# Controlling Cross-Lingual Answer Distributions in Language Models: Enabling Transfer of Factual Preferences

**Lukas Ellinger, Alexander Manev, and Georg Groh**
School for Computation, Information and Technology
Technical University of Munich, Germany
{lukas.ellinger, alexander.manev}@tum.de, grohg@cit.tum.de

## Abstract

Multilingual large language models exhibit systematic differences in their outputs across languages, even when representing the same underlying knowledge. Prior work has primarily focused on evaluating or reducing such inconsistencies. In this work, we instead study whether cross-lingual behavior can be controlled: specifically, whether answer distributions associated with other languages can be expressed under English prompting. To this end, we construct a human-annotated factual dataset and a cultural scenarios dataset, and compare intervention methods including persona prompting, activation steering, and preference-based fine-tuning. We evaluate how these methods affect answer distributions and their generalization to culturally grounded settings. Our results show that answer distributions can be systematically shifted toward those observed in other languages, with persona prompting consistently outperforming more complex intervention methods.

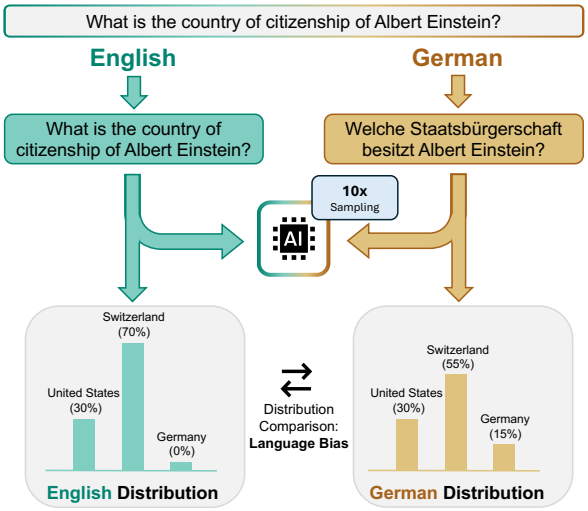


Figure 1: Cross-lingual variation illustrated as a shift in answer distributions. For the same question, we generate multiple responses per prompt (10 samples) and estimate answer probabilities based on frequency. While all answers are factually plausible, their output probability differs depending on the prompt language.

## 1 Introduction

Multilingual large language models (LLMs) are designed to operate across languages and are often expected to exhibit consistent behavior for the same underlying task. However, prior work has shown that their outputs can vary systematically depending on the prompt language, even when the underlying task remains unchanged (Shafayat et al., 2024; Shcharbakova et al., 2025; Wang et al., 2025a).

Importantly, these differences affect not only which outputs are generated, but also how likely those outputs are. Figure 1 illustrates this effect for the question about the country of citizenship of Albert Einstein. Under English prompting, the model generates only "Switzerland" and "United States," whereas German prompting additionally yields "Germany." The resulting answer distributions therefore also differ in their relative frequencies, e.g., "Switzerland" accounts for roughly 70% of English but only 55% of German generations.

Most prior work has focused on measuring such cross-lingual variation (Roh et al., 2025; Wang et al., 2025a; Jiang et al., 2020; Kassner et al., 2021) or reducing it via alignment methods (Agarwal et al., 2025; Bu et al., 2025).

However, such variation is not always undesirable. In many cases, it reflects meaningful differences in how information is expressed across languages, making it desirable to control rather than eliminate it. For example, users may wish to access knowledge as it is typically expressed in another linguistic or cultural context while interacting in their own language (Goldman et al., 2025). Different languages tend to emphasize different aspects or contributors of the same fact, reflecting cultural or educational narratives—even when all answers are factually correct (Kim and Kim, 2025; Calvo-Bartolomé et al., 2025).

Thus, we take a complementary perspective.



*Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026)*, pages 35–49

Rather than focusing on whether a model produces the "correct'' answer, we consider the distribution over answers, estimated via repeated sampling. We refer to this distribution as the model's *factual preferences*, capturing which answers a model tends to favor. We do not interpret these distributions as explicit beliefs of the model, but as empirical generation tendencies under repeated sampling. This shifts the focus from individual predictions to answer distributions and raises a central question: can factual preferences associated with one language be expressed under a different, fixed prompt language?

To investigate this, we fix the prompt language to English and test whether answer distributions observed in German, Spanish, and Bulgarian can be reproduced.

Our contributions are as follows:

- We study whether answer distributions from German, Spanish, and Bulgarian can be expressed under English prompting, enabling cross-lingual transfer of factual preferences.
- We construct two datasets: a human-validated factual dataset for measuring distribution shifts, and a cultural scenarios dataset for testing generalization to culturally grounded preferences.
- We compare intervention methods, including prompting, activation steering, and preference-based fine-tuning (DPO).
- We demonstrate that answer distributions can be shifted toward those of other languages under fixed prompting, with simple prompting outperforming more complex methods.
- We release code and datasets to support reproducibility and enable their usage for further research.

## 2 Background and Related Work

We situate our work within three lines of research: cross-lingual inconsistencies in LLMs, evaluation of cultural and multilingual behavior, and methods for aligning or steering model outputs.

**Cross-lingual Inconsistencies in LLMs.** Although the Semantic Hub Hypothesis posits a shared, language-agnostic representation space, scaffolded by the model's dominant training language (typically English) (Wu et al., 2025), empirical evidence shows that LLM outputs vary systematically across languages (Shafayat et al., 2024; Shcharbakova et al., 2025). Recent work further indicates that factual knowledge utilization depends on the language used during internal reasoning, with performance improving when the language of thought aligns with the source of knowledge (Kang and Kim, 2025).

These differences arise from both how knowledge is stored and how it is generated. Factual knowledge is not uniformly shared: models encode a mixture of language-independent, cross-lingually shared, and transferred knowledge (Zhao et al., 2024). In addition, while knowledge is encoded in a largely language-independent space, models transition to language-specific representations in the final layers, where decoding introduces factual and cultural variation depending on the prompt language (Wang et al., 2025a).

**Evaluation of cultural and cross-lingual behavior.** To assess whether models capture culturally grounded knowledge beyond surface-level translation, several benchmarks have been proposed. Datasets such as FORK (Palta and Rudinger, 2023), NormAd (Rao et al., 2025), and CulturalBench (Chiu et al., 2025) evaluate cultural alignment, typically within English-only settings.

Complementary to this, cross-lingual evaluation benchmarks such as XLQA (Roh et al., 2025), KLAR (Wang et al., 2025a), mLAMA (Kassner et al., 2021), and X-FACTR (Jiang et al., 2020) assess how model outputs vary across languages.

These approaches focus on evaluating model behavior and improving consistency. In contrast, we focus on controlling it: we investigate whether factual answer distributions from one language can be expressed under a fixed prompt language, enabling cross-lingual transfer of factual preferences.

**Alignment and Intervention Methods.** Efforts to mitigate cross-lingual inconsistencies often rely on parameter updates, including fine-tuning for multilingual consistency (Agarwal et al., 2025), representation-level alignment (Bu et al., 2025), and cross-lingual transfer of supervision signals (Liu et al., 2025). As our setting naturally yields preference pairs, we consider Direct Preference Optimization (DPO) (Rafailov et al., 2023) as a parameter-based approach for aligning response distributions across languages.

Complementary to these approaches, inference-time control techniques have been explored. Activation engineering methods, such as representa-

tion steering (Zou et al., 2025; Turner et al., 2024) and Contrastive Activation Addition (CAA) (Rimsky et al., 2024), enable targeted manipulation of model activations without updating parameters. In addition, prompt-based methods, including persona prompting, provide a lightweight alternative for influencing model outputs at inference time. Prior work shows that such approaches can shift value alignment and cultural framing (Wang et al., 2025b), although their effects can be inconsistent and may interfere with factual accuracy (Zheng et al., 2024; Lutz et al., 2025).

While these techniques have been used to steer behaviors such as truthfulness or stylistic preferences, their use for shifting factual preferences across languages remains underexplored.

## 3 Methodology

Our goal is to evaluate whether factual answer distributions associated with one language can be expressed under a fixed prompt language. Concretely, given a question and a target language, we estimate the distribution over plausible answers by repeated sampling, and treat this empirical distribution as the model's factual preferences in that language. We then investigate how closely these preferences can be matched under English prompting.

To this end, we construct two datasets: a factual dataset (Section 3.1) and a cultural scenarios dataset (Section 3.2). We then evaluate answer distributions using distributional metrics (Section 3.3) to analyze the effect of our intervention methods (Section 3.4).

### 3.1 Fact Dataset

To study cross-lingual factual preferences, we require a dataset that allows for controlled comparison of answers across languages.

We take inspiration from KLAR (Wang et al., 2025a), which provides parallel factual triples across languages. However, manual inspection revealed several issues, including incorrect ground-truth objects and translation errors. While KLAR aims to avoid questions with multiple correct answers, we observed that such cases still occur in practice.

We therefore curate our dataset manually, building on the general setup of KLAR while revising and expanding it. All objects are verified against Wikidata, using Wikipedia as a fallback. We explicitly retain and expand cases with multiple valid answers (e.g., individuals with multiple citizenships), as these are central to our analysis of answer distributions.

| Relation | Total | Unq | Multi |
|---|---|---|---|
| capital | 289 | 262 | 27 |
| city of origin | 26 | 26 | 0 |
| country of citizenship | 125 | 81 | 44 |
| country of origin | 36 | 30 | 6 |
| languages | 255 | 128 | 127 |
| occupation | 48 | 6 | 42 |
| religion | 111 | 79 | 32 |
| **Total** | 890 | 612 | 278 |

Table 1: Dataset statistics showing total, unique, and multiple subject occurrences per relation.

We construct the dataset in four languages where we have native or near-native proficiency: English, German, Spanish, and Bulgarian. We first curate the dataset in English and then translate the entries into the target languages using the DeepL API[1]. All translations are subsequently manually reviewed and aligned with the target-language Wikipedia to ensure semantic accuracy.

We focus on seven relations: *country of citizenship*, *languages*, *religion*, *occupation*, *capital*, *city of origin*, and *country of origin*. These relations cover both well-defined factual knowledge and cases prone to cultural or geopolitical variation.

For each relation, we construct open-ended question formats rather than the popular multiple-choice format, which is prone to positional bias (Li et al., 2024). To mitigate prompt sensitivity (Errica et al., 2025), we design five prompt variants per relation.

The final dataset consists of 890 questions per language. Of these, 278 (31.24%) have multiple correct answers. Table 1 shows the distribution across relations. For dataset details, see Appendix B.

### 3.2 Cultural Scenarios Dataset

To evaluate whether the interventions generalize beyond the factual dataset, we construct a cultural scenarios dataset targeting broader cultural transfer.

To the best of our knowledge, no existing resource covers our target languages and setup. We therefore construct the dataset with the assistance of Gemini 3.1 Pro and manually validate all instances, following prior work that has used LLMs for cultural data generation (Wang et al., 2024). We create 50 scenarios per target language (German, Spanish, Bulgarian), for a total of 150 samples. The full

[1]https://www.deepl.com/en/products/api

prompt is provided in Appendix C.

Each instance consists of an English question describing a culturally grounded scenario, with two answer options: one reflecting an English-centric norm and the other reflecting the target-language norm. For example, to probe Spanish cultural norms regarding dinner times, the English-centric option is "6:30 p.m. or 7:00 p.m.", while the target-centric option is "9:30 p.m. or 10:00 p.m.".

### 3.3 Evaluation Approach

We conduct all experiments using Gemma 3 12B Instruct (Gemma Team et al., 2025), a mid-sized instruction-tuned model that provides a realistic and computationally tractable evaluation setting.

Exact versions of all models used for generation and evaluation are provided in Appendix A.

#### 3.3.1 Fact Dataset

For each dataset entry, we prompt the model multiple times, each time requiring a single explicit answer. For each generation, we randomly select one of the predefined prompt variants. This yields a set of generations from which we estimate the answer distribution. We sample 10 generations at temperatures 0.8 and 1.2, resulting in distributions similar to those shown in Figure 1.

To enable the comparison of outputs across languages, we map model outputs to a canonical English form using an LLM-based extraction step (GPT-5.1).

Based on these canonicalized outputs, we distinguish between *contrastive* instances, where the sets of canonical answers differ between English and the target language, and *aligned* instances, where both languages yield the same answer set but may differ in the distributions over these answers. For evaluation and training, we split the dataset into train, validation, and test sets at the subject level. Contrast pairs for CAA and DPO are constructed only from contrastive instances to study explicit answer-level shifts, while aligned instances are used to evaluate more subtle distributional preference shifts in the absence of answer-level differences.

We compare answer distributions using the Jensen–Shannon distance (JSD), i.e., the square root of the Jensen–Shannon divergence (base 2) (Lin, 1991). JSD ranges from 0 to 1, where lower values indicate more similar distributions.

In addition, we report the Gold Match Rate (GMR), which measures the proportion of generations that match the gold answer after canonicalization. Higher GMR indicates stronger agreement with the reference answers.

We provide details, including the prompts for generation and extraction, in Appendix D.

**Persona System Prompt**

```
Adopt the perspective of a typical person
born, raised, and currently living in
<country>. Your worldview, background
knowledge, and behavioral expectations are
entirely grounded in <demonym> culture,
education, and society. When asked any
question-even in English-rely on what a
typical <demonym> would consider natural,
correct, and factual.

+ Default System Prompt (see Appendix D)
```

Figure 2: Persona-based system prompt used to condition the model on a target language perspective. The model is instructed to adopt the viewpoint of a typical individual from the specified country when generating answers.

#### 3.3.2 Cultural Scenarios Dataset

The cultural scenarios dataset is evaluated in a binary-choice setting, where the model selects the scenario associated with the target language. We compute accuracy based on the selected option.

To mitigate positional bias in multiple-choice settings (Li et al., 2024), we randomize the order of answer choices. We evaluate under three temperature settings (0, 0.8, and 1.2). For the stochastic settings (0.8 and 1.2), we sample ten generations per instance.

### 3.4 Interventions

We evaluate three intervention methods for controlling answer distributions: persona prompting, activation steering via Contrastive Activation Addition (CAA), and Direct Preference Optimization (DPO). These methods differ in how they influence model behavior, ranging from inference-time prompting and representation-level manipulation to parameter updates.

**Persona Prompting** Persona prompting conditions the model on a target cultural perspective by modifying the system prompt. Specifically, we instruct the model to adopt the viewpoint of a typical individual from the target country (see Figure 2). While the input question remains in English, the model is guided to generate answers according to what would be considered natural and appropriate

within the target cultural context. This intervention operates entirely at inference time and does not require any parameter updates.

**CAA Steering** Unlike prompting, which operates purely through instruction context, Contrastive Activation Addition (CAA) (Rimsky et al., 2024) intervenes directly in the model's hidden states by adding a directional vector to the residual stream during inference. This vector is derived from contrast pairs and represents a shift toward a target preference in activation space. During generation, the vector is injected at a selected model layer and scaled by a steering multiplier controlling the intervention strength.

We construct these contrast pairs from the *contrastive* instances defined in Section 3.1. For each instance, we form pairs via the cross-product between target-language answers and English answers, excluding identical pairs. Each pair contrasts a target-language-consistent answer with an English-consistent alternative, providing an explicit signal for shifting answer preferences.

We sweep over all model layers and steering multipliers on the validation split and select configurations based on JSD to the target-language distribution. Consistent with prior work suggesting that later layers increasingly reflect language-specific decoding behavior (Wang et al., 2025a), we observe that the strongest steering effects consistently emerge in late layers.

Further details on data construction and implementation are provided in Appendix E.

**Direct Preference Optimization (DPO)** We apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) as a parameter-based intervention. We use the same contrast pairs as in CAA steering, treating target-language answers as *chosen* and English-preferred answers as *rejected*.

This allows us to test whether cross-lingual answer distributions can be internalized through parameter updates, rather than induced during inference. We provide further details in Appendix F.

## 4 Results

We analyze cross-lingual differences in answer distributions and evaluate whether these can be shifted through targeted interventions. First, we quantify cross-lingual differences. We then evaluate whether interventions can shift answer distributions toward target-language behavior on both *contrastive* instances and *aligned* instances. Finally, we assess whether these intervention effects transfer to culturally grounded scenarios.

| | EN | DE | ES |
|---|---|---|---|
| **DE** | 0.14 | | |
| **ES** | 0.18 | 0.19 | |
| **BG** | 0.21 | 0.21 | 0.24 |

Table 2: Pairwise JSD between answer distributions across languages (darker indicates greater divergence).

| | EN | DE | ES | BG |
|---|---|---|---|---|
| **GMR** | 0.88 | 0.87 | 0.88 | 0.82 |

Table 3: Overall Gold Match Rate (GMR) per language.

### 4.1 Cross-Lingual Distribution Differences

Table 2 shows pairwise JSD between answer distributions across languages. We observe consistent divergence across all language pairs, indicating that model outputs depend on the input language rather than reflecting a single language-invariant distribution. The lowest divergence is observed between English and German, while the highest divergence occurs for pairs involving Bulgarian.

In addition to the observed distributional differences, we also observe differences in overall correctness across languages, as shown by GMR in Table 3. While English, German, and Spanish achieve comparable performance (0.87–0.88), Bulgarian shows a lower overall match rate (0.82). We provide a per-relation breakdown for English-to-target differences in Appendix G.

Motivated by these findings, we evaluate whether these distributions can be controlled by shifting English outputs toward a selected target language.

### 4.2 Distribution Shifts on Contrastive Instances

Table 4 reports JSD between answer distributions under English prompting and target-language distributions on contrastive instances. In addition, Table 5 reports GMR to analyze how these shifts relate to agreement with gold answers.

Interventions significantly reduce divergence for Bulgarian and German, with the largest gains observed for German. In contrast, Spanish shows smaller, non-significant improvements despite its lowest baseline divergence.

Among all methods, prompting consistently

| Method | BG | DE | ES |
|---|---|---|---|
| Baseline | 0.58 | 0.52 | 0.47 |
| Prompt | **0.49** (-0.09)* | **0.40** (-0.13)* | **0.41** (-0.06) |
| Steered | 0.50 (-0.08)* | 0.44 (-0.09)* | 0.44 (-0.03) |
| DPO | 0.50 (-0.08)* | 0.46 (-0.07)* | 0.46 (+0.01) |

Table 4: JSD between answer distributions under English prompting and target language distributions on the test split of the factual dataset (lower = better alignment). Parentheses show changes from baseline; * indicates significant changes ($p < 0.05$). Best results per language are shown in bold.

| Method | BG | DE | ES |
|---|---|---|---|
| *Source* | 0.77 | 0.69 | 0.63 |
| *Target* | 0.65 | 0.60 | 0.80 |
| Prompt | **0.72** (-0.05) | **0.68** (-0.01) | **0.64** (+0.01) |
| Steered | 0.71 (-0.06) | 0.61 (-0.08) | 0.63 (-0.00) |
| DPO | 0.71 (-0.06) | 0.62 (-0.07) | 0.60 (-0.04) |

Table 5: GMR across languages and intervention methods. Parentheses indicate absolute changes relative to the English source baseline. Higher values indicate stronger agreement with gold answers. Best intervention result per language is shown in bold.

| Method | BG | DE | ES |
|---|---|---|---|
| Prompt | 19.3 (-1.0) | 26.2 (-0.1) | 30.5 (-0.4) |
| Steered | 32.7 (-0.4) | 22.9 (-0.4) | 29.1 (-0.3) |
| DPO | 7.9 (-0.3) | 17.7 (-0.2) | 18.6 (-0.2) |

Table 6: JSD ($\times 10^{-3}$) between the intervention-induced answer distributions and the target-language distributions on aligned instances. The value in parentheses denotes $\text{JSD}_{\text{target}} - \text{JSD}_{\text{source}}$, where negative values indicate that the resulting distribution is closer to the target-language distribution than to the original English distribution.

achieves the strongest distribution shifts toward the target language, yielding the lowest JSD values across target languages. Steering also improves target-language similarity across all languages, but remains consistently weaker than prompting. DPO shows a less stable pattern: while comparable to Steering for Bulgarian and German, it slightly worsens target-language similarity for Spanish.

Comparing Table 4 and Table 5, Prompting achieves the strongest shifts toward the target distributions while consistently preserving the highest GMR across interventions, particularly for German. Steering and DPO yield weaker distributional shifts and larger reductions in GMR.

Spanish differs from Bulgarian and German in that the target-language distribution achieves a higher GMR than the English source distribution. Accordingly, shifting the distribution toward the target decreases GMR for Bulgarian and German, but can increase it for Spanish. Prompting most closely follows this pattern, slightly improving GMR for Spanish while preserving higher GMR than Steering and DPO for Bulgarian and German.

Overall, these results show that answer distributions can be shifted toward a target language while largely preserving agreement with gold answers. Notably, prompting not only produces the strongest overall shifts toward the target distributions but also maintains the highest GMR across intervention methods.

### 4.3 Distribution Shifts on Aligned Instances

In Table 6, we analyze aligned instances, i.e., subjects for which the English and target-language answer sets coincide and only the distribution over these answers may differ.

Across all languages and intervention methods, both JSD and $\Delta$ values remain very small. This indicates that, when both languages share the same answer space, the resulting distributions remain comparatively similar overall. Nevertheless, all methods consistently move the resulting distributions closer to the target-language distributions than to the original English distributions, as reflected by the negative $\Delta$ values.

Among all methods, DPO consistently produces the smallest JSD values across target languages, yielding distributions that are closest to the target-language distributions while also remaining comparatively close to the original English distributions. In contrast, Prompting and Steering produce larger JSD values, indicating broader distributional shifts that move the resulting distributions further away from both the original English and target-language distributions than DPO.

GMR remains nearly unchanged across interventions and languages, which is expected since aligned instances share the same answer sets across languages.

### 4.4 Generalization to Cultural Scenarios

We evaluate in Table 7 whether the observed distribution shifts transfer beyond our factual dataset to culturally grounded scenarios.

Prompting yields large gains across all target languages, improving accuracy over the baseline (+0.40 to +0.67). In contrast, steering and DPO show little to no improvement, with performance

| Method | BG | DE | ES |
|---|---|---|---|
| Baseline | 0.46 | 0.32 | 0.20 |
| Prompt | **0.86** (+0.40) | **0.86** (+0.54) | **0.87** (+0.67) |
| Steered | 0.46 (+0.00) | 0.32 (+0.00) | 0.21 (+0.01) |
| DPO | 0.45 (-0.01) | 0.32 (+0.00) | 0.21 (+0.01) |

Table 7: Accuracy on the cultural scenarios dataset under English prompting. Values are averaged across temperatures ($T \in \{0, 0.8, 1.2\}$). Parentheses denote changes relative to the baseline. Best results per language are shown in bold.

remaining close to the baseline across languages.

Additionally, the observed patterns remain nearly unchanged across temperature settings, indicating that the generalization behavior reflects stable intervention effects rather than sampling variability.

These results indicate that persona prompting not only shifts answer distributions on factual instances but also generalizes to broader, culturally grounded settings. More complex interventions fail to induce such transfer, suggesting that they primarily capture dataset-specific patterns rather than broader target-language-conditioned behavior.

## 5 Discussion

Our results show that both answer distributions and correctness depend on the language used in the prompt. Bulgarian is consistently furthest away from the other languages in terms of JSD and also exhibits the lowest overall GMR. This may reflect weaker multilingual representation quality due to limited training data availability (Qin et al., 2025).

However, cross-lingual variation is not limited to lower-resource settings and cannot be reduced to differences in correctness alone. Even for comparatively high-resource languages such as English, German, and Spanish, which achieve highly similar GMRs, we still observe substantial differences in answer distributions. Thus, multilingual models often produce equally plausible and factually correct answers across languages while differing in which answers they prefer to generate.

Such variation is not necessarily undesirable. Different languages and cultural contexts may emphasize different aspects or contributors of the same underlying fact, even when all answers remain factually plausible (Kim and Kim, 2025; Calvo-Bartolomé et al., 2025). From this perspective, cross-lingual variation becomes not only a consistency problem, but also a controllable representational property. Our results suggest that multilingual models encode multiple language-conditioned answer distributions that can, at least partially, be transferred under a fixed prompt language.

**Distribution Shift.** We separate our analysis into contrastive and aligned instances. Contrastive instances differ in their answer sets between source and target language, whereas aligned instances share the same answer set but differ in the distributions over these answers. This distinction allows us to separately study explicit answer-level shifts and more subtle distributional preference changes.

On contrastive instances, the interventions show that cross-lingual answer distributions can be shifted toward a target language, although the effect differs across languages and methods. Prompting achieves the strongest and most consistent shift across target languages.

Interestingly, merely changing the contextual framing of the prompt is sufficient to substantially shift answer probabilities under a fixed prompt language, without any explicit alignment training. This suggests that multilingual models already encode language-conditioned answer preferences that can be activated through prompting, whereas Steering and DPO depend more strongly on the factual alignment distributions used during training.

The GMR results further show that distribution matching is not equivalent to maximizing correctness. Since source and target distributions can differ in GMR, shifting toward the target distribution may also shift GMR toward the target-language value. This pattern is visible for Bulgarian and German, where target-language GMR is lower than English, and for Spanish, where it is higher. Prompting most closely follows this target-directed behavior. It not only achieves the strongest distributional alignment, but also introduces the least degradation in agreement with gold answers. This highlights that prompting achieves stronger target-language matching with fewer side effects on overall factual answer behavior than Steering and DPO.

Aligned instances exhibit a different behavior pattern. Since the answer sets are identical across languages, overall JSD values remain very small across all interventions. Nevertheless, all methods still shift distributions slightly toward the target language despite the absence of explicit answer-level differences between source and target distributions. This indicates that the interventions capture broader distributional tendencies beyond merely favoring target-specific answers.

At the same time, the interventions differ in how strongly they perturb the original distributions. DPO consistently remains closest to both the source and target distributions, whereas Prompting and Steering induce somewhat larger shifts while still remaining directionally closer to the target language. One possible explanation is that the stronger shifts produced by Prompting and Steering also increase generations outside the dominant answer support shared across both language distributions.

**Generalization beyond factual distributions.** The cultural scenarios evaluation exhibits a substantially different pattern from the factual distribution alignment experiments. Prompting is the only intervention that consistently produces large improvements over the baseline across target languages, whereas Steering and DPO remain almost unchanged.

This further strengthens Prompting as the overall strongest intervention in our evaluation setting. It seems to induce broader, target-language-consistent behavior through contextual framing. In contrast, Steering and DPO appear more closely coupled to the factual alignment distributions used during training.

**Future Work.** An important direction for future work is understanding how these language-conditioned distributions emerge during multilingual training and whether they can be manipulated more systematically. In particular, it remains unclear whether the observed effects primarily reflect cultural framing, training-data imbalance, retrieval dynamics, or language-specific decoding behavior (Wang et al., 2025a). Extending the analysis to additional model families and lower-resource languages may further clarify how multilingual representation structure shapes factual preference distributions.

## 6 Conclusion

In this work, we investigated whether cross-lingual answer distributions can be controlled under a fixed prompt language. Rather than viewing multilingual variation solely as a problem of inconsistency, we studied whether language-conditioned answer distributions can be transferred and expressed under English prompting.

Across all evaluated interventions, simple persona prompting consistently produced the strongest alignment with target-language distributions while also preserving the highest agreement with gold answers and generalizing most effectively to culturally grounded scenarios.

More broadly, our findings point toward a complementary perspective on multilingual alignment: rather than solely minimizing cross-lingual variation, future systems may benefit from enabling controllable access to language-conditioned representational tendencies encoded within the model.

## Limitations

**Model Scope and Generalization.** All experiments were conducted using Gemma 3 12B Instruct. While this model provides a computationally tractable evaluation setting, the observed factual preference distributions and intervention behaviors may depend on model-specific properties. It therefore remains unclear to what extent the results generalize to substantially larger frontier models or to smaller and less capable models.

**Language Coverage.** Our experiments focus on three target languages: German, Spanish, and Bulgarian. These languages were selected because they are covered by native or near-native speaker proficiency within the author team, enabling reliable manual curation and validation of the dataset. However, they do not represent the full diversity of multilingual knowledge representations, and other languages may exhibit qualitatively different preference structures or steering behavior.

**LLM-Based Evaluation Pipeline.** Parts of the evaluation pipeline rely on GPT-5.1-based answer extraction and canonicalization. While this substantially improves robustness compared to exact string matching, occasional normalization errors in ambiguous cases cannot be ruled out entirely. Since exhaustive manual verification was infeasible, minor inaccuracies in the reported metrics may remain.

**Cultural Scenario Evaluation.** The cultural scenarios dataset provides a controlled setting for evaluating whether intervention effects transfer beyond factual QA. However, it captures only a limited subset of culturally grounded behaviors and should therefore be interpreted as a proxy evaluation rather than a comprehensive measure of cultural alignment. The scenarios intentionally use high-signal cultural contrasts to make representational shifts in LLM behavior observable and are not intended as sociological modeling or normative claims about cultures.

## Acknowledgments

We gratefully acknowledge Dino Hromic and Jaeyoung Yoo for their contributions to the curation of the Fact Dataset as part of their lab course project.

## References

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. 2025. Aligning LLMs for multilingual consistency in enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 117–137, Suzhou (China). Association for Computational Linguistics.

Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. 2025. AlignX: Advancing multilingual large language models with multilingual representation alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6460–6489, Suzhou, China. Association for Computational Linguistics.

Lorena Calvo-Bartolomé, Valérie Aldana, Karla Cantarero, Alonso Madroñal de Mesa, Jerónimo Arenas-García, and Jordan Lee Boyd-Graber. 2025. Discrepancy detection at the data level: Toward consistent multilingual question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22013–22054, Suzhou, China. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. 2025. What did I do wrong? quantifying LLMs' sensitivity and consistency to prompt engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *Preprint*, arXiv:2502.21228.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Eojin Kang and Juae Kim. 2025. When language shapes thought: Cross-lingual transfer of factual knowledge in question answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 4868–4873, New York, NY, USA. Association for Computing Machinery.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Sean Kim and Hyuhng Joon Kim. 2025. A dual-layered evaluation of geopolitical and cultural bias in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 580–595, Vienna, Austria. Association for Computational Linguistics.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Weihao Liu, Ning Wu, Wenbiao Ding, Shining Liang, Ming Gong, and Dongmei Zhang. 2025. Selected languages are all you need for cross-lingual truthfulness transfer. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8963–8978, Abu Dhabi, UAE. Association for Computational Linguistics.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models.

In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 2373–2403. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. XLQA: A benchmark for locale-aware multilingual open-domain question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28809–28821, Suzhou, China. Association for Computational Linguistics.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-FAct: Assessing factuality of multilingual LLMs using FActscore. In *First Conference on Language Modeling*.

Hanna Shcharbakova, Tatiana Anikina, Natalia Skachkova, and Josef van Genabith. 2025. Cross-lingual fact verification: Analyzing LLM performance patterns across languages. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1137–1147, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering Language Models with Activation Engineering. *Preprint*, arXiv:2308.10248.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. 2025b. Multilingual prompting for improving LLM generation diversity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6367–6389, Suzhou, China. Association for Computational Linguistics.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

## A Model Access

To support reproducibility, Table 8 lists all models used in this paper, including their names, exact versions, and access providers.

## B Fact Dataset

Table 9 reports the number of contrastive instances per relation and language across the train, validation, and test splits. Since contrastive instances are constructed independently for each target language, the resulting dataset sizes differ between languages.

To avoid subject leakage, the train, validation, and test splits are constructed at the subject level, ensuring that each subject does not appear in multiple splits.

Table 10 summarizes the number of aligned instances per relation and language. Since these instances do not contain cross-lingual distribution shifts, they do not provide meaningful supervision signals for intervention training and are therefore excluded from DPO and CAA training. Similar to the contrastive split, we observe substantial variation across relations, with *capital* containing the largest number of aligned examples across all languages. This is expected, as this relation is comparatively constrained and leaves little room for multiple plausible correct answers or culturally dependent variation.

## C Cultural Scenarios Dataset Generation

Figure 3 shows the prompt used to generate the cultural scenarios dataset with Gemini 3.1 Pro. Following prior work on LLM-assisted cultural data generation (Wang et al., 2024), we use an AI-in-the-loop setup in which scenarios are first generated and subsequently validated by human annotators before inclusion in the dataset.

Answer ordering is randomized to avoid positional biases between English-centric and target-centric responses.

## D Evaluation

Figure 4 shows the default system prompt used for generations across all experiments. To ensure that generations remain comparable across languages and intervention methods, the prompt enforces single-answer responses without additional explanation or reasoning.

Figure 5 presents one representative English question template for each evaluated relation. These templates define the semantic structure of the factual queries used throughout the dataset.

Figure 6 shows all five prompt variants for the *capital* relation across English, German, Spanish, and Bulgarian. During generation, prompts are sampled randomly from these paraphrased variants to reduce dependence on a single wording and to estimate answer distributions across semantically equivalent formulations.

## E CAA

### E.1 Contrast Pair Construction

CAA steering vectors are constructed from the *contrastive* instances defined in Section 3.1. For each instance, we form contrast pairs by combining target-language-consistent answers with English-consistent alternatives while excluding identical pairs.

To ensure stable and directly comparable activations, we formulate contrast pairs as binary-choice prompts with single-token outputs ("A" or "B") rather than open-ended textual completions. This avoids variability from multi-token generation and allows steering vectors to be computed from a single answer token.

A representative contrast pair is shown below:

**Prompt shown to the model:**

Which country is Albert Einstein a citizen of?

Choices:
A. United States
B. Germany

**Reference labels used for training:**

*positive:* B
*negative:* A

Each pair contrasts a target-language-preferred answer with an English-preferred alternative, thereby defining an indirect steering signal through the contextualized answer tokens ("A" vs "B").

### E.2 Steering Vector Computation

We compute steering vectors following the standard CAA formulation of Rimsky et al. (2024). For each contrast pair, we run the prompt twice through the model: once with the positive continuation and once with the negative continuation.

At layer $L$, we extract the hidden-state activation corresponding to the answer token and compute the difference between the positive and negative activations. The steering vector is obtained by averaging these difference vectors across all contrast pairs. All vectors are L2-normalized before inference.

| Name | Version | Access Provider |
|---|---|---|
| Gemma 3 12B Instruct | N/A | locally |
| Gemini 3.1 Pro | N/A | Gemini API |
| GPT-5.1 | gpt-5.1-2025-11-13 | OpenAI API |

Table 8: Specific model versions used in our experiments. For each model we provide the exact version and the access provider.

| | Train | | | Validation | | | Test | | |
|---|---|---|---|---|---|---|---|---|---|
| **Relation** | **BG** | **DE** | **ES** | **BG** | **DE** | **ES** | **BG** | **DE** | **ES** |
| capital | 38 | 21 | 27 | 4 | 2 | 3 | 4 | 2 | 3 |
| city of origin | 14 | 11 | 10 | 1 | 1 | 1 | 1 | 1 | 1 |
| country of citizenship | 41 | 28 | 30 | 5 | 3 | 3 | 5 | 3 | 3 |
| country of origin | 10 | 9 | 7 | 1 | 1 | 1 | 1 | 1 | 1 |
| languages | 80 | 63 | 65 | 9 | 7 | 8 | 9 | 7 | 8 |
| occupation | 28 | 19 | 16 | 3 | 2 | 2 | 3 | 2 | 2 |
| religion | 48 | 56 | 46 | 6 | 6 | 5 | 6 | 6 | 5 |
| **Total** | 259 | 207 | 201 | 29 | 22 | 23 | 29 | 22 | 23 |

Table 9: Number of contrastive instances per relation and language across the train, validation, and test splits. Contrastive instances correspond to subject–relation pairs whose answers differ between English and the respective target language.

| **Relation** | **BG** | **DE** | **ES** |
|---|---|---|---|
| capital | 243 | 264 | 256 |
| city of origin | 10 | 13 | 14 |
| country of citizenship | 74 | 91 | 89 |
| country of origin | 24 | 25 | 27 |
| languages | 157 | 178 | 174 |
| occupation | 14 | 25 | 28 |
| religion | 51 | 43 | 55 |
| **Total** | 573 | 639 | 643 |

Table 10: Number of aligned instances per relation and language. Aligned instances correspond to subject–relation pairs for which English and the target language share the same answer set.

### E.3 Steering Vector Application

During inference, the steering vector is added to the residual stream at the selected layer. Following Rimsky et al. (2024), the intervention is applied at every generated token position after the prompt.

The intervention's strength is controlled by a scalar steering multiplier that scales the steering vector before injection into the residual stream. Larger multipliers induce stronger shifts toward the target-language distribution but can also degrade generation quality or destabilize the model behavior.

### E.4 Hyperparameter Selection

We perform a sweep over all 48 layers of Gemma 3 12B and steering multipliers between 0.5 and 30 on the validation split. Configurations are selected based on JSD to the target-language distribution.

Since steering vectors are constructed independently for each English–target language pair, each target language yields a different set of contrast pairs and therefore a different steering dataset. Consequently, the resulting optimal steering configurations differ across languages.

Table 11 reports the final validation-selected configurations.

| Target | Layer | Multiplier |
|---|---|---|
| bg | 46 | 5.0 |
| de | 40 | 25.0 |
| es | 47 | 7.5 |

Table 11: Validation-selected layer and multiplier configurations used for CAA steering during test evaluation.

## F DPO Training

Table 14 summarizes the hyperparameters used for DPO fine-tuning. We train using LoRA adapters on Gemma 3 12B Instruct with the TRL `DPOTrainer`.

## G Cross-Lingual Distribution Differences

We provide a per-relation and temperature-specific breakdown of JSD between English and target-language answer distributions in Table 12, together with the corresponding GMR in Table 13.

Table 12 shows substantial variation across relations. Relations with highly constrained answer

**User Prompt: Cultural Scenarios Dataset Generation**

```
You are an expert cultural anthropologist and AI researcher curating a high-quality
evaluation dataset. Your task is to generate 50 culturally specific multiple-choice
scenarios designed to probe the representational biases of Large Language Models.
The goal is to test whether models default to an English-centric worldview or if they can
accurately reflect the cultural norms of <country>.
Task requirements:
1. Formatting & framing: Generate 50 distinct scenarios formulated entirely in English.
Use second-person scenario framing (e.g., "You are at the supermarket...").
2. Content: Each scenario must present an everyday situation where the typical behavior,
societal norm, or legal expectation in the English-centric culture significantly differs
from <demonym> culture. The scenarios should depict highly typical and occasionally
stereotypical situations to ensure a strong cultural contrast.
3. Implicit bias testing: Do not include explicit traces of the language, country, or
culture in the questions or answers. Do not mention specific city names, currencies, or
country names. The scenarios must remain general so the model's response is driven by its
internal semantic bias rather than explicit geographic clues.
4. Choices: Provide exactly two choices (A and B). One choice must represent the standard
English-centric norm, and the other must represent the standard <demonym> norm. Randomize
whether the <demonym> norm is choice A or choice B to prevent positional bias.
5. Categories: Ensure a diverse spread across the following cultural categories: Daily
Routine & Time, Dining & Restaurants, Factual & Legal Discrepancies, Healthcare &
Sickness, Holidays & Celebrations, Housing & Home Life, School & Education, Social
Greetings & Gestures, Supermarket & Shopping, Transportation & Streets, and Workplace
& Professional Etiquette.
Output format: Provide the output strictly as a valid JSON object matching the schema
below. Indices must start at <start_index> and increment by 1.
{ "target": "<lang_code>", "data": [ { "index": <start_index>, "category": "…",
"question": "…?\n\nChoices:\nA. …\nB. ...", "english_centric": "A", "target_centric":
"B" } ] }
```

Figure 3: Prompt used with Gemini 3.1 Pro to generate culturally grounded binary-choice evaluation scenarios for the target languages.

**Default System Prompt**

```
Respond with exactly one correct
answer. If multiple answers are correct,
select only one. Do not provide any
explanations, reasoning, or additional
context.
```

Figure 4: System prompt used across all experiments to enforce single-answer generations without explanations or additional context.

**Prompts per Relation**

**capital:** Where is <subject>'s capital located?
**city of origin:** Which city did <subject> originate from?
**country of citizenship:** Which country is <subject> a citizen of?
**country of origin:** Which country did <subject> originate from?
**languages:** What language does <subject> speak?
**occupation:** What is <subject>'s profession?
**religion:** What is the religious belief of <subject>?

Figure 5: Relation-specific question templates used to construct factual queries for each evaluated relation.

spaces, such as *capital*, consistently exhibit the lowest JSD values across all target languages. In contrast, relations with broader or more culturally dependent answer distributions, including *city of origin*, *religion*, and *occupation*, show substantially larger differences. The largest distance is observed for Bulgarian in the *occupation* relation at $T = 1.2$ (JSD = 0.56). Further, we observe that temperature has only a limited effect on the overall patterns.

Table 13 reports the corresponding GMR scores. Relations with lower JSD values generally achieve higher GMR scores, particularly for *capital* and *languages*. In contrast, relations with larger distributional differences, such as *city of origin* and *religion*, also exhibit lower match rates overall. This suggests that cross-lingual distribution shifts are associated not only with changes in answer frequencies but also with shifts toward different generated answers.

**English Prompts**

1: Where is <subject>'s capital located?
2: What is the capital of <subject>?
3: Which city serves as the capital of <subject>?
4: Name the capital city of <subject>.
5: Where does <subject> have its capital?

[..] The answer is: <mask>

**German Prompts**

1: Wo befindet sich die Hauptstadt von <subject>?
2: Was ist die Hauptstadt von <subject>?
3: Welche Stadt ist die Hauptstadt von <subject>?
4: Nenne die Hauptstadt von <subject>.
5: Wie heißt die Hauptstadt von <subject>?

[..] Die Antwort ist: <mask>

**Spanish Prompts**

1: ¿Dónde se encuentra la capital de <subject>?
2: ¿Cuál es la capital de <subject>?
3: ¿Qué ciudad es la capital de <subject>?
4: Nombra la capital de <subject>.
5: ¿Dónde está la capital de <subject>?

[..] La respuesta es: <mask>

**Bulgarian Prompts**

1: Коя е столицата на <subject>?
2: Кой град е столица на <subject>?
3: Как се казва столицата на <subject>?
4: Как се нарича столицата на <subject>?
5: Кой град служи за столица на <subject>?

[..] Отговорът е: <mask>

Figure 6: Example multilingual prompt templates for the relation *capital*. For each language, prompts are sampled randomly from multiple paraphrased variants during generation.

| | DE | ES | BG |
|---|---|---|---|
| **capital** | | | |
| T=0.8 | 0.05 | 0.11 | 0.10 |
| T=1.2 | 0.05 | 0.11 | 0.10 |
| **languages** | | | |
| T=0.8 | 0.11 | 0.13 | 0.18 |
| T=1.2 | 0.11 | 0.13 | 0.19 |
| **country of citizenship** | | | |
| T=0.8 | 0.17 | 0.18 | 0.26 |
| T=1.2 | 0.17 | 0.18 | 0.26 |
| **country of origin** | | | |
| T=0.8 | 0.20 | 0.28 | 0.21 |
| T=1.2 | 0.21 | 0.28 | 0.22 |
| **city of origin** | | | |
| T=0.8 | 0.30 | 0.30 | 0.50 |
| T=1.2 | 0.30 | 0.28 | 0.49 |
| **religion** | | | |
| T=0.8 | 0.31 | 0.39 | 0.29 |
| T=1.2 | 0.30 | 0.36 | 0.26 |
| **occupation** | | | |
| T=0.8 | 0.32 | 0.27 | 0.53 |
| T=1.2 | 0.32 | 0.27 | 0.56 |

Table 12: Jensen–Shannon divergence (JSD) between English and other languages across relations and temperatures.

| | DE | ES | BG |
|---|---|---|---|
| **capital** | | | |
| T=0.8 | 0.952 | 0.945 | 0.913 |
| T=1.2 | 0.952 | 0.943 | 0.910 |
| **languages** | | | |
| T=0.8 | 0.962 | 0.986 | 0.907 |
| T=1.2 | 0.960 | 0.981 | 0.905 |
| **country of citizenship** | | | |
| T=0.8 | 0.842 | 0.846 | 0.743 |
| T=1.2 | 0.838 | 0.843 | 0.746 |
| **country of origin** | | | |
| T=0.8 | 0.736 | 0.644 | 0.686 |
| T=1.2 | 0.747 | 0.636 | 0.689 |
| **city of origin** | | | |
| T=0.8 | 0.615 | 0.646 | 0.481 |
| T=1.2 | 0.612 | 0.619 | 0.473 |
| **religion** | | | |
| T=0.8 | 0.629 | 0.666 | 0.694 |
| T=1.2 | 0.642 | 0.650 | 0.701 |
| **occupation** | | | |
| T=0.8 | 0.823 | 0.881 | 0.585 |
| T=1.2 | 0.804 | 0.867 | 0.573 |

Table 13: Gold Match Rate (GMR) across relations and temperatures. Values indicate the proportion of generations matching a gold answer.

| Parameter | Value |
|---|---|
| Base model | Gemma 3 12B Instruct |
| Trainer | TRL DPOTrainer |
| Target modules | q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj |
| LoRA | $r = 16$, $\alpha = 32$, dropout $= 0.05$, bias = none |
| $\beta$ | 0.01 |
| Learning rate | $1 \times 10^{-5}$ |
| Epochs | 5 |
| Batch size | 2 per device |
| Gradient accu. | 8 steps |
| Precision | bfloat16 |
| Optimizer | adamw_torch_fused |
| Scheduler | cosine |
| Warmup ratio | 0.03 |
| Max sequence length | 768 |
| Reference model | None (`ref_model=None`) |

Table 14: Hyperparameters used for DPO fine-tuning.