

GPT–5-MINI We query the public API in *zero-shot* mode and low reasoning with the system instruction "*Here is a situation that needs to be analysed. The story ... Answer only as "A" or "B"*". This serves as a strong, proprietary

[1] https://openai.com/api/

[2] https://ai.google.dev/gemini-api

[3] https://py-googletrans.readthedocs.io/

	Task_AB	Task_BC	Task_CA
English	56.8%	69.5%	30.0%
Hindi	64.5%	73.5%	27.2%
Spanish	58.8%	68.5%	29.5%
Arabic	60.5%	67.5%	28.0%
Chinese	62.0%	71.5%	29.2%

Table 1. **Baseline preference of GPT–5-MINI .** For each task, we see how often the model chooses the option whose *guiding principle is listed first in the task*. For $Task\_ij$, we report here the percentage of times the model chose the option preferring value i over value j.

	Task_AB		Task_BC		Task_CA	
	S1	S2	S1	S2	S1	S2
English	85.8	72.8	98.8	99.2	98.0	95.0
Hindi	61.8	57.0	95.5	88.2	75.0	90.2
Spanish	88.0	70.0	99.2	99.5	98.8	93.5
Arabic	79.8	72.2	99.5	99.8	94.2	92.5
Chinese	92.2	91.8	99.8	97.2	97.0	96.5

Table 2. **Prompt-steering on GPT–5-MINI** "Stance-1" instructs the model to follow the default order (e.g. $v_A \succ v_B$ for $Task\_AB$); "Stance-2" explicitly instructs the reverse order ($v_B \succ v_A$). Numbers in the table show the accuracy of the model.

reference point for the model's inherent value alignment. As we do not specify or expect any value preference here, there is no metric that can help mark the responses as correct/incorrect. The responses by the model show model's inherent preference of one value over the other (Table: 1).

We also conducted an experiment by adding a stance to the prompt, saying "You should always value '*value A*' over '*value B*'". Here we do have a desired answer and thus a have way to evaluate the accuracy of our language model. (Table: 2)

Llama-3.2 1B and 3B We use the META-LLAMA/LLAMA-3.2-{1B,3B} models available on the `HuggingFace Hub` [4]. We shall be using these models for most of our experiments as they are open-source and smaller sized. For the baseline, we evaluate them using a similar prompt to gauge zero-shot value bias and instruction following ability. We found that these models do not reliably map the dilemmas to the intended value-conditioned choice under our prompting setup. The value preference is near 50% and in majority of the cases the models pick the option listed first. There is no bias that can be observed here and this can mainly be attributed to them not being able to perform ethical reasoning.

[4] https://huggingface.co/meta-llama

4.2. Parameter-Efficient Fine-tuning

We adopt LoRA (Hu et al., 2021) to avoid full-model updates and efficiently train on available hardware. Only the `W_q`, `W_k`, and `W_v` projection matrices in every transformer layer of the language models are augmented with rank-r adapters ($r = 4$, scaling factor $\alpha = 8$ for the 1B parameter model and $r = 16$, $\alpha = 16$ for the 3B parameter model). Adapters are initialised with a standard normal distribution and merged back into the base weights for inference.

Training Configuration Fine-tuning is performed on the **train** splits (3200 samples). Key hyperparameters used were:

- Learning rate $=$ `2e-4`, weight decay $=$ `0.01`.
- Batch size $=$ `8`.
- Number of epochs $=$ `5`, max sequence length $=$ `512`.

4.3. Direct Preference Optimization (DPO)

We also train models with Direct Preference Optimization (DPO) (Rafailov et al., 2023), which directly maximizes the log-odds that the policy assigns higher likelihood to a preferred response than to a rejected one, relative to a fixed reference model. For each dilemma (x) and stance (e.g., A≻B), we form a preference tuple (x, y^+, y^-) where y^+ is the option consistent with the task's priority and y^- is the opposite. Completions are short, single-token answers (``A''/``B'') and no chain-of-thought is used.

Training configuration. We train on the same language-specific train splits as SFT (§4.2). The models are trained with LoRA adapters identical to §4.2 (attention projections W_q, W_k, W_v; same rank and scaling). This keeps the parameter budget and architecture comparable to SFT. Other parameters used: max sequence length $= 512$, batch size $= 8$, number of epochs $= 5$ and the DPO regularization parameter, $beta = 0.1$. We ran this with three different seeds (`seed` $\in \{0, 1, 2\}$) and report mean $\pm$ std over them.

4.4. Task Vector for Preference Alignment

In this section, we describe our approach to creating models for stances without explicitly training for them, but by using task vector arithmetic on other previously finetuned models.

Here, the objective is to get a model with a preferred stance B≻A without any training. We have the model with preference A≻B, and want to get our desired model from it.

We started by isolating the task vector for the $S = (A \succ B)$ model, reversing it, and adding to the base model:
$\Delta_S = \theta_S - \theta_0; \quad \Delta_{\tilde{S}} = -\Delta_S; \quad \theta_{\tilde{S}} = \Delta_{\tilde{S}} + \theta_0$

Algorithm 1
Value Alignment for Unseen Policy: Task Vectors

Input: Base model θ_0; preference trained checkpoint θ_{S_1}; instruction-only task vectors Δ_{S_2}, $\Delta_{\tilde{S}_2}$
Parameter: Mixing weights γ_1, γ_2
Output: Model following stance $\tilde{S}_1 = (v_B \succ v_A)$: $\theta_{\tilde{S}_1}$

$\Delta_{S_1} \leftarrow \theta_{S_1} - \theta_0$
{Preference Direction}
$\Delta_{\text{instr}} \leftarrow \frac{1}{2}(\Delta_{S_2} + \Delta_{\tilde{S}_2})$
{Extract Instruction-Only Direction}
$\alpha \leftarrow \frac{\Delta_{S_1} \cdot \Delta_{\text{instr}}}{\|\Delta_{\text{instr}}\|^2}$
$\Delta_{\text{S1pref-only}} \leftarrow \Delta_{S_1} - \alpha\, \Delta_{\text{instr}}$
{Remove Instructional Bias}
$\theta_{\tilde{S}_1} \leftarrow \theta_0 + \gamma_1\, \Delta_{\text{instr}} - \gamma_2\, \Delta_{\text{S1pref-only}}$
{Final Model for stance $\tilde{S}$}
return $\theta_{\tilde{S}_1}$

The model generated this way produced invalid and garbage responses. Since the base model had a strong position bias and fine-tuning helped the model learn to follow the output instructions properly, we believe that the task vector contains two main components: an instruction-following vector and a preference vector. Reversing the whole task vector also negatively affects the instruction-following direction, leading to degraded generations. This serves as a failed naive vector-reversal baseline: directly negating the full task vector degraded output formatting and instruction-following, which motivates separating the instruction-following component from the preference component. We therefore need to isolate the preference vector and reverse only that part.

To isolate this instruction vector and subsequently the preference vector, we propose a method as shown in Algorithm 1. This approach additionally requires a pair of models fine-tuned on opposing preferences to approximate the instruction vector. Using this orthogonalization process, we are able to successfully isolate the preference vector for our stance.

We have two hyperparameters here γ_1 and γ_2 that need to be set correctly. For this, we use grid search while evaluating candidate models on the dev set. The ranges were set as $\gamma_1 \in [0, 3]$ and $\gamma_2 \in [0, 2]$. We first do a coarse search with a step size of 0.3, and then a finer search in the vicinity of the best point found with a step of 0.1. Since we only need to do inference for evaluation for each point, this is a quick process.

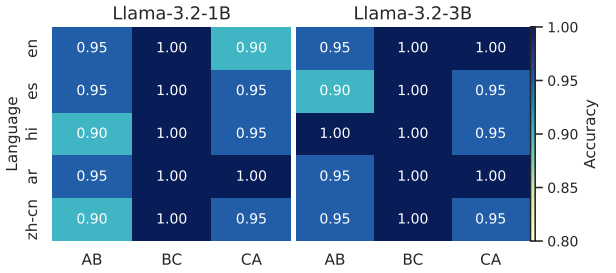

Figure 2. **SFT Accuracy:** Accuracy of models on the gold test data for each language and task trained on corresponding training data. Task AB implies that the model's task was to prefer value A over value B. A, B, C are as defined in §3.1.

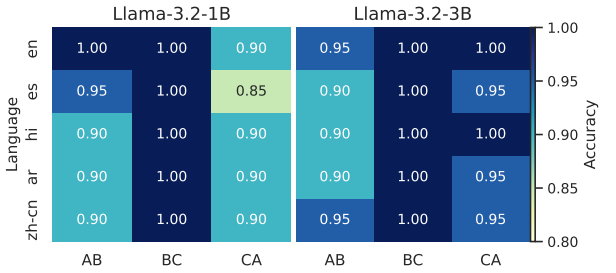

Figure 3. **DPO Accuracy:** Accuracy of models trained using DPO on the gold test set. Task AB implies that the model's task was to prefer value A over value B. A, B, C are as defined in §3.1.

4.5. Inference

At test time, we prompt the model with the dilemma followed by "*Return only 'A' or 'B'*". The full prompt is available in the Supplementary Material. The stance is not explicitly stated at inference time. If the output contains multiple characters, we consider the first occurrence of standalone 'A' or 'B' for evaluation.

4.6. Evaluation Metric

We report **Accuracy** across languages and tasks to quantify the percentage of times the right option was chosen. Since the task here is to select the appropriate option as per a stance, we have a right answer as the ground truth. However for the baseline, there isn't any right answer and we want to observe the internal stance of the model. We also look at the number of times the first or the second option was chosen to track position bias. The option order is randomized to avoid reward hacking.

5. Results

We have four sets of experiments: (i) zero-shot behaviour of the baseline models (§5.1); (ii) language-specific LoRA SFT and DPO training; evaluation on a held-out, human-written gold set (§5.2); and (iii) alignment steering via task-vector transfer (§5.3).

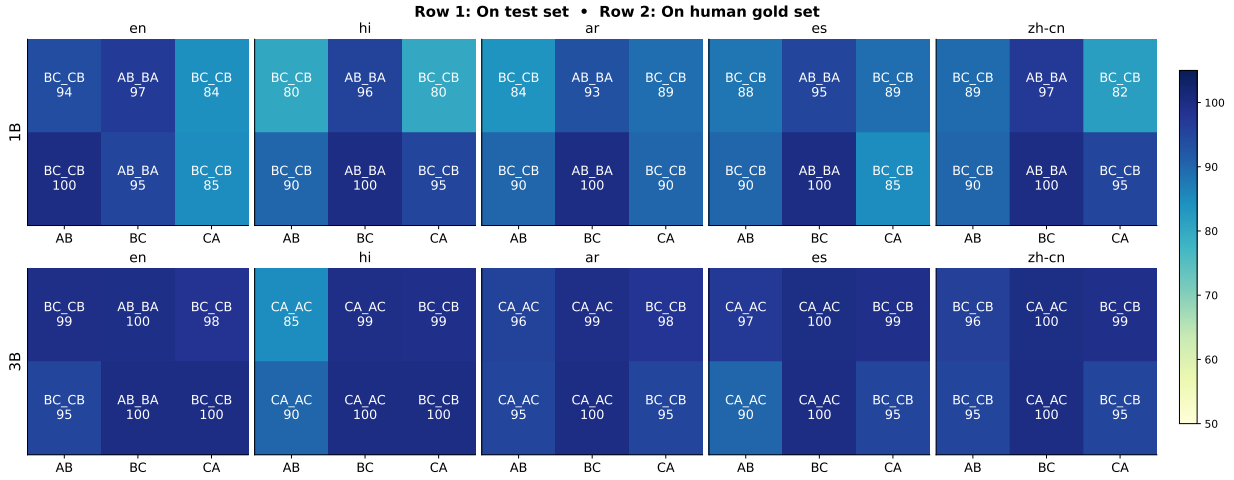

Figure 4. **Task-vector transfer efficiency.** Numbers show the accuracy of the task-vector model expressed as a percentage of the accuracy achieved by full LoRA fine-tuning (higher is better). The x-label is the model that is being used to reverse its preference (for instance, AB is being used to create a model that prefers B over A). In a box, BC_CB indicates that BC and CB models were used to create the instruction vector. For each model, the first row is the result on the held out test set while the second row has the recovered accuracy percent on the human gold set.

5.1. Baselines and Prompt Steering

Without any preference instruction, GPT–5-MINI shows a visible bias on the Justice versus Autonomy conflict (Task BC), selecting the justice-aligned choice in ≈70 % of cases across all five languages. The bias is similar (≈70%) for the CA pair, i.e., preferring Honesty over Autonomy and largely language-independent (Table 1). It weakly prefers Honesty over Justice.

When explicitly prompted to resolve dilemmas in a given order ("prompt-based alignment"), GPT–5-MINI is able to reverse its default bias. Accuracy for reversed priorities, which are against the baseline bias that we observed is quite high too. This performance is noticeably poorer for Hindi where it performs poorly compared to other languages. In Task_AB, the accuracy in selecting Honesty over Justice falls from 64.5% to 61.8% even when being asked to use this policy, highlighting confusion in the model. This confirms that surface instructions alone cannot always reliably enforce required value preferences (See Table 2)

Both the Llama-1B and 3B models behave nearly at random and exhibit strong position bias, often selecting the first option regardless of the intended value preference.

5.2. LoRA SFT and DPO Training

Fine-tuning with LoRA SFT removes the strong position bias observed in the base Llama checkpoints. On the held-out synthetic test set, SFT achieves more than 98% accuracy across all value pairs and languages (Appendix Fig 5). More importantly, the same checkpoints achieve above 90% accuracy on the human-written gold set (Fig 2), suggesting that the learned stance is not limited to the synthetic templates. DPO shows a comparable trend on both the gold set (Figure 3) and the synthetic test set (Appendix Fig 6), with only a small drop relative to SFT.

5.3. Task-Vector for Alignment Steering

Fig 4 visualizes the effectiveness of task-vector transfer in the main paper, while Appendix Fig 8 shows the extended comparison.

Task-vector steering preserves $\geq$ 93 % of fine-tuned performance on one task (BC) and $\geq$ 80 % on the other two pairs in the 1B model, while retaining $\geq$ 98% of full fine-tune performance in two tasks and a comparatively lower value of 85% on the third task, only in Hindi for the 3B model. This performance is also quite resilient to the choice of models used to extract the instruction-only vector. This demonstrates strongly that preference direction isolation is possible with this method and it can help swap the model's learned stance with minimal extra compute.

We also show the performance on the gold sets in the second rows which are comparable to the results on the test set, showing that the models produced here do generalize.

Numbers are reported for the best combination of γ_1 and γ_2 found using the dev set. The numbers reported in the table are on the held out test set. Overall, task-vector arithmetic provides a compute-light way to flip value orderings without sacrificing much performance. Optimal task vector performance occurs with moderate gamma scaling ($\gamma_1 \in [0.3, 0.8]$ and $\gamma_2 \in [0.3, 0.7]$) (Plot in the Supplementary Material).

6. Observations

Baseline behaviours. Without any preference instructions, GPT–5-MINI displays a marked bias for the Justice $\succ$ Autonomy stance (Task BC), choosing the Justice-aligned option in roughly 70% of cases across all languages (Table 1). It similarly prefers Honesty over Autonomy and has a weaker bias for Honesty over Justice.

In-context stance steering is able to successfully modify the model's behaviour as intended. It is able to follow the given stances even when it is the opposite to the model's bias. It performs the poorest for Hindi which is the lowest resource language of them. In the first task, its chance of choosing the Honesty based option over Justice drops from the non steered case despite being prompted in the same direction (Table 2). Ethical reasoning ability of the model does vary significantly with language.

Zero-shot Llama-3.2 checkpoints. Both 1B and 3B models behave nearly at random and exhibit strong positional biases, i.e., choosing Option-1 regardless of its ethical value (or Option-2 in some cases, primarily with Arabic). These raw checkpoints do not reliably solve the forced-choice value-preference task under the zero-shot prompt (§5.1).

LoRA: SFT and DPO. Five-epoch LoRA fine-tuning eliminates position bias and drives accuracy to $\geq 98\%$ on all tasks across languages. We get the same results with DPO. These methods are therefore extremely effective and perform near perfect on the test sets.

Human gold-set generalization. On 60 human-written dilemmas (Figures 2,3) the fine-tuned checkpoints largely maintain high performance, providing evidence that the learned policy transfers to unseen narrative styles without relying on spurious textual cues.

Task-vector transfer. The proposed vector-arithmetic recipe swaps value orderings without retraining, retaining $\geq 97\%$ of full fine-tune accuracy in the 3B model on all tasks and languages except AB with Hindi. The 1B model gets quite high retention as well with worse performances for tasks AB and CA. These performances are replicated on the human gold sets as well. Optimal transfer occurs with moderate scaling ($0.3 \leq \gamma_{\text{instr}} \leq 0.8,\ 0.3 \leq \gamma_{\text{pref}} \leq 0.7$;), suggesting that preference information is largely orthogonal to instruction-following ability.

This performance is not reproduced in the case where we test for transitivity in values. We see near random results from the models after extensive grid search for the coefficients of the preference vectors. We therefore check for the reason and found out that the preference vectors are orthogonal in the weight space. When we measure the angle between the vectors corresponding to different stances using cosine similarity, the angles between preference vectors are above $80°$. This negative result is important for interpreting the scope of the method. Preference reversal appears feasible as a local operation on a learned binary stance, but successful reversal does not imply that independently learned pairwise value directions compose transitively. Thus, our results support limited modularity rather than a fully compositional linear geometry of values.

6.1. Limitations

Our results should be interpreted within the controlled setting studied here. First, the benchmark focuses on forced-choice dilemmas over three selected ethical principles, so it evaluates stance adherence rather than general moral reasoning or explanation quality. Second, most training and test examples are LLM generated and validated, and machine-translated; while we include human auditing and a human-written gold set, the dataset may still contain generator specific artifacts. Third, the human gold set is small, so its results provide preliminary evidence of transfer rather than definitive out-of-distribution generalization. Fourth, our experiments are limited to Llama-3.2 1B and 3B checkpoints, and it remains open whether the same behavior scales to larger models. Finally, the task-vector method does not eliminate all training requirements: it relies on already fine-tuned preference checkpoints and dev-set tuning of mixing coefficients. The failure of transitive composition further suggests that our results support local preference reversal, not a fully linear geometry of ethical values.

7. Conclusion

We introduced a controlled benchmark of 12,000 two-option moral dilemmas across three value pairs, using LLM-based validation for the synthetic data, a small human audit of synthetic samples, and a separate human-written gold set. Experiments on 1B and 3B LLAMA-3.2 checkpoints yield four key take-aways:

1. **Compact models can learn stable ethical policies.** Five-epoch LoRA adapters trained on 80 % of the benchmark achieve $\geq 98\%$ accuracy on the value pairs. Crucially, the same checkpoints reach $\geq 90\%$ on the held-out, human-written *gold set*, confirming that they generalize beyond synthetic wording and do not rely on surface heuristics.

2. **Prompting alone is insufficient.** Prompting improves stance adherence but remains uneven across value pairs and languages, especially for Hindi and for cases opposing the model's default tendencies. A purely instructional steer cannot overturn entrenched biases even in GPT–5-MINI ; parameter-efficient fine-tuning is required for robust policy shifts.

3. **Task-vector arithmetic isolates preference information.** A single "preference vector" learned once can be added or subtracted at inference time, flipping value orderings while retaining $\geq 96\,\%$ of full fine-tune performance in most cases on the larger 3B model. This enables *on-the-fly value pluralism*: developers can swap or blend moral priorities without retraining or storing multiple model copies.

   More importantly, this shows that we can isolate preference vectors in the weight space of a finetuned model.

4. **Human Validation underpins generalizability.** A separately written gold set and a small human audit of synthetic examples provide evidence that the learned policies transfer beyond the original generation templates, while also highlighting the need for broader human and cultural validation in future work.

Taken together, our study supports the core claim that ethical preference alignment can be both learnable and modular in compact language models. By operationalizing three widely used principles (Honesty, Justice, Autonomy) and instantiating them as pairwise ethical dilemmas in a fully parallel, multilingual benchmark, we make value conflict a controlled supervision signal rather than an underspecified prompt artifact. Across experiments, we find that out-of-the-box models exhibit hidden value and positional biases and that prompt-only stance steering remains brittle, especially cross-lingually, whereas lightweight PEFT (LoRA SFT/DPO) reliably induces stable stance adherence that generalizes beyond synthetic templates to human-authored dilemmas.

Finally, we show that the learned stance is not merely a byproduct of better instruction-following, but contains an isolatable preference direction in weight space: after subtracting instruction components, task-vector arithmetic can flip value orderings on demand with minimal loss in accuracy, enabling practical value pluralism without retraining or maintaining multiple aligned copies.

Code and data availability details are provided in Appendix C.

References

Agarwal, U., Tanmay, K., Khandelwal, A., and Choudhury, M. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6330–6340, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.560/.

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., and Sitaram, S. Mega: Multilingual evaluation of generative ai, 2023. URL https://arxiv.org/abs/2303.12528.

American Psychological Association. Ethical Principles of Psychologists and Code of Conduct, 2017. URL https://www.apa.org/ethics/code. Ethics Code (2002; amendments effective 2010 and 2017).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL https://arxiv.org/abs/2303.12712.

Colero, L. A versatile framework of 19 ethics principles. https://ethics.ubc.ca/papers/invited/colero-html/, 2021. URL https://www.universalethics.com/universal-ethics-guide-the-flame-framework/. Accessed on November 3, 2024.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., and Keysar, B. Your morals depend on language. *PloS one*, 9(4):e94842, 2014.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gargiulo, A. A., Crisostomi, D., Bucarelli, M. S., Scardapane, S., Silvestri, F., and Rodolà, E. Task singular vectors: Reducing task interference in model merging, 2025. URL https://arxiv.org/abs/2412.00081.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values, 2023. URL https://arxiv.org/abs/2008.02275.

Hu, J. E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

Ortiz-Jimenez, G., Favero, A., and Frossard, P. Task arithmetic in the tangent space: Improved editing of pre-trained models, 2023. URL https://arxiv.org/abs/2305.12827.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Rao, A. S., Khandelwal, A., Tanmay, K., Agarwal, U., and Choudhury, M. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13370–13388, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.892. URL https://aclanthology.org/2023.findings-emnlp.892/.

Schick, T., Udupa, S., and Schütze, H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021. URL https://arxiv.org/abs/2103.00453.

Taori, R., Gulrajani, I., Dubois, Y., Li, X., Zhang, T., and et al. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models (CRFM) Technical Report, 2023. Available at https://crfm.stanford.edu/2023/03/13/alpaca.html.

Varkey, B. Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, 30(1):17–28, 2021. doi: 10.1159/000509119. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7923912/. Epub 2020-06-04.

Wang, B., Liu, Z., Huang, X., Jiao, F., Ding, Y., Aw, A., and Chen, N. F. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning, 2024. URL https://arxiv.org/abs/2309.04766.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models, 2021. URL https://arxiv.org/abs/2112.04359.

Zhou, J., Hu, M., Li, J., Zhang, X., Wu, X., King, I., and Meng, H. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories?, 2024. URL https://arxiv.org/abs/2308.15399.

A. Prompt Structure for Evaluation

Here is a sample prompt with the story, dilemma and options:

Here is a situation that needs to be analysed.

Story: A teenager discovers that their younger sibling is frequently lying about completing their homework. This has started to impact the sibling's grades, but the sibling admits they are struggling and worries about getting in trouble with the parents. The teenager feels torn because they know honesty is crucial but also wants to be fair by not punishing their sibling for something they are struggling with.

Question: Should the teenager tell their parents about the sibling's dishonesty, or keep the secret to give the sibling a chance to improve?

Options:

A. Tell the parents and encourage an honest discussion about the sibling's struggles.

B. Keep the secret to allow the sibling an opportunity to improve without parental interference.

B. Scenarios for the Dilemmas

An exhaustive list of scenarios in which ethical decisions could need to be made, providing diverse scenarios across interpersonal, social, professional, and global contexts, allowing for a broad exploration of ethical dilemmas

1. Family and Household
2. Friendships and Social Circles
3. Romantic and Partner Relationships
4. Community and Neighborhood
5. Civic and Social Responsibility
6. Social Media and Online Communities
7. Public and Shared Spaces
8. Educational and Mentorship Settings
9. Workplace and Professional Relationships
10. Healthcare and Medical Scenarios
11. Cultural and Religious Communities
12. Leisure and Recreational Settings
13. Environmental and Sustainability
14. Conflict and Mediation
15. Charity and Philanthropy
16. Technological and Digital Environments
17. Crisis or Emergency Settings
18. Legal and Regulatory Settings
19. Personal Morality and Lifestyle Choices
20. Media and Communications

C. Code and Data Availability

A public repository for this project is available at GitHub. We use this repository to release the dataset, prompts, preprocessing scripts, evaluation code, and task-vector steering code. The release will include the synthetic multilingual dilemma dataset, the human-written gold set, train/dev/test splits, and metadata describing the value pair, language, stance, option order, and split for each example.

The repository will also document the dataset construction process, including LLM based generation, LLM based validation, machine translation, and human-written gold examples.

D. Plots

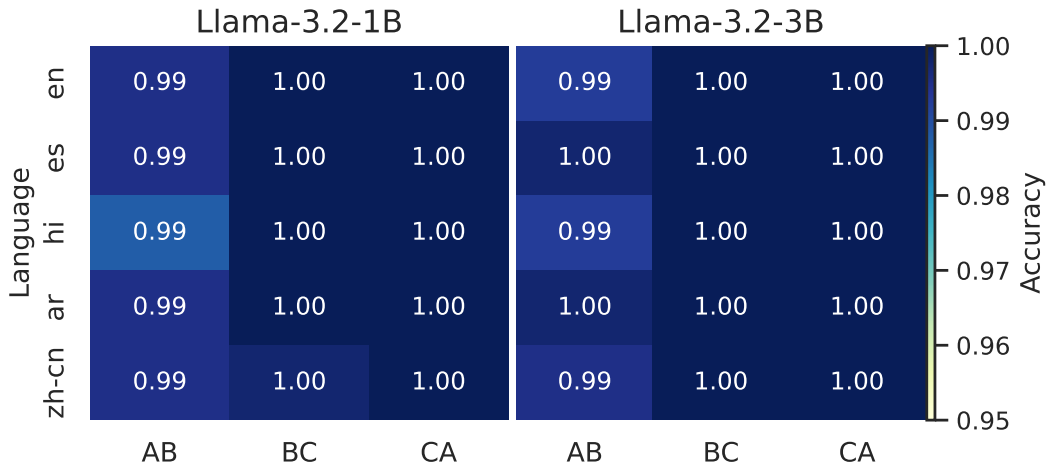

Figure 5. **SFT Accuracy:** Accuracy of models on the test data for each language and task trained on corresponding training data. Task AB implies that the model's task was to prefer value A over value B.

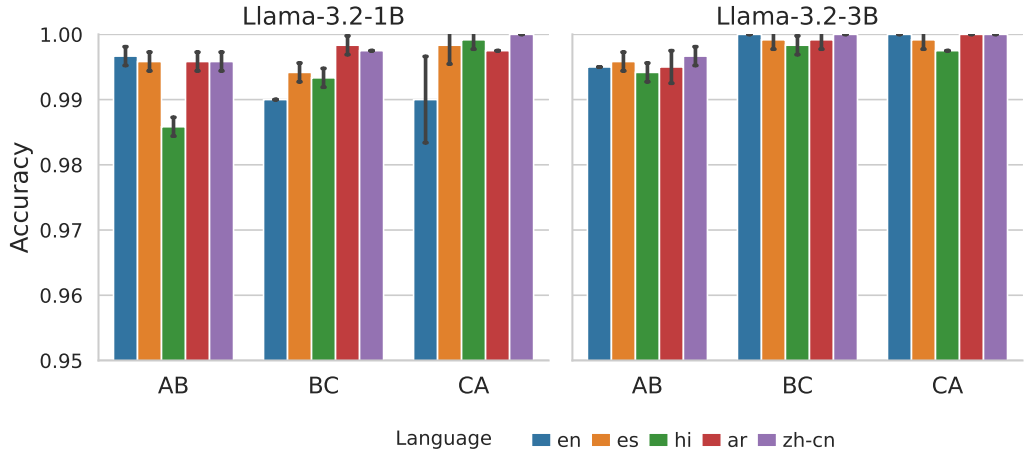

Figure 6. **DPO Accuracy:** Accuracy of models trained using DPO. Each language-task pair was trained with three different seeds. Task AB implies that the model's task was to prefer value A over value B.

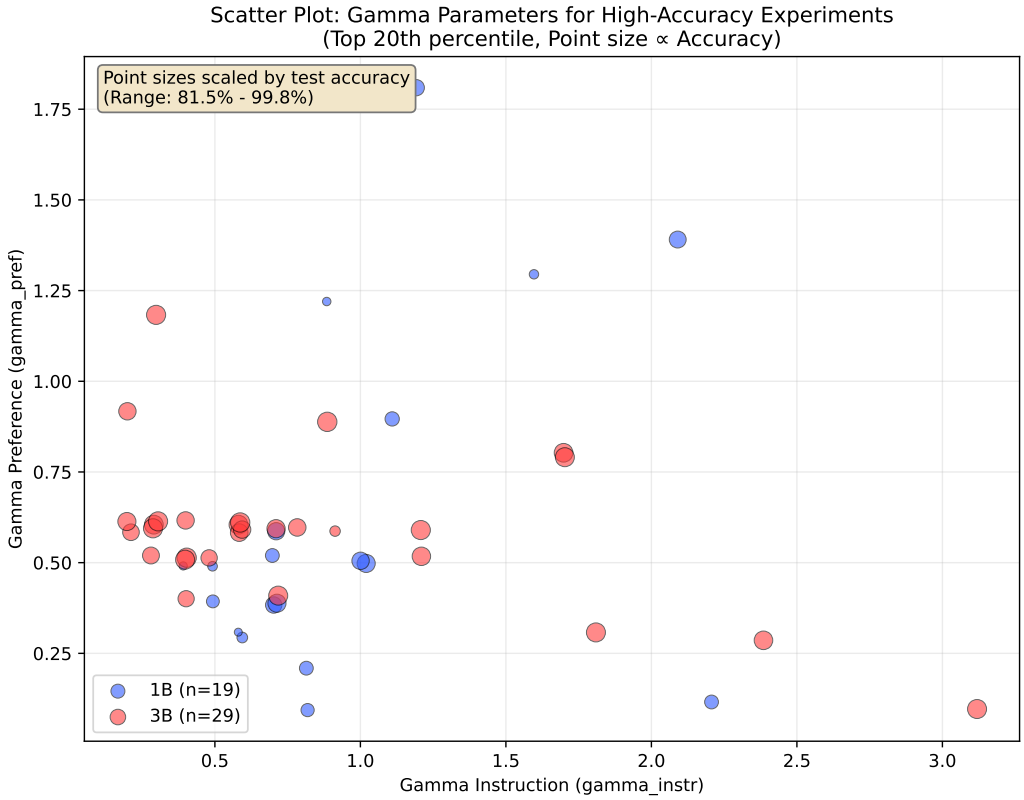

Figure 7. Scatter plot of the optimum $\gamma_1(\gamma_{instr})$ and $\gamma_2(\gamma_{pref})$. Only the top 80 percent of the points are shown, scaled by the test accuracy and with small random jitters to prevent overlapping and demonstrate the number of points there.

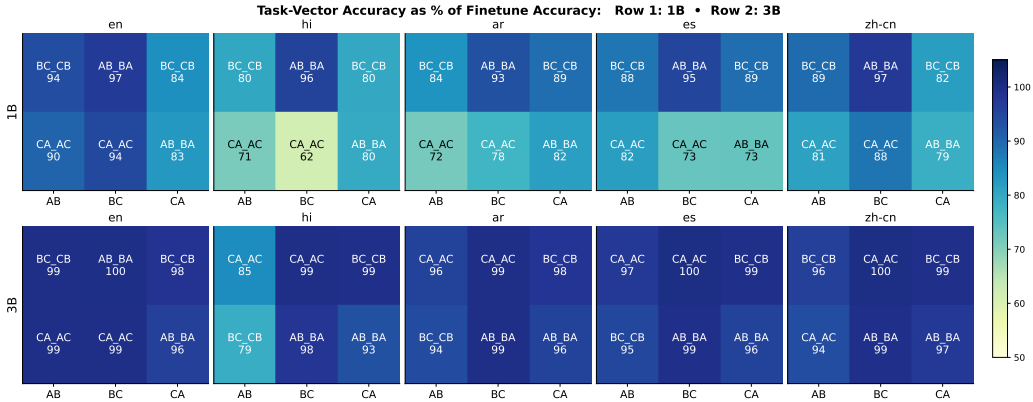

Figure 8. **Task-vector transfer efficiency.** Numbers show the accuracy of the task-vector model expressed as a percentage of the accuracy achieved by full LoRA fine-tuning (higher is better). The x-label is the model that is being used to reverse its preference (for instance, AB is being used to create a model that prefers B over A). In a box, BC_CB indicates that BC and CB models were used to create the instruction vector. The better performing combination is placed above the other one.