Trans-border Representation Probe: How LLMs Represent the Dai-Thai Community

Minglu Sun
Khoury College of Computer Sciences, Northeastern University

1. Abstract

This study examines how large language models represent trans-border communities whose identities resist nation-state categorization. Focusing on the Dai-Thai community spanning China, Thailand, Myanmar, Laos, and Vietnam, the probe whether LLMs encode "methodological nationalism"—the assumption that nation-states are natural units of cultural analysis.

Using matched-size models (Llama-3.3-70B vs. Qwen-2.5-72B) to eliminate capability confounds, I analyze 44 responses across 11 bilingual prompts. Mixed-methods validation combines manual 5-dimension coding with multilingual embedding analysis.

Core Findings:

1. **Identity ossification**: All models struggle with identity fluidity (M=1.95/3.00)—the weakest dimension universally
2. **Language > Origin**: Embedding analysis shows query language (similarity: 0.649) shapes responses more strongly than model origin (0.509)
3. **Symbolic annihilation**: The same model provides complete trans-border information in Chinese while entirely erasing Southeast Asian distribution in English
4. **Validation**: Significant correlation between embedding and manual coding (r = -0.369, p = 0.002)

2. Introduction & Related Work

2.1 The Problem: Algorithmic Nationalism

Existing AI fairness research predominantly examines biases along gender, race, and political dimensions, implicitly assuming that cultural identity aligns with nation-state boundaries. This assumption systematically excludes communities whose identities are inherently trans-border—what Scott (2009) terms "Zomia" populations: highland peoples across Southeast Asia who have historically resisted state incorporation.

The Dai-Thai community exemplifies this blind spot. Numbering over 20 million across China, Thailand, Myanmar, Laos, and Vietnam, they share linguistic roots, religious practices (Theravada Buddhism), and cultural traditions (e.g., the Water Splashing Festival / Songkran). Yet nation-state frameworks classify them as distinct ethnic groups: "Dai" (傣族) in China, "Thai" in Thailand, "Shan" in Myanmar.

Research Questions:

- RQ1: Do LLMs force fluid trans-border identities into fixed national categories?
- RQ2: How does query language affect trans-border representation?
- RQ3: Does language or model origin more strongly shape representation patterns?

2.2 Related Work

Probing Language Model Representations. This study builds methodologically on CommunityLM (Jiang et al., 2022), which uses prompt-based probing to elicit community-specific worldviews from language models. While CommunityLM probes partisan worldviews (Democrat vs. Republican), I extend this paradigm to **cross-national cultural representation**, examining how models trained in different geopolitical contexts represent communities that transcend those very boundaries.

Cross-lingual Discrepancies in AI Systems. Jiang et al. (2024a) demonstrate systematic discrepancies between parallel English and Chinese content in LLM outputs, focusing on named entity representation. Our study extends this finding to show that such discrepancies affect not only factual content but also **cultural representation of borderland communities**—with one language receiving complete information while another receives erasure.

Embedding-based Analysis. Following the methodological logic of ConGraT (Brannon, Jiang et al., 2024), which uses embedding similarity for community detection in graph-text data, and Time-Aware Document Embeddings (Jiang et al., 2024b), which analyzes factor interactions in embedding space, I employ multilingual sentence embeddings to computationally validate manual coding and discover latent clustering patterns.

Cultural Representation Theory. I operationalize Hall's (1997) concept of **symbolic annihilation**—the systematic absence or misrepresentation of marginalized groups in media—to examine how LLMs may render trans-border identities invisible through omission or forced categorization.

---

3. Methodology

3.1 Experimental Design

I compare Llama-3.3-70B (Meta, US-origin) and Qwen-2.5-72B (Alibaba, China-origin) via OpenRouter API. Both models are open-source, instruction-tuned, and of comparable capability (~70B parameters), enabling isolation of origin-country effects from capability differences.

Following CommunityLM's insight that declarative prompts reduce model hedging, I designed 11 prompts across four categories that force models to take positions on identity fluidity: factual knowledge (e.g., "Where do Dai people primarily live?"), cultural continuity (e.g., "Are the Dai Water Splashing Festival and Thai Songkran the same festival?"), identity classification (e.g., "Can a person be both Dai and Thai?"), and narrative framing (e.g., "Describe the history of the Dai people"). Each prompt was administered in English and Chinese, yielding 44 total responses.

3.2 Coding and Validation

Manual coding assessed five dimensions on a 1-3 scale: trans-border recognition, identity handling, cultural continuity, narrative framing, and factual accuracy. To validate manual coding computationally, I conducted embedding analysis using paraphrase-multilingual-MiniLM-L12-v2, selected for its cross-lingual semantic alignment across 50+ languages.

Full prompt list and coding rubric available at: github.com/ooodddee/Trans-border-Representation-Probe

---

4. Results

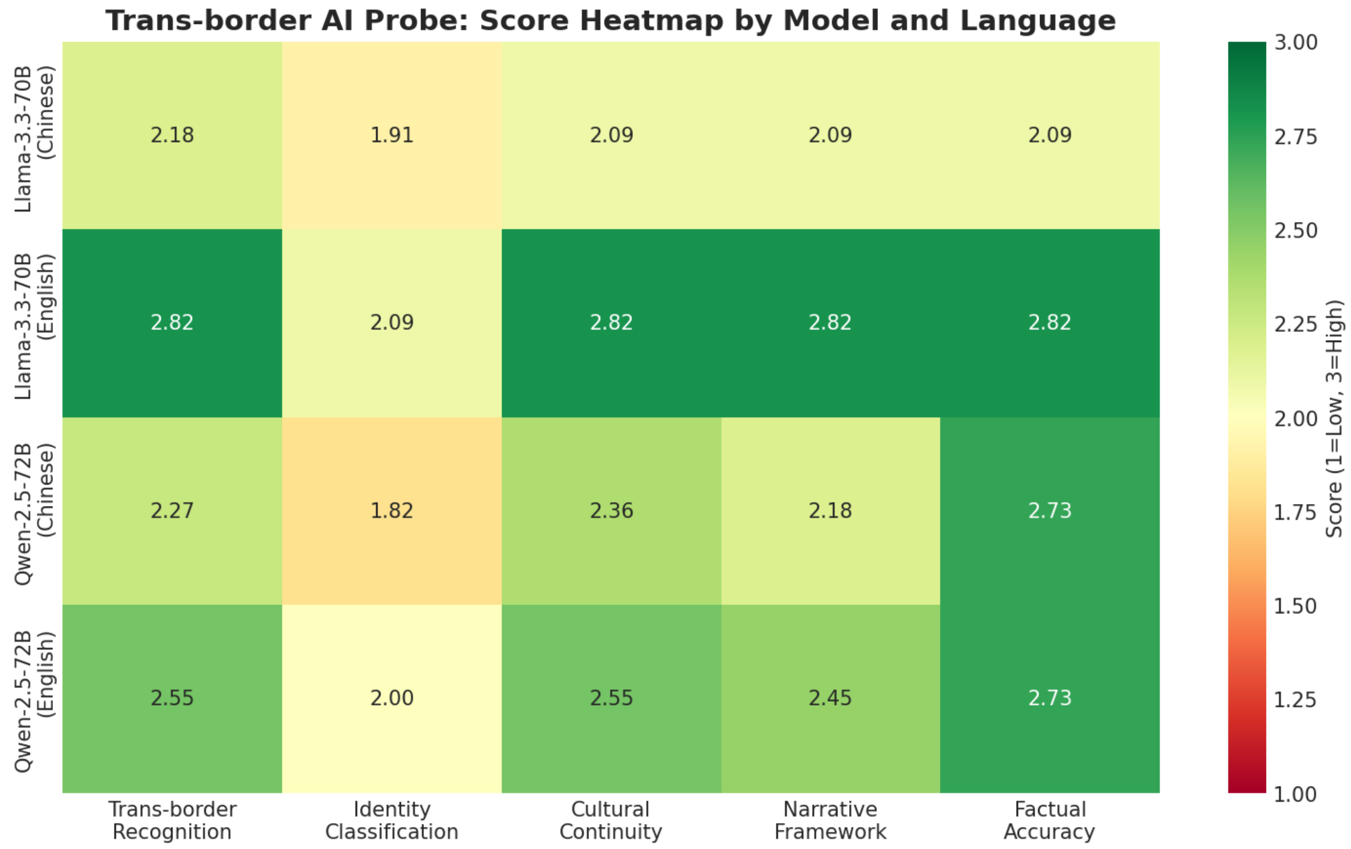
4.1 Quantitative Summary

Table 1: Average Scores by Model and Language

Model	Language	Trans-border	Identity	Cultural	Narrative	Accuracy	Mean
Llama-3.3-70B	English	2.82 (0.40)	2.09 (0.30)	2.82 (0.40)	2.82 (0.40)	2.82 (0.40)	**2.67**
Llama-3.3-70B	Chinese	2.18 (0.75)	1.91 (0.54)	2.09 (0.70)	2.09 (0.70)	2.09 (0.83)	**2.07**
Qwen-2.5-72B	English	2.55 (0.69)	2.00 (0.63)	2.55 (0.69)	2.45 (0.69)	2.73 (0.47)	**2.45**
Qwen-2.5-72B	Chinese	2.27 (1.01)	1.82 (0.75)	2.36 (0.81)	2.18 (0.98)	2.73 (0.47)	**2.27**

Scale: 1 = Poor, 2 = Partial, 3 = Good. Standard deviations in parentheses.

The identity dimension shows universally low scores (range: 1.82–2.09), confirming **identity ossification** as a systematic pattern independent of model origin or query language.



4.2 Finding: Symbolic Annihilation Across Languages

Beyond inconsistency, I identify **symbolic annihilation**—where trans-border presence is entirely erased in one language while fully acknowledged in another.

Case Study: Qwen-2.5-72B on "Where do Dai people primarily live?"

Language	Response	Score
Chinese	"主要聚居在中国云南省...少数分布在**缅甸、老挝、泰国、柬埔寨、越南**等东南亚国家"	3 (Complete)
English	"primarily live in the southwestern part of China, mainly in Yunnan Province... one of the 56 officially recognized ethnic groups in China"	1 (Erased)

The same model provides complete trans-border information in Chinese but **entirely omits Southeast Asian distribution in English**. This constitutes symbolic annihilation through omission, rendering the transnational community invisible to English-language users.

4.3 Finding: Language Dominates Over Model Origin

Embedding analysis reveals that **query language clusters responses more strongly than model origin**.

Table 2: Embedding Similarity Analysis

Comparison Type	Cosine Similarity
Same Language, Different Model	**0.649**
Same Model, Different Language	0.509

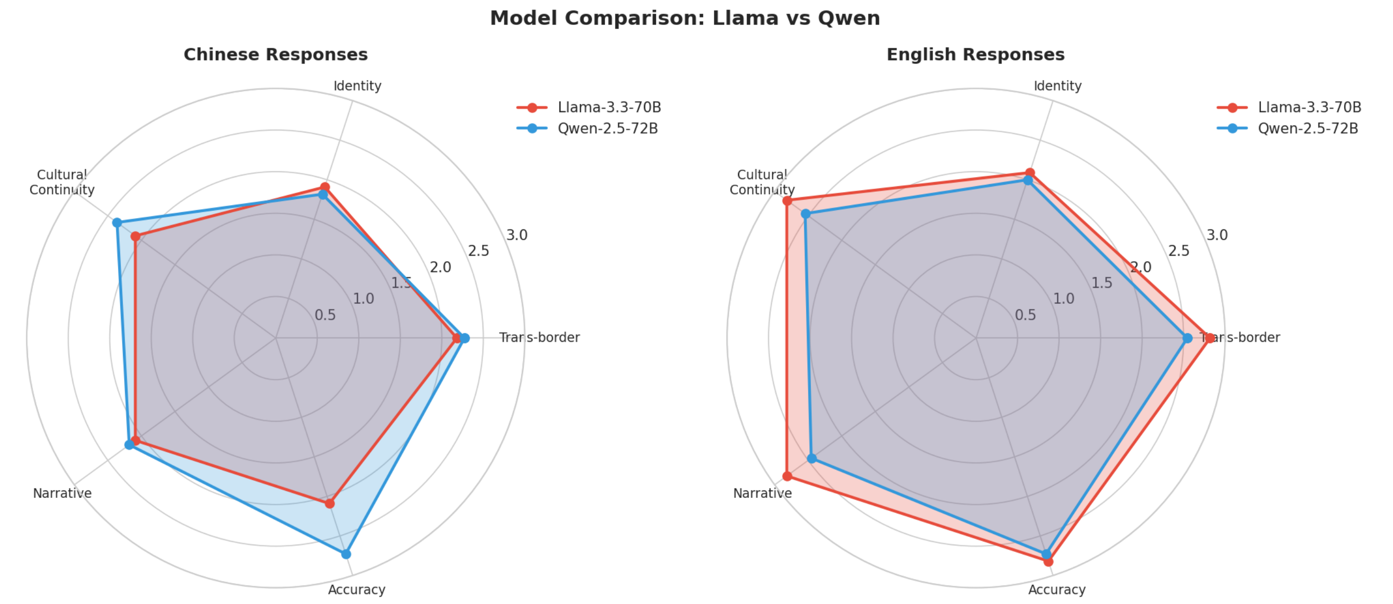
Responses in the same language (regardless of model) are significantly more similar than responses from the same model across languages. This finding suggests that query language fundamentally restructures how LLMs represent trans-border communities, independent of the model's geopolitical origin.

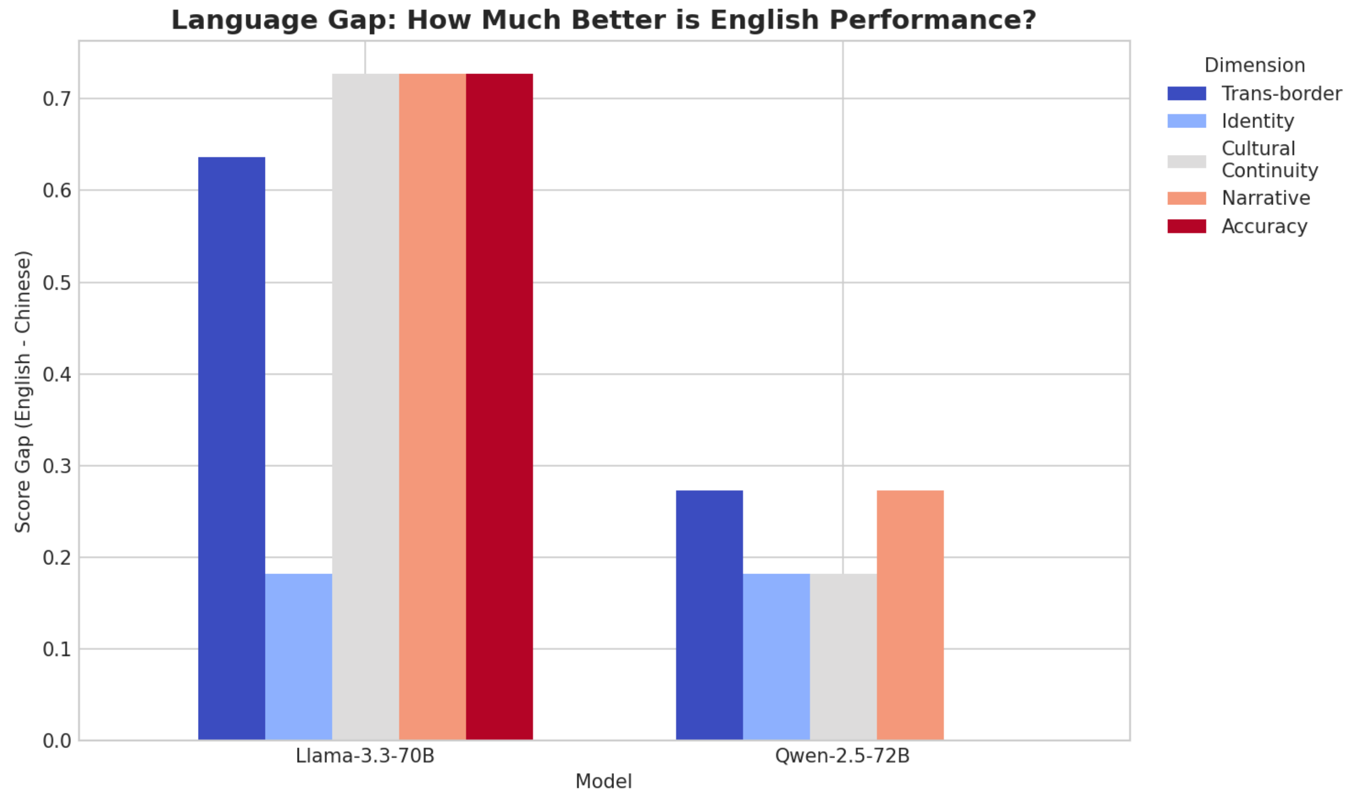
Validation. The significant negative correlation between embedding similarity and manual score differences (r = -0.369, p = 0.002) confirms that computational and human coding approaches capture related aspects of representation quality, providing triangulated validation.

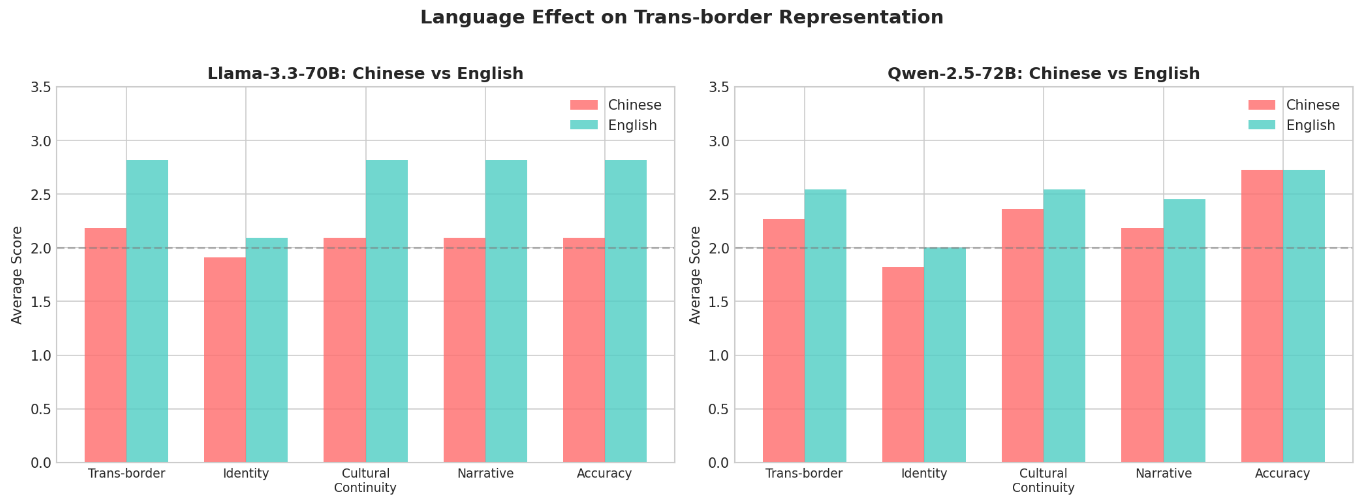
5. Discussion & Conclusion

5.1 Implications

The findings reveal a systematic blind spot in LLM representation: **identity ossification**. When queried about communities with fluid, trans-border identities, models default to nation-state frameworks, forcing classification into single national categories. This pattern persists across model origins (US vs. China) and represents a form of algorithmic nationalism embedded in training data organization.


Model Comparison: Llama vs Qwen
Chinese Responses
English Responses
Llama-3.3-70B
Qwen-2.5-72B
Identity
Cultural Continuity
Narrative
Accuracy
Trans-border
0.5
1.0
1.5
2.0
2.5
3.0


Language Gap: How Much Better is English Performance?
Score Gap (English - Chinese)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Llama-3.3-70B
Qwen-2.5-72B
Model
Dimension
Trans-border
Identity
Cultural Continuity
Narrative
Accuracy

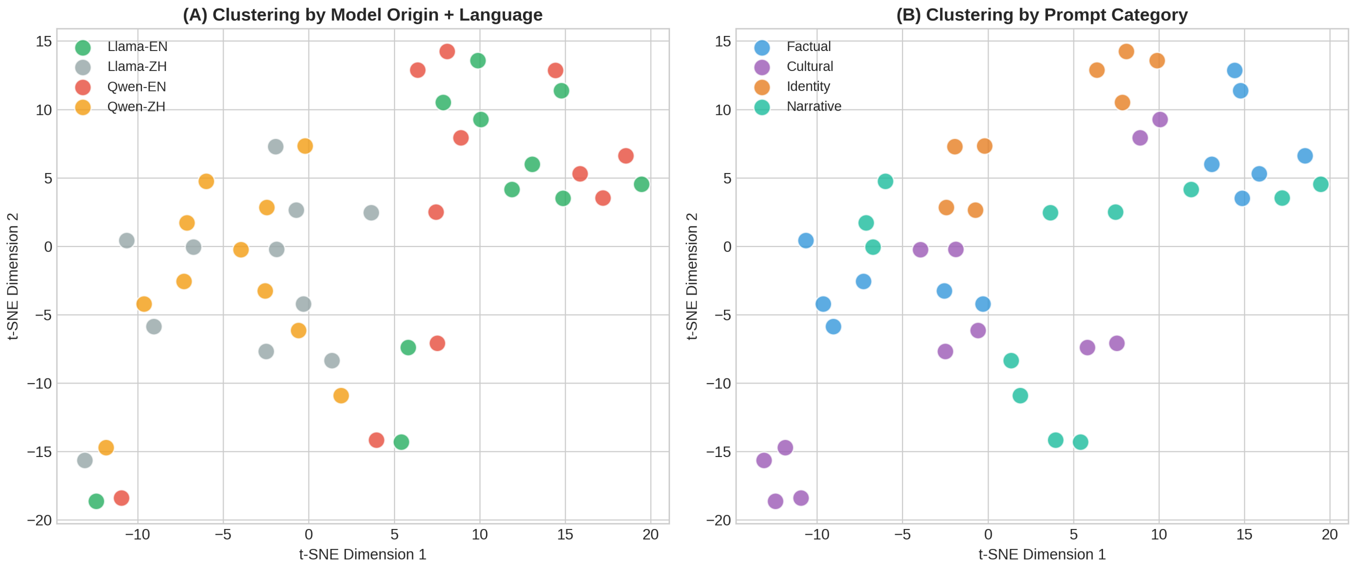

Language Effect on Trans-border Representation
Llama-3.3-70B: Chinese vs English
Qwen-2.5-72B: Chinese vs English
Average Score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Chinese
English
Trans-border
Identity
Cultural Continuity
Narrative
Accuracy

These visualizations demonstrate the stark contrast: Llama-3.3-70B exhibits a substantial gap between its English performance (M=2.67) and Chinese performance (M=2.07), suggesting that English fine-tuning optimizations do not transfer to Chinese. In contrast, Qwen-2.5-72B maintains more consistency across languages (English M=2.45, Chinese M=2.27), reflecting its native-language optimization for Chinese. This interaction pattern—where model origin and query language jointly shape output quality—highlights that geopolitical design decisions embedded during training are not neutral but fundamentally alter what information different user populations receive.

The radar chart above visualizes this pattern clearly: across all four model-language combinations (Llama English, Llama Chinese, Qwen English, Qwen Chinese), the "Identity" dimension consistently forms the innermost arc—systematically narrower than all other dimensions. This geometric pattern provides direct visual evidence that identity handling is not merely weaker in some contexts, but represents a universal structural limitation in how LLMs process trans-border identity.

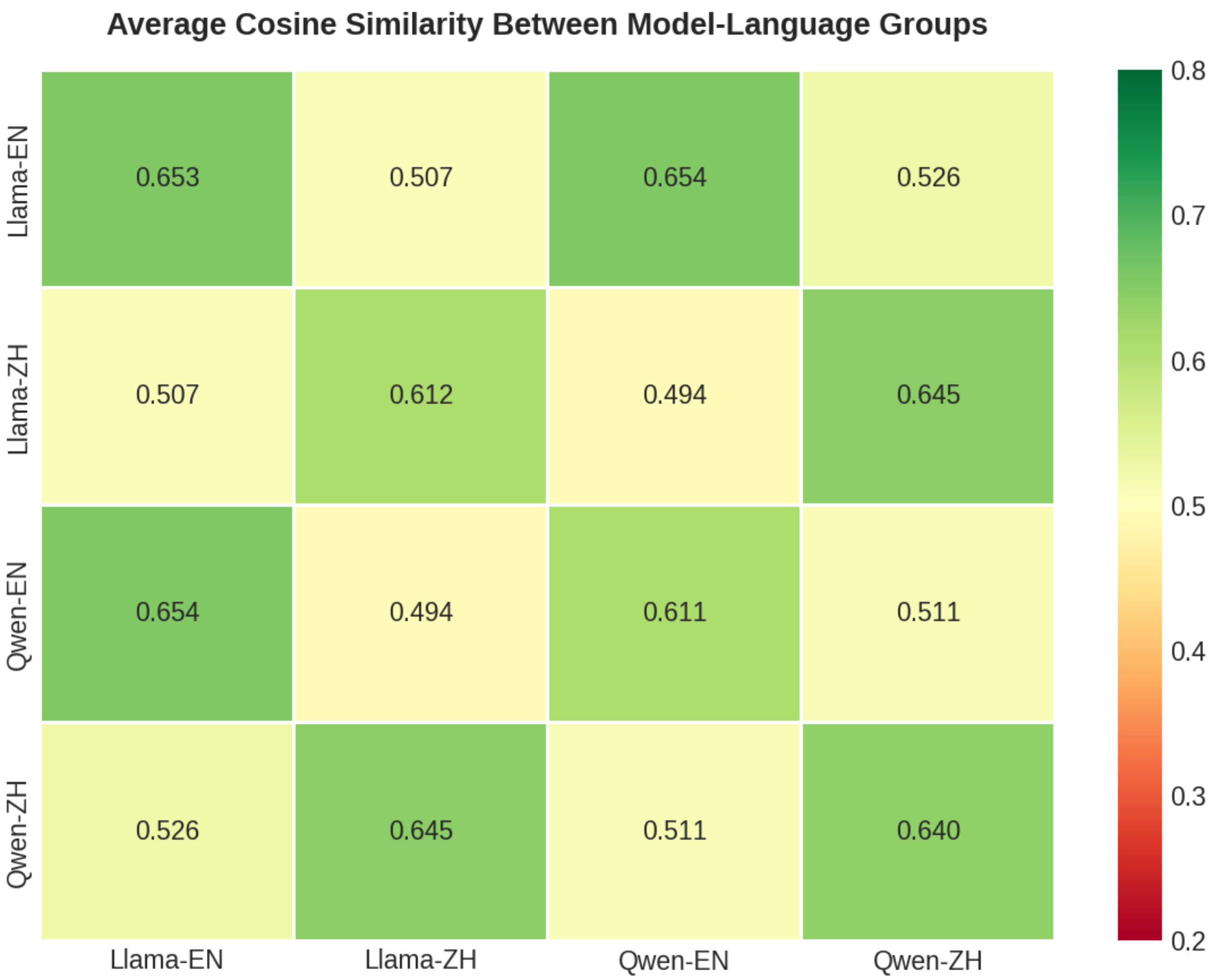
The symbolic annihilation finding is particularly concerning: users querying the same model in different languages receive fundamentally different representations of reality. English-language users are presented with a China-bounded view of the Dai community that erases its transnational distribution—information readily available in the model's Chinese responses.

Embedding-Based Clustering: Language Dominates Over Model Origin.** However, the most striking computational finding emerges from embedding-space analysis: when responses from all four model-language combinations are embedded and visualized via t-SNE dimensionality reduction, they cluster primarily by language rather than by model origin.



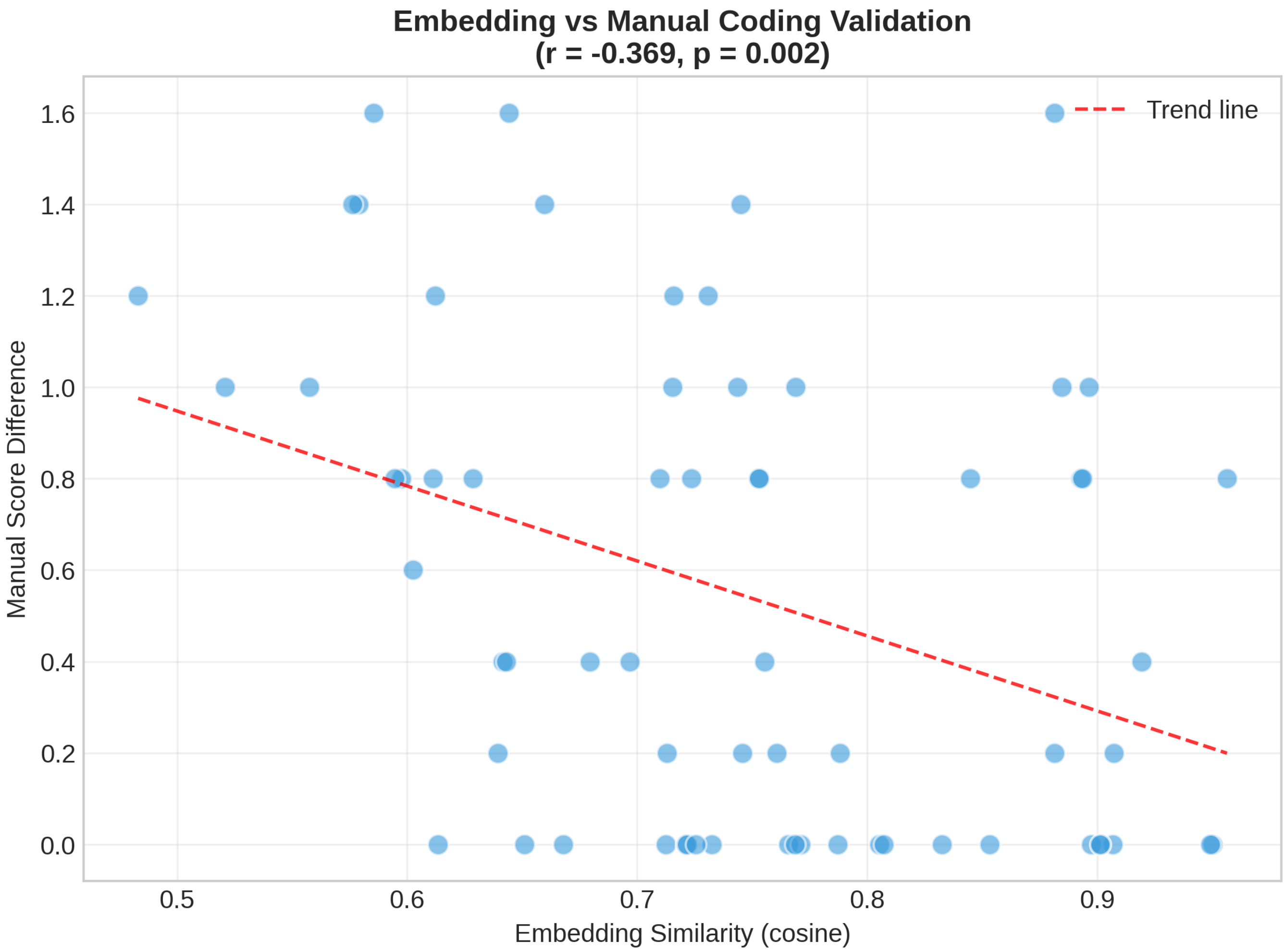
The t-SNE plot reveals clear color-based clustering (English responses cluster together regardless of model; Chinese responses cluster together regardless of model), with no coherent Llama vs. Qwen clustering. This spatial arrangement provides direct evidence that query language acts as a dominant structural force reshaping the representation space itself—more so than the model's geopolitical training context.

To quantify this observation, I computed cosine similarity matrices for response embeddings:



The heatmap demonstrates numerically what the t-SNE plot shows visually: same-language pairs (regardless of model) achieve mean cosine similarity of **0.649**, while same-model pairs across languages achieve only **0.509**. This 27% increase in within-language similarity confirms that language fundamentally restructures the representation space more powerfully than model origin.

Automatic Validation via Embedding-Manual Coding Correlation. To validate that embedding-based findings capture meaningful variation rather than surface-level linguistic similarity, I computed Pearson correlation between cosine similarity measures and differences in manual coding scores:



The correlation of r = -0.369 (p = 0.002) is statistically significant and of moderate effect size, indicating that responses more similar in embedding space tend to receive more similar manual coding scores. This convergence between computational and human judgment provides crucial triangulated validation: the embedding model is not simply flagging linguistic surface similarity, but capturing semantically meaningful variation that human coders independently identify.

5.2 Limitations

This study has several limitations. First, the sample size (n=44) limits statistical power. Second, single-coder annotation introduces potential bias; future work should establish inter-rater reliability. Third, the embedding model used for validation may itself carry biases that inflate same-language similarity. Finally, and most critically, **the coding schema reflects academic frameworks rather than community-defined criteria** I have not yet validated whether these dimensions capture what affected communities consider adequate representation.

5.3 Future Work

Model expansion. I plan to extend this analysis to GPT-4o, Claude 3.5, and additional Chinese-origin models (Baichuan, Yi) to test generalizability of findings.

Community validation. A critical next phase involves participatory workshops in Xishuangbanna and Dehong, Yunnan, recruiting Dai community members to evaluate AI outputs and develop community-defined harm taxonomies. This addresses the fundamental limitation that current AI auditing is researcher-defined rather than community-centered.

Global extension. The Trans-border Representation Probe framework can be applied to other cases of algorithmic nationalism: Kurdish communities across Turkey/Syria/Iraq/Iran, Sámi peoples across Nordic countries, and Rohingya across Myanmar/Bangladesh.

Code & Data: github.com/ooodddee/Trans-border-Representation-Probe