Towards Multilingual Mechanistic Interpretability

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multilingual language models achieve strong averages yet often behave unpredictably across languages, scripts, and cultures. We argue that mechanistic explanations for such models should satisfy a *causal* standard: claims must survive causal interventions and must *cross-reference* across environments that perturb surface form while preserving meaning. We formalize *reference families* as predicate-preserving variants, and we introduce *triangulation*, an acceptance rule requiring (i) invariance of the conditional law of a task score given the internal states of a proposed subgraph and (ii) directional stability and sufficient magnitude of interventional effects across references. To supply candidate subgraphs, we adopt *automatic circuit discovery* (edge attribution patching, position-aware circuit discovery, and sparse subgraph selection), and we *accept or reject* those candidates by triangulation. Our proposal situates mechanistic interpretability within the theory of causal abstraction and complements causal mediation analyses by focusing on falsifiable cross-environment invariance.

1 Introduction

The success of multilingual language models (MLLMs) has disguised a persistent pattern: large average gains mask instability across languages, writing systems, and cultures. When one isolates language- or culture-specific subsets, model rankings can invert; when inputs mix languages within a sentence, models often leak or rely on brittle shortcuts. These observations suggest that many analyses of internal states are, at best, associational: they reveal where information is encoded but not whether it *causes* behavior.

We take a simple position. A mechanistic explanation should be accepted only if it remains valid under interventions and *cross-references* across environments that keep meaning fixed while perturbing nuisance attributes such as language and script. Multilinguality offers exactly these environments. The literature on invariant causal prediction establishes how stability across environments can reveal causal parents, while cross-referenceability clarifies when effects move between populations that differ in specified ways. Mechanistic interpretability provides the tools for local interventions on internal states. What has been missing is a standard that integrates these ingredients into an *acceptance rule* for mechanism claims.

We propose such a standard and call it *triangulation*. Triangulation evaluates a proposed subgraph in two complementary ways. First, it demands that the conditional law of a task score given the subgraph's internal states be invariant across predicate-preserving references. Second, it requires that causal interventions on those states—replacing them with activations drawn from the references—push the score in directions that are consistent across references and large enough to rule out chance. In practice, this rule filters out mechanisms that owe their apparent success to language identity, script, register, or other surface cues.

2 Related Work

Mechanistic interpretability for LLMs. Interventional tools such as causal tracing and path/edge patching underpin LLM mechanistic interpretability, but outcomes are highly sensitive to corruption choices and metrics, motivating on-manifold constraints and stricter protocols. Recent evidence further shows that circuit "faithfulness" scores can be brittle to seemingly minor ablation details, reinforcing the need for acceptance criteria that go beyond single-environment patch scores [Miller et al., 2024].

Automatic circuit discovery. We use "automatic circuit discovery" broadly for pipelines that algorithmically produce candidate subgraphs with minimal manual curation. Search-based methods (e.g., ACDC) directly return sparse circuits that preserve behavior on held-out inputs [Conmy et al., 2023]; position-aware variants add token-span sensitivity and an automated schema, improving the size–faithfulness trade-off [Haklay et al., 2025]. Edge-scoring methods (e.g., EAP/EAP-IG) automatically rank edges; coupled with an automatic selection rule (thresholding/pruning or seeding a search on a pre-pruned graph), they also yield circuits. In our pipeline, automatic discovery proposes subgraphs, and triangulation determines acceptance.

Causal mediation and falsification tests. Causal mediation decomposes total effects into natural indirect/direct components through nominated mediators (e.g., attention heads) and has been applied to transformers [Vig et al., 2020]. Causal scrubbing offers behavior-preserving resampling tests to falsify mechanistic hypotheses [Chan et al., 2022]. Our approach shares the falsification ethos but avoids cross-world assumptions by requiring *invariance* of the predictive link and *directional stability* of interventional effects across predicate-preserving references.

Multimodal mechanistic interpretability. For vision–language models, NOTICE introduces a corruption/intervention pipeline for text–image pairs to probe attention-level roles [Golovanevsky et al., 2025]. Tools such as LVLM-Interpret emphasize interactive analysis rather than controlled interventions [Ben Melech Stan et al., 2024]. Explicit cross-environment tests (e.g., language/script flips that preserve the predicate) remain rare in multimodal MI; our triangulation standard fills this gap.

Invariance and causal abstraction. Invariant Causal Prediction formalizes why *causal* parents support stable conditional behavior across environments [Peters et al., 2016]. Causal Abstraction gives a principled account of when low-level interventions should commute with high-level changes, yielding graded faithfulness between circuits and interpretable models [Geiger et al., 2025]. Our acceptance rule operationalizes these principles: only circuits whose predictive link is invariant and whose interventional effects are directionally stable across reference families are accepted as mechanisms.

3 Structural Causal Model

We summarize a forward pass by five endogenous variables $\{\mathcal{R}\}$, C, X, H, M. The symbol $\{\mathcal{R}\}$ denotes a set of reference families; for a base input x we choose a particular family $\mathcal{R}(x) = \{r_1, \dots, r_K\} \subset \{\mathcal{R}\}$. The variable C denotes nuisance attributes influenced by these references. The observed text is X. The internal states are $H = (H_1, \dots, H_J)$ at a specified patch site (e.g. attention head, MLP); the task score (e.g., a logit margin) is M.

Using the language of structural causal models:

$$\begin{aligned} &C = g_C\big(\{\mathcal{R}\}\big), \qquad X = g_X(C), \\ &H_j = f_\ell\big(H_{<j}, X\big) \quad (\ell = 1, \dots, J), \qquad M = f_M(H, X). \end{aligned} \tag{1}$$

We assume that $\{\mathcal{R}\}$ influences M only through (C, X) and that the predicate of interest resides in X rather than in superficial aspects of C. Under this description, a reference family toggles C while keeping the predicate in X intact. Figure 1 shows the corresponding causal graph.

The mathematical role of a reference family is to vary nuisances while preserving the predicate. For a input x and its family $\mathcal{R}(x) = \{r_1, \dots, r_K\}$, we model predicate preservation as a tolerance on

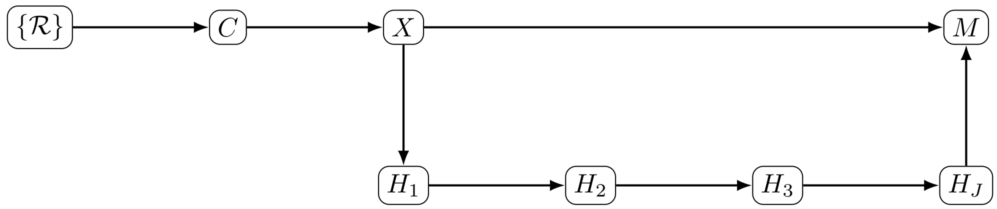

Figure 1: A causal DAG for multilingual mechanisms. The set of reference families $\{\mathcal{R}\}$ determines nuisances C, which shape X. The input X propagates through internal states $H_1, \ldots, H_J$ to M, with a direct dependence of M on X.

the task score,

$$|M(r_k) - M(x)| \leq \varepsilon \quad \text{for all } k \in \{1, \ldots, K\}. \tag{2}$$

A mechanistic hypothesis specifies a set $S \subseteq \{1, \ldots, J\}$ of components (e.g., a handful of heads and MLPs tied together by a plausible path from X to the first decoding step that bears the relevant attribute). Let $a_j(\cdot)$ denote activations of j at a fixed patch site. The causal intervention is to replace an endogenous state with one drawn from a reference, i.e. $\mathrm{do}(H_j = a_j(r_k))$. For a deterministic forward pass, the corresponding change in the score is

$$\Delta M_j^{(k)} = M\Big(x \,\Big|\, \mathrm{do}(H_j = a_j(r_k))\Big) - M(x). \tag{3}$$

4 Triangulation

Triangulation asks for invariance in prediction and for directional stability of interventional effects. We first specify a target behavior M and a reference family $\mathcal{R}(x)$, which casts the task as: identify a sparse subgraph that mediates M and remains valid across the references in $\mathcal{R}(x)$. Candidate subgraphs are proposed by any principled localization or search procedure that scores or isolates routes in the computation graph, optionally with position sensitivity when examples vary in length. From these proposals, a compact subgraph is selected by a sparsity-oriented criterion that preserves M on held-out inputs. This yields candidates for S without hand-crafting; triangulation then accepts only those whose interventional effects are directionally stable and whose predictive link is invariant across the references in $\mathcal{R}(x)$.

The first requirement is an internal version of invariant causal prediction. If S captures the variables on which the mechanism for M depends at the patch site, then the conditional distribution of M given a_S should not change across the references that preserve the predicate,

$$F\left(M(r) \,|\, a_S(r)\right) \;\; \text{is identical for all} \;\; r \in \mathcal{R}(x). \tag{4}$$

In practice, one could fit a single predictor $\widehat{g}$ from a_S to M pooled across references and then verifies that residual distributions are stable.

The second requirement concerns interventions. For each component $j \in S$, denote the effect vector $\boldsymbol{e}_j = (\Delta M_j^{(1)}, \ldots, \Delta M_j^{(K)})$ from (3). Beforehand we preregister a vector $\boldsymbol{c} \in \{-1, +1\}^K$ that codifies the *direction* in which the score should move for each member of the reference family. We accept the mechanism only if

$$\forall k\colon \operatorname{sgn}\big(\Delta M_j^{(k)}\big) = c_k, \qquad \|\boldsymbol{e}_j\|_2 \geq \tau, \qquad \frac{\boldsymbol{e}_j \cdot \boldsymbol{c}}{\|\boldsymbol{e}_j\| \, \|\boldsymbol{c}\|} \geq \gamma, \tag{5}$$

for thresholds $\tau > 0$ and $\gamma \in (0, 1]$, subject to an on-manifold constraint $\|a_j(r_k) - a_j(x)\| \leq \delta$. In other words, triangulation is an *acceptance* rule, not a discovery method. To propose candidates S at scale, we can use automatic circuit discovery (cf. [Conmy et al., 2023, Hanna et al., 2024]) and then filter via triangulation.

5 Proposed Case Study: Inclusive English→French Translation

For a conceptual example, consider English-to-French translation in settings where French admits both binary and inclusive realizations. The goal is to evaluate whether a localized internal mechanism S governing the first gender-bearing decision remains valid across a small, predicate-preserving reference family $\mathcal{R}(x)$ for each base sentence x.

Data and reference construction. For each base item x, we build $\mathcal{R}(x) = \{r_{\text{he}}, r_{\text{she}}, r_{\text{they}}\}$ that toggles source-side ambiguity and target-side realization while preserving denotation. Concretely: (i) minimally paraphrase the English source to flip pronominal ambiguity and scrub overt gender cues while keeping named entities and semantics fixed; (ii) prepare parallel French targets that differ *only* in the gender-marked span (inclusive option(s) vs. binary alternatives), using controlled templates and editor guidelines; (iii) balance occupational and contextual stereotypes around the referent (neutral vs. stereotyped contexts).

Score and locus. Let the *gender-bearing locus* (GBL) be the earliest decoding position where the French realization commits to gender marking. Using teacher-forced decoding and alignment, we identify the GBL and define the task score

$$M \;=\; \log p_\theta(\text{incl} \mid \text{ctx at GBL}) \;-\; \max\big\{\log p_\theta(\text{masc} \mid \cdot),\; \log p_\theta(\text{fem} \mid \cdot)\big\},$$

so that larger M favors an inclusive realization at the GBL.

Candidate localization. For each x, we localize routes from the source cue span to the GBL and generate candidate subgraphs S with a fully automatic pipeline: a principled route/edge localization step (scoring or intervention-based), optional position sensitivity to handle variable pronoun placement, and a sparsity-oriented selection that preserves M on held-out items. Discovery is thus separated from acceptance; no manual head-picking is performed.

Triangulation tests (accept/reject). For each candidate S, we apply two checks without re-estimating any new quantities. (i) *Internal invariance:* using the pooled predictor $\widehat{h}(a_S) \to M$ defined above, we test residual stability across $r \in \mathcal{R}(x)$ exactly as in Eq. (4). Only candidates passing invariance proceed. (ii) *Interventional consistency:* we perform the patches on the components of S and summarize the resulting score changes via the effect vectors defined around Eq. (3); acceptance requires the directional and magnitude criteria of Eq. (5) under the same on-manifold constraint introduced earlier. Thresholds are calibrated with placebo patches, and we report permutation p-values per candidate.

Analyses and diagnostics. We report: (a) pass/fail rates under invariance vs. under directional alignment; (b) effect-size distributions for accepted vs. rejected candidates; (c) robustness by context (neutral vs. stereotyped) and by realization type (inclusive pronoun vs. paraphrase vs. orthographic agreement). Two contrasts are diagnostic: circuits genuinely mediating the cue at the GBL should show stable residuals and consistently signed ΔM across $\mathcal{R}(x)$; nuisance-sensitive routes (e.g., reacting to punctuation or script) should produce near-zero or sign-inconsistent effects and be rejected.

Error taxonomy and ethics. We separate (1) lexicalization failures (no inclusive candidate is available or scored competitively), (2) agreement failures (inclusive cue with binary downstream agreement), and (3) cue misrouting (changes in pronoun position flip the sign pattern). Because inclusive realization remains socially and institutionally contested, we treat it as an *optional, user-controlled* target; evaluation is framed strictly by predicate preservation and cross-reference stability rather than by prescriptive preference.

6 Limitations and scope

Triangulation raises the evidential bar but does not guarantee uniqueness. Several distinct subgraphs may satisfy the acceptance rule when their effects are redundant or when the model implements multiple pathways for the same predicate. The quality of reference families is decisive: if the supposed predicate-preserving rewrites actually change what is being asked, invariance and consistency may mislead. Finally, causal interventions at scale are computationally heavy; prioritization strategies and sampling are advisable.

References

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. LVLM-Intrepret: An Interpretability Tool for Large Vision-Language Models. pages 8182–8187, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/XAI4CV/html/Stan_LVLM-Intrepret_An_Interpretability_Tool_for_Large_Vision-Language_Models_CVPRW_2024_paper.html.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Nitishinskaya Jenny, Ansh Radhakrishnan, Shlegeris Buck, and Nate Thomas. Causal Scrubbing: a method for rigorously testing interpretability hypotheses [Redwood Research]. December 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. ISSN 1533-7928. URL http://jmlr.org/papers/v26/23-0058.html.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Tal Haklay, Hadas Orgad, David Bau, Aaron Mueller, and Yonatan Belinkov. Position-aware Automatic Circuit Discovery. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2792–2817, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.141. URL https://aclanthology.org/2025.acl-long.141/.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. August 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT#discussion.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer Circuit Evaluation Metrics Are Not Robust. August 2024. URL https://openreview.net/forum?id=zSf8PJyQb2&utm_source=chatgpt.com.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, November 2016. ISSN 1369-7412. doi: 10.1111/rssb.12167. URL https://doi.org/10.1111/rssb.12167.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.