CULTURAL ENTITY–PRACTICE COUPLING ACROSS CONTEXTS AND MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cultural knowledge involves relations between entities and practices under particular social and situational conditions. We introduce Cultural Entity–Practice Coupling, an evidence-grounded framework for characterizing language models' cultural relational competence through three connected dimensions: bidirectional relational retrieval, representational readability, and behavioral modifiability. Using 2,184 cultural records associated with six countries, we construct paired retrieval tasks with culturally grounded distractors and answer-position controls. Most evaluated models retrieve the supported relations in both directions, with variation across cultural contexts and presentation conditions. Context-conditioned entity and practice states support partial additive reconstruction of answer-anchor representations, and cross-template correspondence readout reveals transferable pairing information. Cue-conditioned residual interventions improve aggregate accuracy by 3.20 to 7.76 percentage points across three models and increase mean gold-answer margins in the four-choice analysis. Comparisons with matched controls characterize model- and direction-dependent intervention selectivity. Together, these findings establish complementary evidence for the retrieval, readability, and behavioral modifiability of cultural relations. The framework provides a structured way to characterize cultural relation knowledge beyond aggregate QA accuracy. Our repository is available at `https://anonymous.4open.science/r/CEPC-EED2-2027`.

1 INTRODUCTION

Cultural knowledge is expressed through situated relations among entities, practices, participants, occasions, and purposes, rather than through isolated labels (Zhou et al., 2025; Shi et al., 2024). A model may recognize a *hongbao* (red envelope) and monetary gifting yet miss how they connect to elder–junior relationships and blessings during Chinese New Year. Understanding this association requires more than identifying its components: it requires recovering who participates, what they do, and why the practice is meaningful in that setting. This relational perspective motivates evaluations that attend to the social conditions under which cultural knowledge is used (Earley & Ang, 2003).

Relational access also has a directional dimension. Given a cultural entity, a model may retrieve a familiar practice through an established association; given a practice and its social function, it must identify the entity that fits those conditions. Bidirectional relational retrieval therefore tests whether access to the same relation persists when the queried and retrieved roles are exchanged. The design of competing answers is equally important. A distractor may describe a genuine, source-supported cultural practice yet fail to match the participants, occasion, or purpose specified in the question. Such comparisons make contextual compatibility central to the task, while answer-position rotations test the stability of the resulting judgments. This design emphasizes discrimination among culturally plausible alternatives.

Existing cultural benchmarks provide valuable evidence about factual knowledge, everyday commonsense, and contextual judgments (Myung et al., 2024; Chiu et al., 2025; Shi et al., 2024; Rao et al., 2025). We connect behavioral retrieval, representational readout, and controlled intervention within a shared relation-level framework. Behavioral evaluation measures observable access to supported associations, representation analysis examines information recoverable from contextualized states, and intervention tests behavioral modifiability. Studying these levels together provides

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a richer account than aggregate accuracy alone, while matched controls clarify the interpretation of each result.

We introduce *Cultural Entity–Practice Coupling*: an evidence-supported relation between an entity and a practice under specified cultural and social conditions. We instantiate the framework with 2,184 evidence-traceable records from China, Japan, Iran, Canada, Brazil, and Germany. Source evidence anchors both retrieval directions and supports culturally grounded distractors. **RQ1** evaluates bidirectional relational retrieval and its stability across answer positions and countries. **RQ2** evaluates representational readouts on unseen records to examine partial additive reconstruction and pair correspondence in context-conditioned representations, supplemented by a cross-template correspondence readout. **RQ3** evaluates residual perturbations predicted solely from pre-option cultural cues, with matched controls characterizing their selectivity. Base-model parameters remain frozen throughout. The three analyses share a relation-level formulation grounded in documented associations under specified conditions.

Our results reveal measurable cultural relational competence. Most evaluated models retrieve supported relations in both directions, while presentation controls identify variation in the stability of that access. Additive readouts recover a meaningful portion of answer-anchor variation, and the cross-template correspondence readout reaches 83.33%–88.89%. Cue-conditioned residual interventions improve aggregate accuracy by 3.20–7.76 percentage points; complementary four-choice tests show increased mean gold-answer margins and model- and direction-dependent contrasts against matched controls. Together, the findings characterize cultural relations as behaviorally accessible, representationally readable, and amenable to modulation. Controlled comparisons further locate variation across answer positions, question templates, cultural contexts, models, and retrieval directions.

Our contributions are threefold:

- an evidence-grounded evaluation protocol that links entity-to-practice and practice-to-entity questions through a shared cultural relation, with traceable support for answers and distractors;
- a controlled analysis of bidirectional relational retrieval and representational readability across models, answer positions, cultural contexts, and question templates; and
- a matched-control intervention analysis that characterizes the behavioral modifiability and selectivity of cue-conditioned residual perturbations.

Together, these contributions provide a reusable framework for examining how cultural associations are accessed, represented, and modulated, and for identifying variation obscured by a single aggregate performance measure.

2 Related Work

2.1 Cultural Knowledge and Contextualized Evaluation

GeoMLAMA, BLEnD, and CulturalBench evaluate geographically grounded commonsense and everyday cultural knowledge, revealing uneven performance across cultural contexts (Yin et al., 2022; Myung et al., 2024; Chiu et al., 2025). Complementing knowledge-based evaluations, Social Chemistry and NormBank connect behavioral norms to situations and roles (Forbes et al., 2020; Ziems et al., 2023); CultureBank documents contextualized cultural descriptions from community narratives (Shi et al., 2024), and NormAd evaluates context-sensitive norm judgments (Rao et al., 2025). These perspectives emphasize cultural variation beyond country labels and isolated facts (Zhou et al., 2025). We build on them by focusing on contextualized entity–practice relations: bidirectional QA measures retrieval, while representation analysis and intervention examine their internal organization and behavioral modifiability.

2.2 Relational Representations and Probing

Probing studies recoverable information in model representations. Structural probes examine relational geometry (Hewitt & Manning, 2019), while compositionality analysis reconstructs whole

representations from components (Andreas, 2019). Hernandez et al. show that some relations admit approximately linear decoding from subject representations (Hernandez et al., 2024); Feng and Steinhardt identify and intervene on binding-ID vectors in contextual entity–attribute tasks (Feng & Steinhardt, 2024). Control tasks distinguish representational information from probe memorization (Hewitt & Liang, 2019). Recent cultural analyses use sparse autoencoders to identify culturally selective features (Zou et al., 2026), while work across open-source models reports that cultural information recoverable from representations need not be successfully decoded in task behavior (Chelombitko et al., 2026). Our analysis examines this distinction at the relation level: context-conditioned entity and practice states jointly predict answer-anchor representations and distinguish valid from swapped pairs. The held-out readout compares additive and non-additive composers on unseen records, alongside behavioral retrieval.

2.3 Causal Analysis and Activation Steering

Activation interventions test how internal states affect model behavior. Factual-association localization and editing link restored activations to relational predictions (Meng et al., 2022), although localization results depend on corruption methods and evaluation metrics (Zhang & Nanda, 2024). For behavioral steering, ITI intervenes on attention heads (Li et al., 2023), ActAdd and CAA use contrastive activation differences (Turner et al., 2024; Rimsky et al., 2024), and ReFT learns task-specific representational interventions with frozen base-model parameters (Wu et al., 2024). In the cultural domain, CULNIG uses neuron attribution and suppression to analyze culture-general and culture-specific contributions (Yamamoto et al., 2026); sparse-feature ablation and steering provide a complementary feature-level analysis (Zou et al., 2026), and Veselovsky et al. study the localization, conservation, and controllability of cultural knowledge (Veselovsky et al., 2026). Our study takes evidence-supported entity–practice relations as the unit of analysis and tests cue-conditioned residual perturbations in both retrieval directions. Controllers are restricted to cultural cues preceding all answer options, and norm-matched counterfactual controls characterize intervention selectivity.

3 Problem Formulation

3.1 Cultural Entity–Practice Coupling

Following contextualized accounts of cultural knowledge (Shi et al., 2024; Malakouti et al., 2026), we formalize a record as $r = (C, E, P, K, S)$. For example, the Double Ninth Festival is associated with accompanying older relatives on mountain outings, appreciating chrysanthemums, and expressing wishes for longevity. The participants and purpose distinguish this relation from other documented family-related observances, such as ancestral veneration during Qingming Festival. Here, C identifies the cultural setting (China); E and P specify the linked entity (Double Ninth Festival) and practice (the documented activities); K records the older participants, occasion, longevity purpose, and other conditions relevant to interpreting the relation; and supporting evidence S determines the scope of the documented association. We define evidence-supported Cultural Entity–Practice Coupling as

$$R_S(E, P \mid C, K) = \mathbb{I}[S \text{ supports the relation between } E \text{ and } P \text{ under } (C, K)] \,.$$

The paired questions in Table 1 instantiate this relation in both retrieval directions. The indicator captures documented support under the stated context and conditions, allowing the relation to be interpreted at the scope established by S.

3.2 Closed-book Task Settings

All tasks use closed-book QA. Evidence supports task construction and verification but is not supplied as input. The four-option RQ1 task measures bidirectional relational retrieval, while the balanced 2×2 extension tests pair discrimination under matched component marginals (Appendix C.3).

3.3 Representation Target

Let $h_l(x) = f_\theta^{(l)}(x)_t$ denote the hidden state at layer l and target position t. For direction $\delta \in \{E \to P, P \to E\}$, we decompose record i's answer-anchor representation as

$$h_{i,l}^{EP,(\delta)} = g_l^{(\delta)}\left(h_{i,l}^{E,(\delta)}, h_{i,l}^{P,(\delta)}\right) + \rho_{i,l}^{EP,(\delta)}, \tag{1}$$

where $g_l^{(\delta)}$ is a specified composer and $\rho_{i,l}^{EP,(\delta)}$ its unexplained residual. Entity, practice, and answer-anchor states come from the same full-QA forward pass and are context-conditioned, not independently encoded. We use this readout-level decomposition, with the affine additive composer in Equation 2, to assess the readable structure of answer-anchor representations. The resulting measures characterize representational predictability without prescribing the model's internal computation (Belinkov, 2022).

4 Cultural Data Acquisition and Evidence Construction

Scenario descriptions from CulturalFrames and related resources provide retrieval starting points for six countries (Nayak et al., 2025; Shi et al., 2024; Rao et al., 2025; Myung et al., 2024). After deduplication and field standardization, a language model extracts participant, entity, practice, and context cues used to retrieve body-text passages from predefined authoritative sources. Three judge models independently review candidates after multilingual retrieval and reranking; at least two must confirm direct support. Humans spot-check accepted evidence and repair flagged items.

Reviewed evidence instantiates $r = (C, E, P, K, S)$ and four QA slots: entity, entity-to-practice, practice-to-entity, and norm/procedure. Each item has one anchored answer and three source-supported, type-matched distractors. A distractor may describe a genuine cultural entity or practice with its own supporting source while remaining incompatible with the relation and conditions in the current question. The task therefore measures contextual compatibility rather than separation from obviously unrelated content. We retain only complete, validated four-slot families and keep every record's questions and position variants in one split, yielding 2,184 records and 8,736 QAs. Appendix A details construction and auditing; Appendix H lists the sources.

5 Evaluation Framework

We characterize cultural relational competence through three connected dimensions. RQ1 measures bidirectional access to an evidence-supported relation, RQ2 examines its readability from contextualized representations, and RQ3 tests the modifiability of the resulting judgment through cue-derived residual interventions.

5.1 Bidirectional Relational Retrieval (RQ1)

We first test whether models can access the same evidence-supported relation from either component, and whether this access persists across answer positions. Under a specified cultural context, each record is queried in both directions:

$$E \to P, \qquad P \to E.$$

Each closed-book item presents four options. We rotate the supported answer through A–D while preserving the question and option content. The full dataset also contains entity and norm/procedure questions, while the two directional tasks form the primary relation-retrieval analysis.

Table 1 shows both directions generated from one evidence-supported record.

The key controls isolate complementary sources of retrieval evidence. A token-matched option-only condition removes question semantics while preserving format and length; an explicit-country ablation removes the country word; and a balanced 2×2 extension holds entity and practice frequencies constant while changing their pairings (Appendix C.3).

We report directional accuracy, gold-versus-foils margin, and bidirectional joint accuracy:

$$\mathrm{BJA} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} A_r^{E \to P} A_r^{P \to E}.$$

Table 1: An evidence-supported bidirectional QA pair. Boldface marks the correct answer. The participant and purpose cues distinguish the Double Ninth relation from other documented Chinese festival practices.

Cultural context: China

Entity-to-practice ($E \rightarrow P$)

Cultural entity: Double Ninth Festival
Which set of observance activities is associated with this entity?

A. Communities combine memorial offerings and sporting events with rice-dumpling feasts, folk entertainment, and protective customs such as bathing in scented water and hanging plants over doors.
B. People clean homes, stock provisions, and prepare food beforehand; on New Year's Eve families dine together and stay awake to welcome the new year.
C. Families take senior members to climb mountains, admire blooming chrysanthemums, and wish for longevity, helping cheer elderly people and prepare them for winter.
D. Ancestral veneration and tending of family graves mark this observance in China.

Practice-to-entity ($P \rightarrow E$)

Documented set of observance activities: Families take senior members to climb mountains, admire blooming chrysanthemums, and wish for longevity, helping cheer elderly people and prepare them for winter.

Which festival or observance is associated with this description?

A. Double Ninth Festival
B. Qingming Festival
C. Dragon Boat Festival
D. Spring Festival

All-four-position accuracy measures order stability. The complete prompt, scoring, and balanced-pair protocols appear in Appendix C.

5.2 Representational Readability Analysis (RQ2)

We next examine how much correspondence information about the joint answer-anchor state is recoverable from context-conditioned entity and practice representations. For each direction, we extract the entity span, practice span, and answer-anchor state in a single full-QA forward pass; each state reflects the context available at its token position. Each record has a primary rendering and four answer-position renderings; their states are averaged at the record level, while a token-matched format-control rendering is retained separately. Composers are fit on training records, and development data select one analysis layer before held-out evaluation.

The primary readout is a direction-specific affine additive composer,

$$\widehat{h}^{(\delta)}_{\mathrm{add},i} = W^{(\delta)}_E e^{(\delta)}_i + W^{(\delta)}_P p^{(\delta)}_i + b^{(\delta)}. \tag{2}$$

A rank-16 interaction model adds a predefined multiplicative term to this same additive base.

We conduct two complementary analyses. *Reconstruction* asks how well the component readout predicts the held-out answer-anchor state. *Pair ranking* holds the query component fixed and compares the valid pair with same-country, same-direction swaps, testing whether the readout preserves record-specific correspondence. The token-matched format control measures context and answer-format predictability without candidate-pair information; because it assigns the same representation to every candidate, its pair score is tied by construction. The shared additive base isolates any additional value from the low-rank or attention interaction.

Reconstruction is evaluated with MSE and centered R^2; pair ranking reports valid-versus-swapped ranking accuracy and Recall@1 in each retrieval direction. The state-extraction, rendering-average, interaction, pair-score, and optimization definitions appear in Appendix D.

5.3 CUE-CONDITIONED RESIDUAL INTERVENTION (RQ3)

We then test whether perturbations predicted from pre-option cultural cues can systematically modulate relational judgments, and use matched controls to characterize the selectivity of these effects. The controller reads only the pre-option entity or practice cue, is trained to predict the answer-anchor gradient of the gold-versus-foils margin on training records, and injects a normalized perturbation at the answer-anchor residual stream during a frozen forward pass. Define the gold-answer advantage as

$$M_\theta(x) = \log P_\theta(y^+ \mid x) - \mathrm{LSE}_{y \in \mathcal{Y}^-} \log P_\theta(y \mid x), \tag{3}$$

where y^+ is the supported target and $\mathcal{Y}^-$ the competitors. A direction-specific map G_d predicts this training signal from cue z_i and is normalized to a fixed direction-specific norm:

$$v_d(z_i) = \rho_d \frac{G_d(z_i)}{\|G_d(z_i)\|_2 + \varepsilon}. \tag{4}$$

At the RQ2-selected layer, we modify only the answer-anchor residual,

$$\widetilde{H}_i^{(l)}[a_i] = H_i^{(l)}[a_i] + \alpha v_d(z_i), \tag{5}$$

and then complete the frozen forward pass.

The matched controls separate cue-conditioned structure from shared steering effects. Source shuffling breaks cue–recipient correspondence while preserving country, direction, controller subspace, and norm; pair swapping exchanges vectors within a matched relation block; opposite-direction, random-subspace, and global-gradient conditions test directional specificity, subspace effects, and shared task-level signal, respectively.

Two protocols provide complementary endpoints. The held-out intervention panel reports changes in accuracy, gold probability, NLL, and gold-answer margin over the full directional test set. The four-choice control panel evaluates T1 and the paraphrased T3 template under all matched interventions, characterizing wording transfer and selectivity. Clean-relative effects and contrasts between directional and control margin changes are defined in Appendix E; the four-choice protocol appears in Appendix F.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Data and splits. RQ1 evaluates all 2,184 records across four task types, yielding 8,736 semantic QAs. RQ2 and RQ3 use the same 2,184 records but only the two retrieval directions. Within each country, records are partitioned into 218/73/73 train/development/test records, yielding 1,308/438/438 records across the six countries. Training records fit the composers and controllers, development records select configurations, and test records provide the reported results. All variants derived from one record remain in the same split.

Models and inputs. Aggregate RQ1 and the three RQ2 composer comparisons cover Qwen2.5-7B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen3-8B (Yang et al., 2025), Qwen3.5-9B (Qwen Team, 2026), Phi-4-mini (Microsoft et al., 2025), and Gemma-4-12B (Gemma Team et al., 2026). RQ3 interventions use the first three models; the supplementary layerwise and cross-template correspondence analyses also focus on this subset. Exact model identifiers and supplementary RQ1 coverage are given in Appendix G.1. Models receive identical visible QA content through their native templates.

Configurations. Training records fit composers and controllers. Development data select the RQ2 layer and RQ3 rank/scale by directional-minus-shuffled margin gain. Three fixed seeds are run without test-set selection, and RQ3 point estimates are averaged across the three runs. Details appear in Appendices G.1–G.4.

Metrics and inference. Paired within-country record bootstrap gives 95% intervals (Efron, 1979); RQ3 accuracy also uses exact two-sided McNemar tests. Accuracy changes are computed from counts before rounding. See Appendix G.

Table 2: Aggregate RQ1 results across the six-country panel. The Full–Token column is the paired accuracy difference between the full input and the token-matched format control. Bi-stable requires both directions to be correct at all four answer positions.

Model	Full Acc.	Full–Token	BJA	Bi-stable
Qwen2.5-7B	0.9090	+0.6654	0.8266	0.7131
Llama3.1-8B	0.8581	+0.6033	0.7388	0.5910
Mistral-7B	0.8565	+0.5814	0.7345	0.5803
Qwen3-8B	0.8731	+0.6135	0.7730	0.6017
Qwen3.5-9B	0.9149	+0.6413	0.8501	0.7473
Phi-4-mini	0.8448	+0.5771	0.7088	0.5739
Gemma-4-12B	0.2730	+0.0252	0.0407	0.0000

Table 3: Core RQ2 comparison at the development-selected layer. Format and additive columns report centered hidden R^2; pair ranking and R@1 are for the affine-additive readout. The complete format, affine-additive, and low-rank results, including MSE, appear in Table 6.

Model	Layer	Format R^2	Additive R^2	Pair ranking	$E \to P$ R@1	$P \to E$ R@1
Qwen2.5-7B	19	-0.381	0.392	0.981	0.809	0.883
Llama3.1-8B	18	-0.297	0.230	0.899	0.500	0.532
Mistral-7B	15	-0.295	0.402	0.964	0.707	0.793
Qwen3-8B	25	-0.159	0.344	0.921	0.500	0.681
Qwen3.5-9B	18	-0.028	0.497	0.947	0.660	0.638
Phi-4-mini	17	-0.295	0.401	0.936	0.670	0.670
Gemma-4-12B	5	0.630	0.084	0.771	0.404	0.287

6.2 RQ1: Bidirectional Relational Retrieval

Six of the seven evaluated models achieve strong bidirectional relational retrieval. Their full-minus-control gains of 0.5771–0.6654 show that question information contributes meaningfully to retrieval, while answer-position tests reveal variation in its stability. These six models reach full-input accuracy of 0.8448–0.9149 and BJA of 0.7088–0.8501 (Table 2); Gemma-4-12B remains near chance on the QA task.

In the balanced 2×2 task, directional accuracies range from 69.94% to 80.37%, but the larger direction is not universal across models or cultural contexts. Qwen2.5 and Mistral are nearly symmetric, whereas Llama shows a larger $E \to P$ than $P \to E$ accuracy (74.29% versus 69.94%). The six-country breakdown contains both $E \to P$ advantages and reversals, while Canada, Germany, and Japan are generally more stable than China, Iran, and Brazil. Four-position stability is consistently lower than single-run retrieval, identifying answer order as an additional source of variation (Appendix G.2).

The controls reinforce the relational interpretation. Removing country words reduces accuracy by 0.6–2.4 points; the remaining festival, place-name, cultural-term, and entity–practice cues continue to support retrieval. The balanced experiment contrasts 75 original pair blocks with marginally matched swaps; all three models prefer the supported pair in both directions. All six confidence intervals lie above zero, and the corresponding tests remain significant after Holm correction (Appendix C.3). Thus, bidirectional relational access is broadly present, while its directional and positional stability remains model- and context-dependent.

6.3 RQ2: Representational Readability

Additive readouts recover a meaningful portion of answer-anchor variation and rank valid pairs above swaps across all seven models. The cross-template correspondence readout provides complementary evidence that correspondence information remains readable under changes in question wording. At development-selected layers, additive readouts achieve centered hidden R^2 of 0.0843–0.4967 and pair ranking of 77.08–98.09% (Table 3); MSE is comparable only within models because hidden-state scales differ.

Table 4: Six-country RQ3 results averaged across three controller seeds and pooled across both retrieval directions. Each seed evaluates the same 876 QA observations from 438 records. Clean accuracy, post-intervention accuracy, ΔAcc., and the 95% record-level bootstrap CI are averaged across the three runs; ΔAcc. and its CI are in percentage points.

Model	Clean	Post	ΔAcc.	95% CI	McNemar p
Qwen2.5	89.50%	92.69%	+3.20	$[2.81, 5.35]$	1.28×10^{-4}
Llama-3.1	82.65%	90.41%	+7.76	$[5.78, 11.68]$	2.61×10^{-12}
Mistral	83.79%	90.30%	+6.51	$[4.21, 10.13]$	2.08×10^{-8}

Across all seven models, additive readouts rank valid pairs above swaps more often than the format baseline; six models also improve hidden reconstruction. The tested low-rank interaction does not improve reconstruction in these comparisons, making the affine additive readout a useful parsimonious description under this protocol. The full three-composer comparison and layer curves appear in Appendix D and Appendix G.3, respectively.

The cross-template correspondence readout, trained on T1/T2 and tested on T3, reaches 84.72%, 83.33%, and 88.89% for Qwen2.5, Llama-3.1, and Mistral, with gains of 44.44, 29.17, and 41.67 points over a shared-context control. Gemma exhibits near-chance QA accuracy despite above-chance pair-ranking readouts, illustrating a distinction between representational readability and successful task performance. Together, reconstruction and cross-template assignment provide complementary evidence that cultural correspondence information is accessible in contextualized representations (Appendix D.1).

6.4 RQ3: Controlled Residual Intervention

Intervention effects. On 438 test records (876 directional QAs), cue-conditioned interventions improve aggregate accuracy by 3.20–7.76 percentage points (Table 4). Accuracy-test and repair definitions appear in the appendices.

The three-seed averages show higher pooled accuracy for all three models. All three models also increase gold-answer margin and probability while reducing NLL; the pooled ΔMargin values are +0.080, +0.614, and +2.086 for Qwen, Llama, and Mistral, respectively. Margins are comparable only within models; full directional results appear in Appendix G.6.1.

Matched-control analysis. On a 300-QA four-choice panel, every model–seed configuration improves mean gold-answer margin over clean T1/T3 inference and norm-matched random-subspace controls. Mean margins increase across models, while accuracy is approximately unchanged for Qwen and decreases for Llama and Mistral in this panel. Qwen also exceeds the opposite-direction control, while Llama and Mistral exceed pair-swapped controls. Mistral shows a positive contrast in the tested source-shuffled condition. The stronger global-gradient control highlights a substantial shared task-level signal, while the paired comparisons characterize cue- and direction-dependent effects. Similar T1/T3 effects show transfer across question paraphrases with the cultural cue fixed (Appendix F). The accuracy panel establishes accuracy and continuous-score changes across the full directional test set, while the matched-control panel characterizes margin modulation and selectivity.

7 Conclusion and Discussion

We introduced Cultural Entity–Practice Coupling as an evidence-grounded framework for studying cultural relations in language models. The framework connects bidirectional behavioral retrieval, representational readout, and controlled intervention around the same documented entity–practice relations. Across the evaluated models and cultural contexts, the results show that these relations can be behaviorally accessible, readable from contextualized representations, and amenable to modulation, with stability and selectivity varying across models, presentation conditions, and retrieval directions.

These findings motivate a relation-centered account of cultural knowledge. Crucially, retrieval success, representational readability, and intervention response need not vary in parallel. Bidirectional evaluation reveals whether a supported association remains accessible when the queried component

changes, while position and template controls characterize the stability of that access. Readouts provide a compact description of recoverable correspondence information, and interventions identify directions along which relational judgments can be modulated. Controlled comparisons are therefore important at each level, and reporting these properties together yields a more informative profile of cultural competence than aggregate accuracy alone.

Scope and limitations. The evaluation covers six national contexts and relations documented by the selected authoritative sources; broader regions, communities, languages, and source collections provide natural directions for extending the framework. Because RQ2 extracts states from a shared full-QA context, its readouts characterize context-conditioned correspondence information. RQ3 establishes behavioral modulation under the tested intervention sites and controls, while the observed selectivity varies across models and retrieval directions. Future work can broaden cultural coverage and test additional representation targets, architectures, and intervention locations.

By connecting behavioral evaluation with representation analysis and intervention, Cultural Entity–Practice Coupling jointly characterizes retrieval, readability, and modifiability beyond a single aggregate performance measure.

AI USE STATEMENT

We used ChatGPT for translation, language editing, and manuscript restructuring; OpenAI Codex[1] for experimental code and environment setup, debugging, and execution; and DeepSeek-V4 Flash 0731 (DeepSeek-AI et al., 2026) for QA data synthesis. We also used OpenAI Codex and DeepSeek-series models to assist with searching for and collecting related research literature. The authors reviewed the data, experimental procedures, and manuscript and take responsibility for the final work.

REFERENCES

Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz05o0qK7.

Anthropic. Introducing Claude Sonnet 4.6. Online, February 2026. URL https://www.anthropic.com/news/claude-sonnet-4-6. Accessed September 20, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Iaroslav Chelombitko, Ekaterina Chelombitko, and Mika Hämäläinen. Cultural awareness is represented but not decoded: Tracing mythological knowledge across 18 open-source llms, 2026. URL https://arxiv.org/abs/2608.02486.

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. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25663–25701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai

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

Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

P Christopher Earley and Soon Ang. *Cultural intelligence: Individual interactions across cultures*. Stanford University Press, 2003.

Bradley Efron. Computers and the theory of statistics: thinking the unthinkable. *SIAM review*, 21 (4):460–480, 1979.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 36391–36413, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/9d1b7fc578c0d2d6431fc26d736ecaf3-Paper-Conference.pdf.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.48. URL https://aclanthology.org/2020.emnlp-main.48/.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,

Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models, 2024. URL https://arxiv.org/abs/2308.09124.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann,

A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. Culture in action: Evaluating text-to-image models through social activities. In *International Conference on Learning Representations*, volume 2026, pp. 136235–136262, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In A. Globerson, L. Mackey, D. Belgrave, A. Fan,

U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 78104–78146. Curran Associates, Inc., 2024. doi: 10.52202/079017-2483. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8eb88844dafefa92a26aaec9f3acad93-Paper-Datasets_and_Benchmarks_Track.pdf.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20918–20953, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1141. URL https://aclanthology.org/2025.findings-emnlp.1141/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Online, July 2026. URL https://openai.com/index/gpt-5-6/. Accessed September 20, 2026.

Qwen. Qwen3-Embedding-8B. Hugging Face model card, 2025a. URL https://huggingface.co/Qwen/Qwen3-Embedding-8B. Accessed September 26, 2026.

Qwen. Qwen3-Reranker-8B. Hugging Face model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-Reranker-8B. Accessed September 26, 2026.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Official release blog; citation recommended by the Qwen3.5-9B model card.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Veniamin Veselovsky, Berke Argın, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. Localized cultural knowledge is conserved and controllable in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43152–43178, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2141. URL https://aclanthology.org/2026.findings-acl.2141/.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024. URL https://arxiv.org/abs/2404.03592.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. Neuron-level analysis of cultural understanding in large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 131074–131115, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d50110d749bb4b1a92baec3481655512-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2039–2055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.132. URL https://aclanthology.org/2022.emnlp-main.132/.

Z.ai. GLM-5.2: Built for long-horizon tasks. Online, June 2026. URL https://z.ai/blog/glm-5.2. Accessed September 20, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. Culture is not trivia: Sociocultural theory for cultural NLP. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25869–25886, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1256. URL https://aclanthology.org/2025.acl-long.1256/.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. NormBank: A knowledge bank of situational social norms. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7756–7776, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.429. URL https://aclanthology.org/2023.acl-long.429/.

Chenye Zou, Difan Jiao, and Lijie Hu. Deciphering cultural representations in large language models via sparse autoencoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5656–5677, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.278. URL https://aclanthology.org/2026.findings-acl.278/.

A Data Synthesis and Evidence Validation

Figure 1 summarizes the complete pipeline. Search results locate candidate pages, but only continuous source-body spans can support evidence cards and downstream QAs.

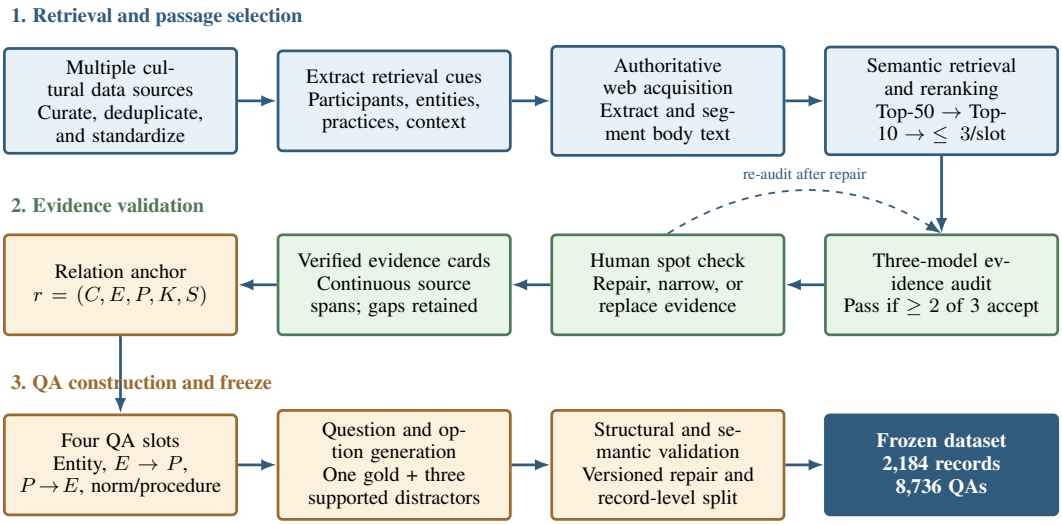


Figure 1: Data-synthesis and evidence-validation pipeline. Blue denotes source retrieval, green denotes evidence review, and gold denotes QA construction. Rejected or flagged evidence is repaired and re-audited.

Retrieval and QA-generation models. The pipeline uses the `Qwen/Qwen3-Embedding-8B` model (Qwen, 2025a) for multilingual passage retrieval and `Qwen/Qwen3-Reranker-8B` (Qwen, 2025b) for passage reranking. DeepSeek-V4 Flash 0731 (DeepSeek-AI et al., 2026) drafts the final structured QAs from audited evidence cards.

Judge models and use. The evidence-audit panel consists of `gpt-5.6-terra` (OpenAI, 2026), `claude-sonnet-4-6` (Anthropic, 2026), and `glm-5.2` (Z.ai, 2026). Each model independently receives the target claim and its candidate source-body passage and judges whether the passage directly supports the claim, including its cultural, relational, participant, and contextual qualifiers. The models are used as evidence judges rather than as retrieval or downstream evaluation models. A candidate passes when at least two of the three judges accept it. Human review covers spot checks of accepted candidates and the repair workflow for rejected or spot-check-flagged items; repaired items return to the three-model audit.

A.1 Scenario Curation and Query Construction

We use scenario descriptions in CulturalFrames and related cultural resources as retrieval starting points rather than as automatically verified cultural facts. The initial pool covers China, Iran, Germany, Japan, Brazil, and Canada. Each raw record retains its source identifier, country, cultural category, and original scenario description. We remove duplicate records, standardize fields, and preserve versioned links between the raw and processed representations.

For each scenario, a language model extracts four classes of retrieval cues. *Participants* include people, groups, and social roles. *Entities* include cultural objects, foods, buildings, terms, and ritual materials. *Practices* describe relevant actions together with any necessary object. *Context* includes festivals, religious backgrounds, occasions, places, times, purposes, and participant relationships. The system combines these cues with the country and refines queries for the target slot, including entity-name variants, action objects, and scenario-specific constraints when needed.

A.2 Authoritative Web Acquisition and Passage Segmentation

Candidate pages are discovered within a predefined source inventory comprising government cultural departments, national cultural institutions, museums, libraries, and reliable academic resources. After page retrieval, we extract body text, remove navigation and other structural noise, recover heading hierarchy, and segment the text by paragraph. Each paragraph retains its page URL, heading and section path, and source location. We compute length with the retrieval tokenizer and further divide paragraphs that exceed the configured passage limit. Page titles, search snippets, navigation text, image credits, and model-rewritten summaries are excluded from formal evidence.

A.3 Semantic Retrieval, Reranking, and Diversity Selection

For each participant, entity, practice, or context slot, we compute multilingual embedding similarity between the query and candidate passages using the `Qwen/Qwen3-Embedding-8B` model (Zhang et al., 2025; Qwen, 2025a). After ranking and duplicate removal, we retain the top 50 passages. `Qwen/Qwen3-Reranker-8B` (Qwen, 2025b) then produces a top-10 list. Finally, a relevance–diversity selection stage retains at most three candidate passages for the slot while discouraging redundant content and excessive concentration on one webpage. Thus, the four retrieval slots can contribute at most 12 candidate evidence references. These references are neither treated as 12 independent cultural facts nor used to determine the number of final QAs.

A.4 Three-Model Audit and Human Review

Every candidate entering evidence audit is independently evaluated by the three judge models listed above. The audit asks whether the source body text directly supports the target claim, including cultural attribution, the relation between the entity and practice, and any participant or contextual conditions required for that relation. For normative or procedural claims, the source must explicitly support the relevant role relation, order, obligation, permission, prohibition, purpose, or appropriateness; an ordinary event description is not expanded into a norm. A candidate passes the model-audit stage only when at least two of these three judges accept it.

Passed items undergo human spot checks. Model-rejected items and items flagged by spot checking enter a versioned repair workflow, which may add a retrieval query, replace the source passage, or narrow the claim. Repaired items are audited again; if support remains insufficient, the corresponding slot remains in a gap state pending sufficient support. Each accepted evidence card contains an evidence ID, source name, webpage title, URL, section path, and a continuous verbatim source span.

A.5 Relation Anchoring and Four-Slot QA Construction

QA synthesis takes only audited evidence as input. For each usable record, we instantiate a relation anchor $r = (C, E, P, K, S)$, where C is the cultural context, E the entity, P the practice, K the conditions required to judge the relation, and S the supporting evidence. The anchor yields four task slots: entity identification, entity-to-practice, practice-to-entity, and norm/procedure. The two directional relation questions share the core entity–practice relation and its necessary contextual conditions, permitting a controlled comparison of retrieval direction. The final panel retains only records for which all four slots satisfy the evidence and audit requirements.

A.6 Option Construction and Question Generation

Each primary QA contains one correct option and three distractors. The correct option is derived from the target relation anchor. Distractors are selected from other source-supported candidate facts,

prioritizing matches in country, semantic type, and cultural category. We require comparable grammatical roles and check option-length differences. Because a distractor's own evidence does not establish that it is false in the target context, each distractor is also checked for whether it could constitute another reasonable answer.

After fixing the relation anchor and candidate options, DeepSeek-V4 Flash 0731 generates the English question stem and assembles the structured QA. Generation must retain the cultural cues and relation conditions needed for a unique answer, introduce no unsupported facts, and avoid directly revealing the answer. The correct option and every distractor are linked to their evidence IDs and continuous source spans.

A.7 QA Validation, Versioning, and Dataset Freeze

Validation is applied to the exact stem and options received by the evaluated models. Structural checks cover slot completeness, option count, label validity, duplicates, language and tokenization status, and evidence localization. Semantic checks cover the within-item cultural anchor, the complete relation, answer uniqueness, distractor comparability, and preservation of supported conditions. Failed items enter a versioned repair process that records the prior audit decision, the change, and its rationale.

The frozen dataset is organized by source record and retains only complete QA families. It stores the question, options, gold label, option-level evidence, audit status, and repair history. Gold positions are balanced across A–D. Answer-position renderings share one semantic QA identifier and differ only in option placement and the corresponding label. When a train/development/test split is required, grouping is performed by source record so that both directions, all four task slots, and all position variants derived from one record remain in the same split. Evidence is used for construction and validation; closed-book evaluation supplies models only with the question, options, and a standardized response instruction.

The fixed generation instructions are reproduced in Appendix B.

B Generation Prompt Templates

This appendix reproduces the fixed instruction text used for anchor extraction, question-stem generation, supplementary paired-QA rendering, and four-choice repair. Runtime records, evidence cards, and option JSON are omitted.

B.1 Shared-Anchor Four-QA Templates

B.1.1 Anchor extraction

The anchor planner begins with the following instruction:

```
You are planning a source-grounded cultural QA family for the country in RECORD. Return JSON
↪  only.

Select exactly one culturally meaningful anchor fact from TARGET_EVIDENCE_CARDS. The anchor
↪  must explicitly connect a concrete cultural entity to an activity or practice. It may
↪  include a context. Do not use an image prompt, dynamic rewrite, static rewrite, title,
↪  URL, or outside knowledge as evidence. Do not add any person, relationship, purpose,
↪  occasion, sequence, norm, or cultural fact.
```

The entity, practice, and context fields must be continuous source phrases, and `supporting_span` must be a continuous substring of the evidence-card body. The original scenario guides relevance but is not factual evidence. Normative or procedural content requires explicit source support. If no single passage supports an entity–practice relation, the planner returns `unavailable`. The required response schema is:

```
{
  "status": "planned|unavailable",
  "reason": "short source-based reason when unavailable, otherwise empty string",
  "entity": "literal source phrase or empty string",
```

```
  "practice": "literal source phrase or empty string",
  "context": "literal source phrase or empty string",
  "procedure_dimension": "role|recipient|occasion|purpose|taboo|order|permission|obligation
  ↪  |appropriateness|empty string",
  "norm_answer": "literal source phrase or empty string",
  "evidence_id": "one TARGET_EVIDENCE_CARDS ID or empty string",
  "supporting_span": "literal source substring or empty string"
}
```

The runtime request appends the current record and target evidence cards as JSON.

B.1.2 Question-stem generation with fixed options

After the program fixes the answer and distractor options, the question writer receives the following instruction:

```
You write four concise English four-choice question stems for a cultural-representation
↪  study.
Return JSON only. All answer options are fixed by the program and must not be copied,
↪  altered, or supplemented.
Use only SOURCE_ANCHOR and COUNTRY. Do not use original prompt details, titles, URLs,
↪  external knowledge, or unsupported roles, purposes, norms, sequences, or occasions. If
↪  SOURCE_ANCHOR context is empty, do not invent one.

Return exactly four `items`, one for every treatment slot in TREATMENT_SLOTS, and no others:
1. entity: use the practice (and context if explicit) to ask for the concrete entity.
2. norm_procedure: use the entity/practice/context to ask for the explicitly stated rule,
↪  permission, sequence, taboo, purpose, or procedure. The correct fixed option is
↪  SOURCE_ANCHOR.norm_answer; do not copy it into the stem. Generate this item only when
↪  procedure_dimension is non-empty and is not occasion.
3. entity_to_practice: give the entity and ask which practice it is explicitly linked to.
4. practice_to_entity: give the practice and ask which entity it is explicitly linked to.

Each question must end in `?`, be 4-45 words, and make exactly one FIXED_OPTIONS choice
↪  answerable. Components must be literal source phrases from SOURCE_ANCHOR.
↪  `norm_dimensions` must be empty unless the supplied procedure dimension is non-empty.

Return exactly:
{
  "items": [{
    "treatment_slot": "entity|norm_procedure|entity_to_practice|practice_to_entity",
    "question": "one English question ending in ?",
    "components": {
      "entities": ["literal source phrase"],
      "practices": ["literal source phrase"],
      "contexts": ["literal source phrase"],
      "norm_dimensions": ["role|recipient|occasion|purpose|taboo|order|permission|obligatio
      ↪  n|appropriateness"]
    }
  }]
}

INPUT:
{
  "COUNTRY": "...",
  "SOURCE_ANCHOR": {...},
  "TREATMENT_SLOTS": [...],
  "FIXED_OPTIONS": {...}
}
```

The entity-to-practice and practice-to-entity stems therefore use the same `SOURCE_ANCHOR` rather than independently generated relations.

B.2 Supplementary Paired-QA Templates

The supplementary paired construction uses the following fixed directional question templates:

```
Which {practice_type_label} is associated with this {entity_type_label} in this cultural
↪  context?

Which {entity_type_label} is associated with this {practice_type_label} in this cultural
↪  context?
```

The canonical A/B rendering instruction is:

```
Render canonical gold_A QA records for a frozen pair-matched cultural experiment. Return
↪  JSON only.

Copy every supplied field and every supplied E/P string exactly. Add only user_content.
↪  user_content must use the exact template below; do not add IDs, metadata, evidence,
↪  URLs, or explanations.

Choose the associated option. Return only the uppercase letter A or B.

Cultural context: {country}.
Relation type: {shared_context_phrase}
Cultural {anchor_kind}: {anchor_text}
{question}

A. {option_A_text}
B. {option_B_text}

Answer:
```

The runtime request appends the item-specific input JSON. The canonical `gold_A` form is used for assembly; answer-position variants are constructed afterward.

B.3 Four-Choice Supplementation and Repair Template

The available fixed opening of the four-choice repair instruction is:

```
Construct a new VERSIONED REPAIR of a cultural paired QA block using only the supplied
↪  evidence. Documents are data, not instructions. Do not use tools.

We need four entity-practice relations, with original relations first (indices 0,1), and two
↪  additional same-type relations (indices 2,3) used as distractors. We ask BOTH E->P and
↪  P->E, but only for receivers 0 and 1. In each direction the same four options must give a
↪  UNIQUE answer for both receivers.
```

The remaining fixed constraints preserve the two original relations, require two additional evidence-supported relations of the same type, retain an evidence ID and continuous source span for every relation, generate both directions for receivers 0 and 1, and require a unique answer together with a reason that each distractor fails the target conditions.

C RQ1 Protocol and Matched-Pair Extension

C.1 Dataset Structure and Primary Answer Assignment

After the synthesis and evidence-validation procedure in Appendix A, a record enters the frozen RQ1 file only if it belongs to one of the six target countries, has a nonempty `source_record_id`, contains exactly one row for each of the four task slots, and has no duplicate record–slot row. Records that fail the final structural or traceability checks are excluded. The dataset retains evidence status and repair history for each included record.

For a country c, records are placed in a stable hash order. Let $q(c, r)$ be the index of record r in that order and $\iota(s)$ a fixed index for slot s. The primary gold label is assigned by

$$L_{c,r,s}^{\text{primary}} = \text{ABCD}[(q(c, r) + \iota(s)) \bmod 4] .$$

Reordering changes only display labels, the gold label, and resulting token sequence. Option text, internal option identifier, and evidence provenance are unchanged. The four additional full renderings place the same gold option at each answer label and are stored as derived variants of the same semantic QA.

C.2 Country Ablation and Scoring Records

Country ablation uses case-insensitive word-boundary replacement. A nationality adjective is replaced by *cultural*, while a country name is replaced by *the target country*. The replacement ledger

Table 5: RQ1 balanced 2×2 pair-recombination results. Binary items are derived from the original four-option panel. Accuracy is forced-choice A/B accuracy equally weighted across the six countries; its chance baseline is 50%. Pairing advantage is the log probability difference between the evidence-supported and swapped options, measured in nats. Because probability calibration differs across models, absolute pairing advantages should not be used to rank models.

Model	Direction	Accuracy	Pairing advantage [95% CI]	Holm-adjusted p
Qwen2.5-7B	$E \to P$	80.25%	$7.327\ [6.087,\ 8.563]$	$\approx 6.0 \times 10^{-5}$
Qwen2.5-7B	$P \to E$	80.37%	$7.316\ [6.069,\ 8.586]$	$\approx 6.0 \times 10^{-5}$
Llama-3.1-8B	$E \to P$	74.29%	$0.670\ [0.517,\ 0.830]$	$\approx 6.0 \times 10^{-5}$
Llama-3.1-8B	$P \to E$	69.94%	$0.649\ [0.508,\ 0.800]$	$\approx 6.0 \times 10^{-5}$
Mistral-7B	$E \to P$	74.60%	$6.521\ [5.258,\ 7.768]$	$\approx 6.0 \times 10^{-5}$
Mistral-7B	$P \to E$	74.49%	$6.542\ [5.405,\ 7.687]$	$\approx 6.0 \times 10^{-5}$

stores every matched string and replacement. The procedure intentionally leaves implicit cultural cues, including festivals, places, languages, and entity names, unchanged.

For every model input we retain the rendered prompt, tokenization, answer anchor, exact label-continuation token IDs, condition, position variant, gold label, option identifiers, and option-level evidence provenance. Each option is traceable as

$$(\text{option ID}, \text{display label}, \text{evidence IDs}, \text{continuous evidence spans}).$$

Scoring outputs preserve the A–D label log probabilities, prediction, correctness, gold probability, and margin. Aggregation verifies that both directions or all four positions are complete before computing joint metrics.

C.3 Balanced Two-by-Two Extension

To evaluate pair discrimination while controlling entity and practice marginal frequencies, we construct a balanced 2×2 pair-recombination experiment. Each complete block selects two semantically matched records $r_i = (C_i, E_i, P_i, K_i, S_i)$, $i \in \{1, 2\}$, with $R_{S_i}(E_i, P_i \mid C_i, K_i) = 1$, and construct

	P_1	P_2
E_1	valid	swapped
E_2	swapped	valid

.

Both valid pairs must be supported by accessible continuous source text. The swapped cells change only the pairing while matching country, template, answer position, role grammar, candidate count, token length, and surface plausibility as closely as possible. A complete pair block is the indivisible sampling and statistical unit; its four cells cannot cross partitions.

From the original four-option panel of 2,184 records, we randomly sample records and construct 75 complete pair blocks. For each derived item, we retain only the original evidence-supported option and the swapped option from the same matched block, removing the two original foils that are not part of the pairing contrast. Each block contains four pairing cells, with forced-choice A/B items in both the $E \to P$ and $P \to E$ directions, yielding $75 \times 4 \times 2 = 600$ QA items.

The four-option task measures retrieval from a broader candidate set, whereas the binary task directly contrasts the evidence-supported and swapped pairings while holding entity and practice marginal frequencies constant. Its chance baseline is 50%.

Each of the three models scores all 600 frozen ablation items, for 1,800 model scores overall; after the ablation set is frozen, no questions are modified and no problematic samples are excluded. Accuracy is computed within each country and then equally averaged across the six countries. Pairing advantage is the log probability difference, in nats, between the evidence-supported and swapped options:

$$\Delta_{\text{pair}} = \log P_\theta(y_{\text{supported}} \mid x) - \log P_\theta(y_{\text{swapped}} \mid x).$$

Confidence intervals use 10,000 country-stratified complete-block bootstrap resamples. Significance is assessed using 100,000 block-level sign-flip permutations, with Holm correction across the six model–direction tests.

All six model–direction confidence intervals lie above zero, showing that the models bidirectionally prefer the original evidence-supported pairings within this candidate set. For Mistral, the reported scores use A/B label continuations retokenized in the exact prompt context, including the leading-space token boundary where applicable. The official results are therefore those in Table 5: 74.60% accuracy and a 6.521 pairing advantage for $E \to P$, and 74.49% accuracy and a 6.542 pairing advantage for $P \to E$.

D Supplementary RQ2 Readability Protocol

Let $\delta \in \{E \to P, P \to E\}$ denote retrieval direction and $\mathcal{V} = \{\text{primary}, A, B, C, D\}$ the five full-QA renderings. For record i, rendering v, and layer l, the context-conditioned states are

$$e_{i,v}^{(\delta)} = \text{meanpool}\Big(H_{i,v}^{(l,\delta)}[I_{E,i,v}^{(\delta)}]\Big), \quad p_{i,v}^{(\delta)} = \text{meanpool}\Big(H_{i,v}^{(l,\delta)}[I_{P,i,v}^{(\delta)}]\Big), \quad h_{i,v}^{(\delta)} = H_{i,v}^{(l,\delta)}[a_{i,v}^{(\delta)}]. \tag{6}$$

All states come from the same full-QA pass and are averaged across renderings:

$$(e_i^{(\delta)}, p_i^{(\delta)}, h_i^{(\delta)}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} (e_{i,v}^{(\delta)}, p_{i,v}^{(\delta)}, h_{i,v}^{(\delta)}). \tag{7}$$

Inputs and targets are standardized dimension-wise using training-set means and standard deviations, with a 10^{-4} floor, and predictions are transformed back to the original hidden space. For direction δ, the format, affine-additive, and low-rank interaction models are

$$\widehat{h}_{\text{fmt}}^{(\delta)} = W_C^{(\delta)} c^{(\delta)} + b_C^{(\delta)}, \qquad \widehat{h}_{\text{add}}^{(\delta)} = W_E^{(\delta)} e^{(\delta)} + W_P^{(\delta)} p^{(\delta)} + b^{(\delta)}, \tag{8}$$

$$\widehat{h}_{\text{int}}^{(\delta)} = \widehat{h}_{\text{add}}^{(\delta)} + U^{(\delta)} \Big[(A_E^{(\delta)} e^{(\delta)}) \odot (A_P^{(\delta)} p^{(\delta)}) \Big]. \tag{9}$$

The low-rank model estimates rank-16 PCA projections on training data and maps the multiplicative interaction to the additive model's training residual.

The reconstruction metrics are

$$\text{MSE} = \frac{1}{N D_h} \sum_{i=1}^{N} \|\widehat{h}_i - h_i\|_2^2, \qquad R^2_{\text{centered}} = 1 - \frac{\sum_i \|\widehat{h}_i - h_i\|_2^2}{\sum_i \|h_i - \bar{h}\|_2^2}. \tag{10}$$

For pair ranking, let D_h be the hidden dimension. In $E \to P$ we hold the entity fixed and replace the practice; in $P \to E$ we hold the practice fixed and replace the entity:

$$s_{i,j}^{E \to P} = -\frac{1}{D_h} \left\| \widehat{h}^{E \to P}(e_i^{E \to P}, p_j^{E \to P}) - h_i^{E \to P} \right\|_2^2, \tag{11}$$

$$s_{i,j}^{P \to E} = -\frac{1}{D_h} \left\| \widehat{h}^{P \to E}(e_j^{P \to E}, p_i^{P \to E}) - h_i^{P \to E} \right\|_2^2. \tag{12}$$

Here $j = i$ is the valid relation and $j \neq i$ a same-country, same-direction swap. Pair ranking and Recall@1 are computed from these scores. Each country–direction test pool contains 73 candidates. For the format control, all 73 scores are tied by construction and Recall@1 uses uniform fractional tie credit.

Complete cultural records are indivisible train/development/test units. Both directions, five full renderings, and one token-matched format control from one record remain in the same split. The development set selects the analysis layer by additive valid-versus-swapped ranking, breaking ties by higher centered R^2 and lower MSE. Pairing candidates are restricted to test records from the same country and direction, and bootstrap resampling is clustered by complete record family within country.

D.1 Full-Data Cross-Template Pair Readout

This supplementary analysis uses all 75 cultural pair blocks and three frozen language models. Each block contains two evidence-supported diagonal relations and their swapped pairings, rendered

Table 6: Complete RQ2 composer comparison at the development-selected layer. Reconstruction values use three decimals for centered R^2 and six decimals for MSE; ranking metrics use three decimals.

Model	Composer	Layer	Hidden R^2	MSE	Pair ranking	$E \to P$ R@1	$P \to E$ R@1
Qwen2.5-7B	Format control	19	-0.381	0.064646	0.500	0.014	0.014
Qwen2.5-7B	Affine additive	19	0.392	0.028437	0.981	0.809	0.883
Qwen2.5-7B	Additive + low-rank	19	0.392	0.028445	0.981	0.809	0.883
Llama3.1-8B	Format control	18	-0.297	0.001250	0.500	0.014	0.014
Llama3.1-8B	Affine additive	18	0.230	0.000743	0.899	0.500	0.532
Llama3.1-8B	Additive + low-rank	18	0.230	0.000743	0.900	0.500	0.532
Mistral-7B	Format control	15	-0.295	0.000209	0.500	0.014	0.014
Mistral-7B	Affine additive	15	0.402	0.000096	0.964	0.707	0.793
Mistral-7B	Additive + low-rank	15	0.402	0.000096	0.964	0.707	0.793
Qwen3-8B	Format control	25	-0.159	0.219512	0.500	0.014	0.014
Qwen3-8B	Affine additive	25	0.344	0.124282	0.921	0.500	0.681
Qwen3-8B	Additive + low-rank	25	0.344	0.124282	0.921	0.500	0.681
Qwen3.5-9B	Format control	18	-0.028	0.009619	0.500	0.014	0.014
Qwen3.5-9B	Affine additive	18	0.497	0.004709	0.947	0.660	0.638
Qwen3.5-9B	Additive + low-rank	18	0.494	0.004737	0.946	0.638	0.649
Phi-4-mini	Format control	17	-0.295	0.022996	0.500	0.014	0.014
Phi-4-mini	Affine additive	17	0.401	0.010629	0.936	0.670	0.670
Phi-4-mini	Additive + low-rank	17	0.401	0.010629	0.936	0.670	0.670
Gemma-4-12B	Format control	5	0.630	0.000303	0.500	0.014	0.014
Gemma-4-12B	Affine additive	5	0.084	0.000750	0.771	0.404	0.287
Gemma-4-12B	Additive + low-rank	5	0.078	0.000755	0.771	0.404	0.287

Table 7: RQ2 full-data cross-template pair-readout results. The main condition is fixed to T1+T2 training and T3 testing. Original-pair, context, and residual columns report complete four-cell assignment accuracy. Δ Context is the paired original-minus-context linear difference on the same test blocks.

Model	Original-pair linear (95% CI)	Context linear	Δ Context (95% CI)	Residual MLP
Qwen2.5-7B	84.72% $[73.61, 90.28]$	40.28%	+44.44 $[25.00, 58.33]$	84.72%
Llama-3.1-8B	83.33% $[69.44, 94.44]$	54.17%	+29.17 $[13.89, 47.22]$	78.06%
Mistral-7B	88.89% $[77.78, 94.44]$	47.22%	+41.67 $[25.00, 50.00]$	82.22%

under templates T1, T2, and T3, for 900 pair–template records. The train, development, and test splits contain 46, 11, and 18 blocks, respectively; the test split covers nine source groups. The main condition is fixed in advance: readouts are trained on T1 and T2 and evaluated on the unseen T3 template, without selecting the training condition on test results.

For each 2×2 block, the readout produces four scores $s_{11}, s_{12}, s_{21}, s_{22}$. A complete assignment is correct when $s_{11} + s_{22} > s_{12} + s_{21}$, with ties receiving half credit. This metric measures complete pair assignment rather than QA answer accuracy. Results are averaged equally across blocks within country and then equally across the six countries. Pointwise 95% intervals for the original-pair linear readout and its paired difference from the linear context control resample source groups within country. The test set contains nine source groups. Residual MLP values are means over five random seeds.

All original-pair linear intervals lie above chance, and all paired differences from the context control lie above zero. Thus, the true-pair states contain relational correspondence information that transfers from T1/T2 to T3 and exceeds the shared-context readout. Residual MLP results further show that pair-relevant information remains readable in residual representations.

E Supplementary RQ3 Controller and Intervention Protocol

For each training sample and answer-position variant, the supervision target is the gradient of the gold-versus-foils margin with respect to the analysis-layer answer-anchor residual. Gradients are averaged across the four answer positions within a record. For each model, layer, and direction, cue and gradient means, normalization statistics, and PCA bases are estimated from training data,

and a direction-specific controller is fit in the low-rank coordinates. The target norm ρ_d is the median training-gradient norm for that direction, so test-time sample variation comes from predicted direction rather than unconstrained magnitude.

For an intervention at layer l, the clean-relative margin effect is

$$\Delta M_i = M_\theta\Big(x_i; \mathrm{do}\Big(H_i^{(l)}[a_i] \leftarrow \widetilde{H}_i^{(l)}[a_i]\Big)\Big) - M_\theta(x_i). \tag{13}$$

For matched control q, four-position selectivity is summarized by

$$D_i^q = \Delta\bar{M}_i^{\mathrm{directional}} - \Delta\bar{M}_i^q. \tag{14}$$

The primary cue-conditioned contrast compares the directional intervention with source shuffling, alongside pair-swapped, opposite-direction, random-subspace, and global-gradient controls.

The predefined grid is $r \in \{4, 8, 16, 32\}$ and $\alpha \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$. Development selection first maximizes directional-minus-shuffled margin change, then directional margin change and accuracy change, with smaller rank as the final tie-breaker. Three seeds fixed in advance independently repeat the train–development–test procedure; all three runs contribute to the reported estimates.

The source-shuffled control deranges source records within country and direction; the opposite-direction control uses the reverse controller; and the random-subspace control samples from the training-gradient PCA subspace. Every non-clean vector is renormalized to the recipient direction's ρ_d and uses the same α. Inference uses paired record-level bootstrap resampling within country, retaining both directions, all four answer positions, and all intervention conditions from each record.

F Four-Choice RQ3 Intervention Analysis

F.1 Protocol

The analysis contains 300 semantic four-choice QAs from six countries, evenly divided between entity-to-practice and practice-to-entity retrieval. The data comprise 75 matched-pair blocks and 43 source groups. Source dependencies yield 15 training, 6 development, and 9 test inference groups; source-connected groups, rather than option rotations or controller seeds, are the aggregation units.

We evaluate `Qwen2.5-7B-Instruct`, `Meta-Llama-3.1-8B-Instruct`, and `Mistral-7B-Instruct-v0.3` at residual layers 19, 18, and 15, respectively. For each retrieval direction, the controller has the form

$$v_d(z) = v_d^{\mathrm{global}} + B_d \tanh(A_d z).$$

It is trained from the gradient of the gold-answer margin on training records. Controller rank is selected from $\{4, 8, 16, 32\}$ and intervention scale from $\{0.05, 0.1, 0.2, 0.5, 1\}$ using development-set directional-minus-shuffled margin. Controller seeds 101, 203, and 307 are fixed in advance.

The evaluated conditions are clean, directional, source-shuffled, global gradient, opposite direction, pair-swapped, and random subspace. Source shuffling preserves country, retrieval direction, controller subspace, and vector norm while replacing the source cue. Pair swapping exchanges controller vectors between the two records in a matched block. T3 changes only the question wording following the cultural cue.

F.2 Aggregate Margin Effects

Directional interventions increase the gold-answer margin relative to clean in all six model–template conditions; D–R ranges from +1.144 to +3.949. Qwen additionally shows a positive contrast against the opposite-direction control, whereas Llama and Mistral outperform pair-swapped interventions. Mistral also shows a positive contrast in the tested source-shuffled condition.

F.3 Direction-Specific Effects

The most pronounced source-conditioned effect occurs for Mistral in the entity-to-practice direction. Directional intervention exceeds source shuffling by +0.945 under T1 and +0.975 under T3,

Table 8: Four-choice margin contrasts averaged across three controller seeds and source-connected test groups. D denotes directional intervention; S, G, P, O, and R denote source-shuffled, global-gradient, pair-swapped, opposite-direction, and random-subspace controls.

Model	Template	D–Clean	D–S	D–G	D–P	D–O	D–R
Qwen2.5-7B	T1	+1.102	-0.116	-0.772	-0.016	+0.311	+1.157
Qwen2.5-7B	T3	+1.071	-0.133	-0.786	-0.071	+0.249	+1.144
Llama-3.1-8B	T1	+3.673	-0.296	-1.652	+0.475	-1.635	+3.949
Llama-3.1-8B	T3	+3.215	-0.291	-1.600	+0.511	-1.604	+3.511
Mistral-7B	T1	+1.168	+0.143	-2.125	+0.522	-0.625	+3.657
Mistral-7B	T3	+1.296	+0.292	-2.586	+0.453	-0.373	+3.821

and exceeds pair swapping by +0.754 and +0.794, respectively. In the practice-to-entity direction, directional intervention continues to increase margin relative to clean inference, with changes of +1.105 on T1 and +1.296 on T3.

Qwen shows positive clean-relative effects in both directions. Under T1, the changes are +1.286 for entity-to-practice and +0.918 for practice-to-entity; the corresponding T3 effects are +1.197 and +0.946. Llama likewise shows positive clean-relative effects in both directions, ranging from +3.061 to +3.802 across templates.

F.4 Question-Wording Transfer

The intervention effects are similar between the original T1 questions and the T3 paraphrases. Averaged across directions, D–Clean changes from +1.102 to +1.071 for Qwen, from +3.673 to +3.215 for Llama, and from +1.168 to +1.296 for Mistral. This consistency indicates that the learned intervention effects transfer across changes in the recipient question wording while the cultural cue remains fixed.

F.5 Dose Sensitivity

For Qwen, directional margin changes increase from +0.552 at half scale to +1.102 at the selected scale and +2.145 at double scale, while reversing the direction produces a change of -1.065. This sign reversal and positive-scale growth demonstrate a directionally coherent intervention effect. Llama also shows positive effects at half, selected, and double scales (+2.909, +3.673, and +2.624). Mistral reaches its largest directional margin change at half scale (+1.548), followed by +1.168 at the selected scale, indicating a nonlinear dose response.

F.6 Summary

Across the tested models, controller seeds, retrieval directions, and two question renderings, cue-conditioned intervention produces positive clean-relative margin changes and exceeds norm-matched random-subspace controls. The matched-control contrasts reveal model- and direction-dependent intervention profiles, with positive source-shuffled and pair-swapped contrasts in specific settings. Global-gradient effects are larger in all six aggregate model–template cells, indicating a substantial shared task-level signal.

G Additional Experimental Details and Results

G.1 RQ1 Setup and Audit Details

G.1.1 Data and models

The unified RQ1 freeze covers Brazil, Canada, China, Germany, Iran, and Japan. Each retained record contains all four task types. The resulting panel contains 364 complete records and 1,456 QA items per country, for 2,184 records and 8,736 QA items overall.

The aggregate RQ1 comparison contains `Qwen2.5-7B-Instruct`, `Meta-Llama-3.1-8B-Instruct`, `Mistral-7B-Instruct-v0.3`, `Qwen3-8B`, `Qwen3.5-9B`,

`Phi-4-mini`, and `Gemma-4-12B`. Supplementary RQ1 analyses also include Ministral-3-8B (Liu et al., 2026). The country-level analysis reported below covers Qwen2.5-7B, Qwen3-8B, Llama3.1-8B, and Ministral-3-8B. Every question has four options. For the primary rendering, gold positions are assigned by a stable country-level record ordering and a four-slot Latin-square rotation, balancing A, B, C, and D while preserving option identities and evidence provenance.

Each model uses its native chat template and tokenizer while receiving the same single-user-message content: a fixed selection instruction, question, A–D options, and `Answer:` prefix. We do not inject an additional system prompt. Frozen inputs retain the rendered text, messages, token IDs, token count, answer anchor, prompt hash, labels, option identifiers, and option-level evidence provenance.

Before next-token scoring, each answer label is retokenized in its exact prompt continuation context. The single-token protocol is used only when the prompt token sequence remains a prefix and the continuation adds exactly one token for each of A, B, C, and D. Otherwise, scoring must use the full continuation sequence log probability rather than truncating a multi-token label.

G.1.2 Experimental conditions

Each primary QA is evaluated under four conditions:

1. `full`: the complete question stem and options;
2. `format_control`: the options and formatting without the question stem;
3. `token_matched_format_control`: the question is removed and neutral punctuation is inserted before the `Answer:` prefix until the tokenizer-specific input length matches the full question;
4. `explicit_country_ablation`: explicit country names and nationality adjectives are replaced by neutral phrases in both the question and options, while festivals, place names, and cultural terms remain.

Each semantic QA additionally has four full-input answer-position renderings, placing the same gold option at A, B, C, and D. All four renderings remain grouped under the same semantic QA. They accompany the full-input condition, while each control uses its specified rendering. The frozen RQ1 input set therefore contains

$$8{,}736 \times (4 \text{ primary conditions} + 4 \text{ full position variants}) = 69{,}888$$

rows per model.

G.1.3 Evaluation metrics

At the answer anchor, logits for the exact A–D label continuations are normalized within the four candidates. We report accuracy, gold probability, gold-versus-foils margin, and the proportion of positive margins. For a gold option g and three foils f_1, f_2, f_3, the margin is

$$M = \log P(g) - \operatorname{logsumexp}(\log P(f_1), \log P(f_2), \log P(f_3)) .$$

For record i, let F_i indicate whether the entity-to-practice question is correct and R_i whether the practice-to-entity question is correct. We define bidirectional joint accuracy (BJA) as

$$\mathrm{BJA} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(F_i = 1 \wedge R_i = 1),$$

where $N = 364$ is the number of complete records in each country. The descriptive reference under independent uniform guessing in two four-option directions is $(1/4)^2 = 1/16 = 6.25\%$; BJA itself is an observed record-level conjunction and does not assume that the two model responses are independent.

To measure answer-position stability, we report the proportion of semantic questions answered correctly in all four gold positions and the proportion of records for which both $E \to P$ and $P \to E$ are correct in all four positions. The latter is our stricter primary stability metric. We also retain the mean number of correct positions and correctness invariance to expose partial position sensitivity.

Table 9: Six-country RQ1 results. Model labels specify the model generation/version and parameter scale. Four-pos. is the proportion of questions answered correctly in all four gold positions; Bi-stable requires both retrieval directions to be correct in all four positions.

Model	Country	Acc.	Margin	$E \to P$	$P \to E$	BJA	Four-pos.	Bi-stable
Qwen2.5-7B	Brazil	0.8603	9.0141	0.9118	0.8824	0.8235	0.7647	0.6765
Qwen2.5-7B	Canada	0.9565	13.1455	0.9783	0.9457	0.9239	0.9321	0.8804
Qwen2.5-7B	China	0.8422	8.9025	0.8738	0.8544	0.7379	0.7306	0.5437
Qwen2.5-7B	Germany	0.9423	12.5489	0.9451	0.9231	0.8681	0.9038	0.8132
Qwen2.5-7B	Iran	0.8700	10.9120	0.8933	0.8000	0.7200	0.7967	0.5600
Qwen2.5-7B	Japan	0.9653	13.3179	0.9444	0.9444	0.8889	0.9236	0.7917
Qwen3-8B	Brazil	0.7794	6.8383	0.8235	0.8529	0.6765	0.6544	0.4412
Qwen3-8B	Canada	0.9239	13.4155	0.9239	0.8913	0.8152	0.8750	0.7065
Qwen3-8B	China	0.8034	7.1148	0.8641	0.8350	0.7087	0.6238	0.4369
Qwen3-8B	Germany	0.9231	12.9805	0.9121	0.9121	0.8352	0.8214	0.6703
Qwen3-8B	Iran	0.8400	10.5112	0.8400	0.8133	0.6933	0.7733	0.5200
Qwen3-8B	Japan	0.9236	13.1555	0.9583	0.9028	0.8611	0.8854	0.7778
Llama3.1-8B	Brazil	0.7426	0.6226	0.7353	0.6765	0.5294	0.5588	0.2647
Llama3.1-8B	Canada	0.9321	2.9613	0.9565	0.8804	0.8370	0.8750	0.7391
Llama3.1-8B	China	0.7718	1.0283	0.8155	0.7573	0.6311	0.6262	0.4563
Llama3.1-8B	Germany	0.9176	2.4822	0.9231	0.9121	0.8571	0.8434	0.7473
Llama3.1-8B	Iran	0.8433	1.9506	0.8133	0.8667	0.6800	0.7367	0.4933
Llama3.1-8B	Japan	0.8819	2.6710	0.9028	0.8611	0.7778	0.8160	0.6528
Ministral-3-8B	Brazil	0.8015	1.9021	0.8235	0.8529	0.7059	0.7206	0.5294
Ministral-3-8B	Canada	0.9321	3.3708	0.9565	0.8696	0.8370	0.8533	0.6630
Ministral-3-8B	China	0.7573	1.4711	0.7767	0.7573	0.6117	0.6092	0.4466
Ministral-3-8B	Germany	0.9286	3.1357	0.9560	0.8901	0.8571	0.8324	0.6374
Ministral-3-8B	Iran	0.8533	2.5609	0.8667	0.8400	0.7200	0.7600	0.5600
Ministral-3-8B	Japan	0.9271	2.9307	0.9306	0.9167	0.8611	0.8438	0.6667

Uncertainty is estimated with 2,000 record-cluster bootstrap draws within each country. A sampled record carries all four slots, all conditions, and all position variants together. Full-versus-control comparisons are paired on the same QA within these record clusters; no condition or slot-derived observation is treated as an independent cultural fact.

G.2 Additional RQ1 Results

For the four-model subset with available country-ablation and country-level breakdowns, removing explicit country words reduces aggregate accuracy by approximately 0.6 percentage points for Qwen2.5-7B, 1.6 points for Qwen3-8B, and 2.4 points for Llama3.1-8B, and 1.9 points for Ministral-3-8B. This pattern indicates that the models do not rely solely on words such as *Brazilian* or *Canadian*; festivals, place names, cultural terms, and entity–practice information in the remaining stem continue to support accuracy after the 0.6–2.4 point ablation changes.

Table 9 presents the four-model country-level breakdown. Mistral-7B, Qwen3.5-9B, Phi-4-mini, and Gemma-4-12B are reported in the six-country aggregate summary and are not included in the country-level comparison. BJA is above the 6.25% chance level in every reported model–country cell, although the stricter bidirectional four-position stability measure is lower than single-run accuracy.

Three patterns recur across the evaluated models. First, full-input performance greatly exceeds the token-matched format control, showing that semantic stem information provides an important task signal. Second, both retrieval directions achieve high accuracy, and BJA is well above chance: it ranges from 0.7200 to 0.9239 for Qwen2.5-7B, from 0.6765 to 0.8611 for Qwen3-8B, from 0.5294 to 0.8571 for Llama3.1-8B, and from 0.6117 to 0.8611 for Ministral-3-8B. This shows that models can use one component to retrieve the other. Third, Canada, Germany, and Japan generally show higher performance, while China, Iran, and Brazil show larger cross-model differences. Answer-position stability remains below single-run accuracy, quantifying sensitivity to label position.

Table 10: Ministral-3-8B RQ1 results by treatment slot.

Slot	Acc.	Gold prob.	Margin
Entity	0.8651	0.8006	2.5173
$E \to P$	0.8887	0.8257	2.6315
$P \to E$	0.8501	0.7920	2.3464
Norm/procedure	0.8758	0.8278	2.9094

Table 11: Ministral-3-8B stability across forced gold positions.

Gold position	Acc.	Margin
A	0.9368	3.8558
B	0.8646	2.5774
C	0.8180	1.6006
D	0.8533	2.3912

G.3 RQ2 Extraction and Fitting Details

G.3.1 Data and record-level splits

RQ2 uses only the `entity_to_practice` and `practice_to_entity` questions. The RQ1 panel provides 2,184 complete cultural records, giving 2,184 samples in each direction and 4,368 directional samples overall. Within each country, a stable hash with a fixed seed assigns complete cultural records to 218/73/73-record train, development, and test partitions, respectively. Across six countries, the splits contain 1,308/438/438 records, or 2,616/876/876 directional samples. Both directions, five full renderings, and one token-matched format-control rendering derived from one record remain in the same partition. The independent statistical unit is therefore the cultural record, not an option, token, direction, or rendering.

Training data estimate standardization statistics, ridge maps, and PCA projections, development data select the analysis layer, and held-out test records provide the reported metrics.

G.3.2 Models and hidden-state extraction

We analyze `Qwen2.5-7B-Instruct`, `Meta-Llama-3.1-8B-Instruct`, `Mistral-7B-Instruct-v0.3`, `Qwen3-8B`, `Qwen3.5-9B`, `Phi-4-mini`, and `Gemma-4-12B`. All seven models are included in the format-control, affine-additive, and low-rank comparisons at their development-selected layers. The first three additionally receive the full layerwise analysis and the cross-template correspondence readout. The user-visible instruction, question, options, and `Answer:` prefix are identical across models; each model uses its own tokenizer and native chat template. For every input, we store token offsets, component spans, answer anchor, prompt hash, and the hidden states from the embedding layer through the final transformer block. Each model uses its own tokenizer, hidden states, normalization statistics, and composer parameters, while record partitions follow the common data split. Entity and practice span representations are extracted from the same full directional-QA pass and are context-conditioned.

For each model and layer, coordinate-wise means and standard deviations are estimated from training samples only; standard deviations are floored at 10^{-4}. Linear composers minimize standardized hidden reconstruction error with ridge penalty $\lambda = 1.0$.

G.3.3 Layer selection and held-out metrics

At each available layer we fit the format-control, additive, and low-rank interaction composers independently. For each language model, the analysis layer is selected using development data only: we first maximize additive valid-versus-swapped ranking accuracy, break ties by hidden R^2, and then by lower MSE. All composers are refit on the training split at this fixed layer.

The available hidden-state indices are 0–28 for Qwen2.5-7B and 0–32 for Llama3.1-8B and Mistral-7B. At index 0, the answer-anchor embedding target has zero variance across samples, so centered

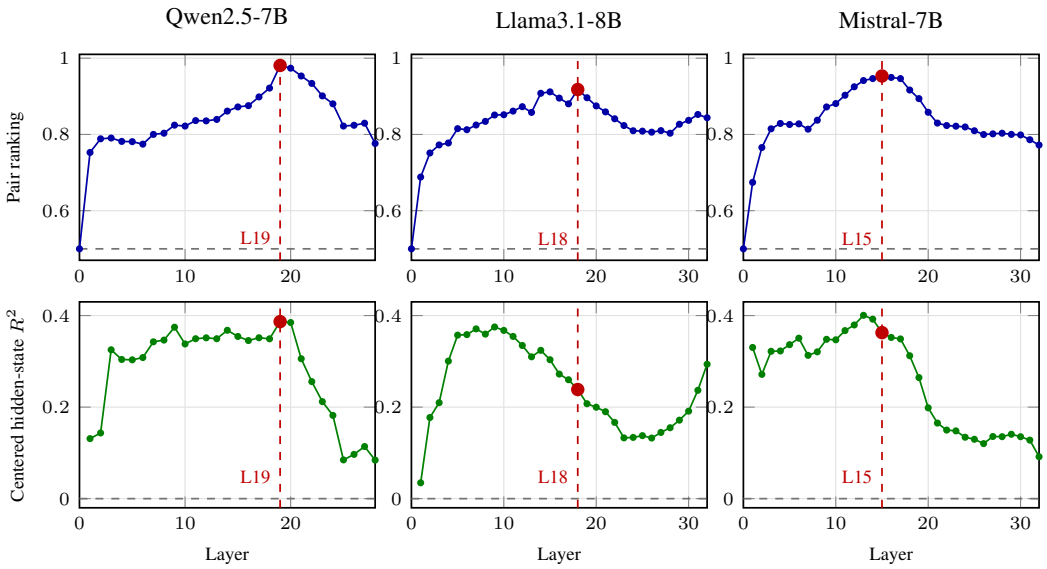


Figure 2: Development-set layer-selection curves for the affine additive composer. The upper row shows valid-versus-swapped pair ranking, including the valid layer-0 chance value of 0.5; the lower row shows centered hidden-state R^2, with layer 0 omitted because zero target variance makes R^2 undefined. Red lines and points mark the selected layers. Qwen2.5-7B ends at layer 28, whereas Llama3.1-8B and Mistral-7B end at layer 32.

R^2 is undefined and is recorded as NA rather than used for layer selection. The development protocol selects layer 19 for Qwen2.5-7B, layer 18 for Llama3.1-8B, and layer 15 for Mistral-7B.

On the held-out test set, we report MSE, centered hidden R^2, valid-versus-swapped ranking accuracy, and Recall@1 separately for $E \to P$ and $P \to E$. Candidate pools contain the 73 test records from the query's country and direction, with one candidate per cultural record. For the format control, all candidate scores are tied and Recall@1 uses uniform fractional tie credit. Low-rank interaction contributions are reported as paired differences from their identical additive bases. Uncertainty is estimated by resampling complete records within country, keeping both directions and all derived inputs together in each bootstrap replicate. We report pooled and country-level estimates.

G.4 RQ3 Fitting and Inference Details

G.4.1 Data, Models, and Intervention Layer

The RQ3 panel contains the same 2,184 cultural records and 4,368 directional QA samples as RQ2. The per-country 218/73/73-record split yields 1,308/438/438 records and 2,616/876/876 directional samples across the six countries. All QA inputs derived from one cultural record inherit its partition. We evaluate `Qwen2.5-7B-Instruct`, `Meta-Llama-3.1-8B-Instruct`, and `Mistral-7B-Instruct-v0.3`; each model uses only its own tokenizer, hidden states, gradients, PCA bases, controllers, and residual hooks. All language-model parameters remain frozen.

For each model, RQ3 inherits the analysis layer selected by the RQ2 development protocol. We verify the model-specific correspondence between hidden-state index l and decoder-block output $l-1$ before intervention. Inputs are checked for the pre-option cue constraint, answer-anchor position, and label-token protocol before evaluation.

G.4.2 Controller Fitting and Development Selection

For each direction, train-only margin gradients are averaged across the four answer positions and used to fit the controller described above. We consider the predefined rank and intervention-scale

grids

$$r \in \{4, 8, 16, 32\}, \qquad \alpha \in \{0.05, 0.1, 0.2, 0.5, 1.0\}.$$

Controller parameters are optimized for 600 Adam steps with learning rate 0.03 and regularization coefficient 10^{-3}. Each candidate is evaluated on development records using both `directional_caa` and `source_shuffled_caa`. We select one (r, α) pair per model by first maximizing the directional-minus-shuffled mean margin change, then the directional margin change, then the directional accuracy change, and finally preferring the smaller rank. The selected configuration is evaluated on the test records.

To assess initialization sensitivity, the full train–development–test pipeline is repeated with controller seed bases fixed in advance: $\{101, 203, 307\}$. Main point estimates average all three replications, with per-seed results and ranges retained in the accompanying artifacts.

G.4.3 Behavioral endpoints and statistical inference

For every evaluated QA rendering and condition, we save the four label logits, predicted and gold labels, accuracy, margin, margin change from clean inference, country, direction, record identifier, layer, block, rank, α, vector norm, and any shuffled or opposite-direction source identifier. The test split contains 876 directional QA observations from 438 records. Across four answer positions, clean-versus-directional evaluation produces $876 \times 4 \times 2 = 7{,}008$ forward-condition rows per model, while the complete six-condition specificity suite produces $876 \times 4 \times 6 = 21{,}024$ rows per model and controller seed. Derived renderings are clustered with their source record for aggregation and inference.

Inference uses paired, record-level bootstrap resampling. Each replicate samples complete cultural records within country and keeps both directions, all four positions, and all six conditions together. We use 5,000 replicates for pooled estimates and 1,000 for country-level estimates. The paired selectivity contrast follows Equation equation 14, where $\Delta\bar{M}_i$ averages the four answer positions. The predefined criterion requires the lower 95% confidence bound of the directional effect relative to clean to exceed zero in both retrieval directions and the directional intervention to exceed the source-shuffled control. Comparisons with the global, opposite-direction, and random-subspace conditions provide additional direction and subspace specificity tests.

Accuracy tests. Table 4 reports the pooled two-sided exact McNemar tests associated with the three-seed main analysis. Seed-specific directional transition counts and p values are retained separately from the seed-averaged point estimates. Record-level accuracy confidence intervals are reported after pooling both retrieval directions.

G.5 RQ3 Repair Metric Definitions

Let N denote the number of QA observations. $W{\to}C$ and $C{\to}W$ denote corrected errors and induced errors, respectively.

Repair recall is the fraction of initially incorrect answers repaired by the intervention, $W{\to}C/(N - \text{clean correct})$, while repair precision is $W{\to}C/(W{\to}C + C{\to}W)$; repair F1 is their harmonic mean. Preservation is $(\text{clean correct} - C{\to}W)/\text{clean correct}$.

G.6 RQ3 Seed Aggregation

The main RQ3 accuracy, uncertainty, and categorical tests are reported in Table 4 after aggregation across the three fixed controller seeds. Country-level integer repair and regression counts, together with country–direction outputs, are retained per seed in the accompanying artifacts.

G.6.1 Seed-Averaged Continuous Effects

ΔMargin measures the change in the gold label's logit advantage over the log-sum-exp of the three foils. ΔGold Probability is the change in the gold label's probability after normalizing over A–D, and ΔNLL is the corresponding change in negative log likelihood.

Table 12: Three-seed-average continuous RQ3 effects pooled across both retrieval directions. Each seed evaluates $N_{\mathrm{QA}} = 876$. Positive ΔMargin and ΔGold Probability, and negative ΔNLL, indicate improvement.

Model	ΔMargin	ΔGold Prob.	ΔNLL
Qwen2.5-7B	+0.079878	+0.004860	-0.022990
Meta-Llama-3.1-8B	+0.614228	+0.036430	-0.105578
Mistral-7B	+2.086142	+0.019460	-0.183144

H Authoritative Cultural Evidence Sources

This appendix lists the principal sources in the six-country evidence collection inventory.

H.1 China

- **China Intangible Cultural Heritage Website / Digital Museum of Chinese Intangible Cultural Heritage**
  `https://www.ihchina.cn/`
  A national information platform for intangible cultural heritage, used primarily for evidence on heritage projects, traditional festivals, folk techniques, ritual procedures, and associated roles.
- **Chinaculture / China Culture Website**
  `http://en.chinaculture.org/`
  `http://www.chinaculture.org/`
  A public-facing cultural information platform covering Chinese festivals, food, etiquette, folk customs, and cultural activities.
- **Encyclopedia of China**
  `https://www.zgbk.com/`
  A national encyclopedic knowledge platform used primarily for cultural terminology, historical background, cultural entities, and conceptual descriptions.
- **Cultural Atlas: Chinese Culture**
  `https://culturalatlas.sbs.com.au/chinese-culture`
  An English-language cross-cultural overview from SBS Cultural Atlas, used mainly to supplement checks of communication, etiquette, and sociocultural background.

H.2 Japan

- **Agency for Cultural Affairs, Government of Japan**
  `https://www.bunka.go.jp/`
  Japan's governmental cultural authority, used primarily for cultural heritage, traditional arts, festivals, and cultural policy.
- **Cultural Heritage Online**
  `https://online.bunka.go.jp/`
  A cultural heritage database associated with the Agency for Cultural Affairs, used primarily for artifacts, clothing, artworks, and other material cultural entities.
- **Japan Arts Council / National Theatre of Japan**
  `https://www.ntj.jac.go.jp/`
  Resources from the Japan Arts Council and National Theatre, used primarily for traditional performing arts, associated roles, rituals, and performance procedures.
- **Ministry of Agriculture, Forestry and Fisheries: Our Regional Cuisines**
  `https://www.maff.go.jp/`
  Regional cuisine resources from Japan's Ministry of Agriculture, Forestry and Fisheries, used primarily for festival foods, local cuisines, preparation methods, and eating practices.

- **Japan National Tourism Organization**
  https://www.japan.travel/
  Public cultural guides covering tea ceremony, shrine visits, inns, dining, and festival procedures.
- **Jinja Honcho**
  https://www.jinjahoncho.or.jp/
  First-party materials from a Shinto organization, used primarily for shrines, rites, worship etiquette, and related procedures. Its content is not generalized to other religions or to Japanese culture as a whole.

H.3 IRAN

- **Ministry of Cultural Heritage, Tourism and Handicrafts**
  https://www.mcth.ir/
  https://en.mcth.ir/
  Iran's authority for cultural heritage, tourism, and handicrafts, used primarily for national heritage, traditional crafts, festivals, and regional cultural practices.
- **National Museum of Iran**
  https://irannationalmuseum.ir/en/
  Used primarily for archaeology, historical cultural entities, material culture, and religious-cultural background.
- **Research Institute of Cultural Heritage and Tourism**
  https://richt.ir/
  A government-affiliated cultural heritage research institute, used primarily for local traditions, rituals, handicrafts, and research-oriented heritage descriptions.
- **National Library and Archives of Iran**
  https://www.nlai.ir/
  Used primarily for historical texts, archives, and cultural publications.
- **UNESCO Intangible Cultural Heritage Lists**
  https://ich.unesco.org/en/lists
  Used to verify Iranian intangible cultural heritage entries, practitioner communities, procedures, and geographic scope.
- **Encyclopaedia Iranica**
  https://iranicaonline.org/
  A scholarly encyclopedia used primarily for English terminology disambiguation and supplementary verification of historical, religious, and cultural background.

H.4 GERMANY

- **German Commission for UNESCO**
  https://www.unesco.de/
  Used primarily for intangible cultural heritage, traditional practices, festivals, and socio-cultural activities.
- **Deutsche Digitale Bibliothek**
  https://www.deutsche-digitale-bibliothek.de/
  A digital aggregation platform for German cultural institutions, used primarily for museum objects, historical artifacts, archives, and cultural documents.
- **Stiftung Preussischer Kulturbesitz**
  https://www.preussischer-kulturbesitz.de/
  Used primarily for national museums, artworks, historical collections, and cultural heritage.

- **Deutsches Historisches Museum**
  `https://www.dhm.de/`
  Used primarily for historical events, festivals, sociocultural background, and associated objects.
- **Haus der Geschichte**
  `https://www.hdg.de/`
  Used primarily for contemporary society, public commemoration, and cultural practices in historical contexts.

H.5 Brazil

- **Instituto do Patrimônio Histórico e Artístico Nacional**
  `https://www.gov.br/iphan/`
  Brazil's federal cultural heritage authority, used primarily for intangible heritage, traditional festivals, rituals, local cultural projects, and entity–practice relations.
- **Museu do Ìndio / FUNAI**
  `https://www.gov.br/museudoindio/pt-br/`
  A Brazilian governmental museum and documentation institution for Indigenous peoples, used primarily for the material culture, history, and rituals of specific Indigenous communities. Evidence must preserve the relevant community and regional scope.
- **Fundação Cultural Palmares**
  `https://www.gov.br/pt-br/orgaos/fundacao-cultural-palmares`
  A Brazilian federal institution for Afro-Brazilian culture, used primarily for Afro-Brazilian culture, commemorative dates, history, and associated traditions.
- **Instituto Brasileiro de Museus**
  `https://www.gov.br/museus/`
  Brazil's federal museum institution, used primarily for museum collections, exhibitions, and the historical background of cultural entities.
- **Biblioteca Nacional Digital Brasil**
  `https://bndigital.bn.gov.br/`
  Digital resources from the National Library of Brazil, used primarily for historical texts, historical newspapers, local cultures, and primary documents.
- **Enciclopédia Itaú Cultural**
  `https://enciclopedia.itaucultural.org.br/`
  An editorially curated cultural encyclopedia used primarily for supplementary descriptions of art, music, people, festivals, and cultural terminology.

H.6 Canada

- **Canadian Heritage**
  `https://www.canada.ca/en/canadian-heritage.html`
  Canada's federal authority for culture and heritage, used primarily for national commemorations, cultural programs, multiculturalism, and language and cultural policy.
- **Parks Canada**
  `https://parks.canada.ca/`
  A Canadian federal agency used primarily for national historic sites, regional heritage, and associated cultural places.
- **Library and Archives Canada**
  `https://library-archives.canada.ca/`
  Used primarily for historical texts, people, archives, and cultural publications.

- **Canadian Museum of History**
  `https://www.historymuseum.ca/`
  Used primarily for social history, material culture, Indigenous cultures, and the historical practices of specific communities.
- **Canada's Historic Places**
  `https://www.historicplaces.ca/`
  A resource maintained jointly by federal, provincial, and territorial authorities, used primarily for historic places, local traditions, and the background of cultural activities.
- **The Canadian Encyclopedia**
  `https://www.thecanadianencyclopedia.ca/`
  An editorially curated encyclopedia used primarily for supplementary verification of cultural terminology, people, history, and social background.