Theoretical Linguistics Constrains Hypothesis-Driven Causal Abstraction in Mechanistic Interpretability

Suchir Salhan
[sas245@cam.ac.uk]

Konstantinos Voudouris
[k.voudouris@helmholtz-munich.de]

Department of Computer Science & Technology, University of Cambridge, Cambridge, U.K.
Helmholtz Institute for Human-Centered AI, Helmholtz Munich, Munich, Germany

Abstract

Mechanistic Interpretability aims to uncover the causal processes that explain language model behaviour by identifying internal representations and circuits. Yet the search space of possible alignment maps—functions linking hidden representations to hypothesized causal variables—remains essentially unconstrained. We show how this lack of structure limits reproducibility, generalisation, and the *explanatory adequacy* of causal claims. In this position paper, we argue that linguistic theory provides theory-driven templates for explicit, falsifiable constraints on candidate alignment maps to move from ad hoc circuit discovery toward a principled, hypothesis-driven science of causal abstraction. We offer case studies to highlight possible restrictions provided by linguistics on alignment maps, focusing on phonology and morphology and the contrastive properties of phoneme-level models, text-trained language models, and multimodal speech-text models. By framing interpretability as a hypothesis-driven, linguistically grounded science, we move closer to a program where mechanistic claims can be cumulative, comparative, and testable across models.

1 Introduction

Cognitive science and mechanistic interpretability face analogous challenges in decomposing complex systems into meaningful components such that behaviour can be explained in terms of causal interactions. Founding doctrines of cognitive science emphasize that helpful decompositions are guided by computational and representational motifs [7, 19, 23]. In the sub-field of machine learning known as mechanistic interpretability (MI), similar principles are emerging through the lens of causal abstraction. MI begins with the postulation of a latent causal process Z that explains observed model behaviour Y. Researchers then attempt to construct an alignment map ϕ, linking hidden representations to hypothesised causal variables. In the limiting case, the space of candidate maps is unrestricted: every well-fitting ϕ is admissible, which, for finite data, is an infinitely large set [3, 9, 10, 16, 20, 22]. Without principled constraints, interpretability risks devolving into a combinatorial search over arbitrary mappings, with little guidance for which structures are scientifically meaningful [15, 24]. This problem is acute in language models, where polysemantic neurons, distributed subspaces, and entangled circuits complicate any claim about linguistic competence. Existing interpretability work has identified agreement circuits, induction heads, and compositional subspaces, but these discoveries remain piecemeal and difficult to generalize. What is missing is a principled framework that can delimit the space of candidate causal maps.

We argue that linguistic theory offers substantive generalisations about the structures that human languages exhibit – from phonological processes (e.g., phonological assimilation) to morphosyntactic

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models.

rules and constraints [8, 12, 13]. These are symbolic generalisations that hold descriptively, independent of the ontogenetic debates about where they come from (i.e., innate vs. acquired). Whether these symbolic generalisations are innately specified or emergent from experience, they define a finite hypothesis space of plausible causal maps. Symbolic computational analyses of cognitive systems that **model rule-based behaviours at multiple structural levels** naturally provide explicit constraints on the kinds of mappings that are cognitively and computationally plausible. The descriptive results of theoretical linguistics can be directly imported into causal abstraction analyses to move from unconstrained "circuit-hunting" toward a falsifiable, hypothesis-driven science of causal abstraction in Deep Neural Networks.

2 Postulating Causal Sources in Mechanistic Interpretability

In Mechanistic Interpretability (MI), a neural network $f : X \to Y$ with hidden states $h \in \mathcal{H}$ can be understood as implementing an abstract causal system. This system maps causal sources Z to outcomes W via a causal model $A : Z \to W$. An **alignment map** $\phi : \mathcal{H} \to Z$ links hidden representations to these causal sources. Interventions and counterfactual analyses on ϕ allow predictions of network behaviour under hypothetical manipulations.

Formally, a causal abstraction exists if, for a network f, there exists an alignment map ϕ and a causal model A such that, for any intervention $I(X)$, the relation $A(\phi(I(X)))$ commutes [9]. However, unconstrained non-linear alignment maps can always produce a correspondence between network activations and arbitrary causal models, rendering causal abstraction vacuous. Sutter et al. (2025) highlight that if f is a universal function approximator, then under standard assumptions any network can be aligned with any algorithm through a sufficiently complex alignment map [24]. This presents the **non-linear representation dilemma**: restricting maps risks missing interpretable ϕ's, while allowing arbitrary maps loses interpretability.

3 The Linguistic Emergentism Hypothesis (LEH) for Hypothesis-Driven Causal Abstraction

We argue that, in light of the non-linear representation dilemma, a principled resolution requires external constraints on which alignment maps are admissible. We propose that **linguistically motivated alignment maps** resolve this dilemma. These maps are restricted to subspaces consistent with descriptive phonological, morphological, and syntactic rules, turning causal abstraction into a falsifiable framework: only alignments compatible with known linguistic templates are considered plausible. Linguistic theory provides explicit constraints on candidate alignments, limiting the combinatorial explosion of admissible mappings. Templates define rule systems, representational structures, and symbolic dependencies guiding ϕ. In phonology, mappings are constrained by frame-level features and rule-based interactions such as feature spreading, assimilation, and opacity. Morphological mappings reflect stem-affix relations and productive finite-state rules, distinguishing memorization from causal implementation. In syntax, mappings are restricted to hierarchical dependencies, such as agreement, binding, and long-distance movement. Restricting mappings to descriptive rule systems makes causal abstraction falsifiable: a proposed rule is either implemented (interventions preserve effects) or rejected. More formally, given a set of domain-specific symbolic rules from linguistic theory $\mathcal{Z}$, an abstract causal system A, and an intervention mapping ϕ^{-1}, an alignment map $\phi : \mathcal{H} \to \mathcal{Z}$ is **linguistically-constrained** if it satisfies

$$f(h \leftarrow \phi^{-1}(I(r))) \approx A(I(r)).$$

This ensures that only mappings consistent with known symbolic rules are admissible; interventions outside this constraint are not mechanistically informative. This defines a falsifiable, mechanistically grounded hypothesis: models implement symbolic generalizations not merely when their outputs align with surface statistics, but when their internal causal structure corresponds to linguistically defined rules. Success is measured both in alignment with linguistic labels and by testing mechanistic necessity (altering a variable disrupts predicted outcomes) and sufficiency (setting it to a target value produces the expected behavior).

Our contribution is to restrict both ϕ and bijective translations τ using these symbolic rules, defining interpretable subspaces for causal variables, limiting geometry and locality of hidden representations, and preventing trivial mappings. We propose that it is possible to also quantify the distribution of a causal variable using a *localization complexity metric* $C(\phi)$, detailed in Appendix B, which measures the number of hidden subspaces needed to implement a causal effect. Lower $C(\phi)$ is consistent with sparse, modular, symbolic representations. Existing interpretability methods provide tools to operationalize this framework to evalute *symbolic emergentism* in Deep Neural Networks: circuit localization, attribution patching, and information-flow analysis identify structural circuits implementing rules, and distributed alignment search (DAS) evaluates causal variable localization. CausalGym [3] exemplifies linguistically-constrained causal abstraction by localizing agreement neurons and distributed circuits mediating long-distance syntactic dependencies. Our proposal generalises this to other linguistic domains (phonology/morphology), and multimodal models, allowing principled, theory-driven exploration of mechanistic structure. This defines a falsifiable, mechanistically grounded hypothesis: models implement symbolic generalizations not merely when their outputs align with surface statistics, but when their internal causal structure corresponds to linguistically defined rules. Unlike classic semantic interpretability, which maps latent vectors to human-interpretable features without falsifiable criteria, our approach evaluates both alignment with linguistic labels and mechanistic necessity and sufficiency. We next illustrate a phonological case study to demonstrate how integrating linguistic constraints with formal causal abstraction yields a hypothesis-driven, falsifiable framework for MI across models and modalities.

3.1 Case Study: Nasal Assimilation and Opacity in Turkish

Turkish provides a clear example of how causal templates can generate testable predictions about language patterns. In this language, a suffixal nasal sound changes depending on the following consonant: it appears as an alveolar before alveolars and as a velar before velars. We can probe whether a model captures these patterns by creating pairs of words that differ only in the relevant context and then testing what happens when we selectively "turn off" or "boost" the relevant feature representations in the model. Similarly, vowel harmony in Turkish can be hidden by another process called final devoicing, creating a situation where the underlying rule is not directly visible on the surface. By performing controlled interventions on the model's internal representations, we can see whether the model truly encodes these hidden rules. Only when both "lesioning" and "stimulation" interventions behave as expected can we conclude that the model's internal structure genuinely reflects the linguistic rules. *Table* 1 illustrates how we can define structural constraints on an alignment map ϕ for Turkish nasal assimilation and vowel harmony opacity.

Table 1: The table shows the surface forms produced by the model (Output), the type of intervention applied to the model's internal representation (Intervention), and the predicted effects of that intervention (Prediction / Counterfactual). "Lesion" means the relevant feature is blocked, and "Stimulate" means it is artificially enhanced.

Process	Input	Output	Intervention	Prediction / Counterfactual
Nasal assimilation	/sen+de/, /sen+ge/	[sende], [seŋge]	Lesion	[senge] (assimilation lost) / No effect
			Stimulate	[seŋde] (over-applied) / No effect
Opacity (harmony)	/kitap+da/, /kitap+ta/	[kitapta], [kitapta]	Lesion	[kitapde] (harmony disrupted) / [kitapta]
			Stimulate	Hidden [+back] (despite devoicing) / No hidden state diff.

4 Architecture and Modality Effects on LEH

By framing these questions in terms of causal abstraction under emergence, we move beyond asking merely whether linguistic rules can be mapped onto model representations, and instead focus on how and why these rules arise within the network. In this context, models themselves function as experimental **model organisms** [16, 20]: different architectures and training regimes serve as distinct "environments" in which we can study the emergence of structured, rule-based behaviour. Self-supervised speech models (S3Ms) are particularly valuable as model organisms because they encode multiple layers of linguistic information—acoustic, phonological, morphological, syntactic, and semantic—simultaneously [17, 21]. Omni-models—self-supervised speech-and-text models that encode multi-level linguistic information—provide the most naturalistic setting for this investigation. Unlike text-only LMs, these models integrate acoustic, phonological, morphological, syntactic, and semantic inputs, more closely approximating the human language learning environment studied in cognitive psychology and developmental linguistics. Yet, despite these capabilities, few studies exploit the tools of theoretical linguistics to probe how such models implement rule-based processes, leaving a gap in mechanistic interpretability research. By returning to well-established phonological and morphological phenomena, we gain a set of descriptive and explanatory benchmarks for evaluating emergent symbolic structures, providing principled constraints for alignment maps and causal interventions. Interventions and counterfactual reasoning in S3Ms make it possible to probe not only the existence of rules, but also their mechanistic implementation, revealing whether the network constructs representations consistent with symbolic generalizations (see *Section* A.1.3 for an additional case study from Yoruba comparing causal abstraction between different model organisms).

We can hypothesize that **mechanistically similar causal structures could emerge even under substantially different training distributions**. Consistent intervention outcomes and counterfactual predictions across phonemic, S3M, and omni-model types (e.g., with lesions blocking nasal assimilation or hidden states preserving vowel harmony despite opacity in the case study) would indicate that mechanistically similar causal abstractions can emerge under divergent training distributions. Consistent intervention outcomes across these models indicate that similar causal abstractions arise even from divergent training regimes. Comparing monolingual and multilingual encoders provides a systematic framework to test developmental universals, asking whether mechanistically similar causal circuits arise across languages despite divergent training distributions and input modalities [2, 5]. In this way, treating models as experimental organisms situates emergent symbolic behaviour within a falsifiable, mechanistically grounded interpretability program. By measuring mechanistic similarity through alignment maps, we can identify which circuits implement abstract linguistic rules within a single architecture and potentially across architectures. When these circuits are conserved across models, this could potentially facilitate interventions, like steering accent features, or facilitating more resource-efficient cross-lingual transfer to low-resource languages by systematically changing behaviour in a controlled, rule-consistent manner.

5 Limitations and Open Problems

Interventions themselves remain fragile, with necessity and sufficiency tests sensitive to design choices. Moreover, the mapping between distributed neural representations and discrete linguistic rules may remain partial at best. While phonology and morphology are tractable domains, extending this method to domains such as semantics and pragmatics remains challenging. Linguistic interpretability is, of course, only a subset of actionable interpretability research. The LEH avoids stipulating where these causal abstractions come from, which are a continued subject of debate in linguistics, restricting itself to descriptive generalisations for application to interpretability research (see *Section* A.2 for further discussion on theory-neutrality of alignment maps).

6 Conclusion

Polysemantic neurons and non-linear representations challenge naive assumptions about the interpretability of deep language models. Interventions, counterfactual analyses, and mechanistic interpretability provide the necessary tools to probe these representations. By constraining alignment maps to **linguistically motivated causal templates**, causal abstraction becomes a falsifiable framework for modelling the emergence of symbolic representations in language models. Across architectures and modalities—including S3Ms as "omni-model"s, phonemic models, and multilingual encoders – we can investigate how mechanistically similar causal structures emerge across different model organisms. Linguistic theory thus provides a blueprint for principled interpretability, linking emergent symbolic behaviour in neural networks to abstract, human-relevant causal mechanisms across multiple levels of structures.

References

[1] D. Archangeli and D. Pulleyblank. Yoruba vowel harmony. *Linguistic Inquiry*, 20(2):173–217, 1989.

[2] C. Arnett, T. A. Chang, J. A. Michaelov, and B. Bergen. On the acquisition of shared grammatical representations in bilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20707–20726, Vienna, Austria, 2025. Association for Computational Linguistics.

[3] A. Arora, D. Jurafsky, and C. Potts. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*, 2024.

[4] E. Baković. Opacity and ordering. In *The Handbook of Phonological Theory*, pages 40–67. 2011.

[5] J. Brinkmann, C. Wendler, C. Bartelt, and A. Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. *arXiv preprint arXiv:2501.06346*, 2025.

[6] S. Bromberger and M. Halle. Why phonology is different. *Linguistic Inquiry*, 20:51–70, 1989.

[7] N. Chomsky. *New Horizons in the Study of Language and Mind.* Cambridge University Press, 2000.

[8] N. Chomsky and M. Halle. *The Sound Pattern of English.* Harper and Row, New York, 1968.

[9] A. Geiger, D. Ibeling, A. Zur, M. Chaudhary, S. Chauhan, J. Huang, ..., and T. Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

[10] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR, March 2024.

[11] Z. Goriely and P. Buttery. Babylm's first words: Word segmentation as a phonological probing task. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 522–539, Vienna, Austria, 2025. Association for Computational Linguistics.

[12] P. Kiparsky. Historical linguistics. In W. O. Dingwall, editor, *A Survey of Linguistic Science*, pages 576–642. University of Maryland Linguistics Program, College Park, 1971.

[13] P. Kiparsky. Abstractness, opacity, and global rules. In O. Fujimura, editor, *Three Dimensions of Linguistic Theory*, pages 57–86. TEC, Tokyo, 1973.

[14] M. Lavechin, Y. Sy, H. Titeux, M. A. Cruz Blandón, O. Räsänen, H. Bredin, E. Dupoux, and A. Cristia. Babyslm: Language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592, Dublin, Ireland, 2023. ISCA.

[15] A. Mueller, A. Geiger, S. Wiegreffe, D. Arad, I. Arcuschin, A. Belfki, ..., and Y. Belinkov. Mib: A mechanistic interpretability benchmark. In *Forty-second International Conference on Machine Learning*, 2025.

[16] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. Distill, 2020.

[17] A. Pasad, C.-M. Chien, S. Settle, and K. Livescu. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12:372–391, 2024.

[18] A. Prince and P. Smolensky. Optimality theory: Constraint interaction in generative grammar. Technical Report RuCCS-TR-2, ROA-537, Rutgers University Center for Cognitive Science, 1993. Published 2004, Malden, MA: Blackwell.

[19] Z. W. Pylyshyn. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1):111–132, 1980.

[20] L. Sharkey, S. Black, and B. Millidge. Current themes in mechanistic interpretability research, 2022.

[21] G. Shen, M. Watkins, A. Alishahi, A. Bisazza, and G. Chrupała. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4250–4261, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[22] C. Shi, N. Beltran Velez, A. Nazaret, C. Zheng, A. Garriga-Alonso, A. Jesson, ..., and D. Blei. Hypothesis testing the circuit hypothesis in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 94539–94567, 2024.

[23] H. A. Simon. *Computational Theories of Cognition*. 1996.

[24] D. Sutter, J. Minder, T. Hofmann, and T. Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? *arXiv preprint arXiv:2507.08802*, 2025.

A Linguistic Templates for Causal Abstraction

We provide more linguistic motivation and detail for how we propose to convert descriptive linguistic generalisations into linguistic templates for causal abstraction, beyond the Case Study in *Section* 3.1. In *Section* A.1, we present some background on rule-based analyses of phonological processes and ordering, before outlining how to derive a more general schema for phonological templates for causal abstraction. We present an additional case study for comparing causal abstraction between different model organisms in *Section* A.1.3.

A.1 Phonological Rules and Rule Ordering

A.1.1 Rule-based Analyses of Phonological Processes

The phonology of a natural language is concerned with the mental rules or constraints that govern how underlying forms (the lexicon) are mapped to surface pronunciations. It generates a set of surface forms that are used in communication via principled transformations that interact. A central insight of the rule-based framework inaugurated by *The Sound Pattern of English* [8] is that these transformations are not arbitrary: they follow ordered rules. Kiparsky's (1971, 1973a) work on opacity, in which later rule applications obscure earlier ones, defined one of the most influential mechanisms in this system [12, 13].

Formally, a phonological rule can be expressed as a context-sensitive rewrite:

$$A \rightarrow B \;/\; C\_D$$

A rule of this form states that a segment A is realized as B when it appears in the environment between C and D. Classic phonological processes illustrate how such rules can operate:

- **Assimilation:** Turkish nasal assimilation applies when a nasal consonant takes on the place of articulation of a following consonant, e.g. /sen+ge/ → [seŋge] versus /sen+de/ → [sende].
- **Harmony:** In Turkish vowel harmony, suffix vowels copy the [back] or [round] features of the root, e.g. /kitap+da/ → [kitapta] 'in the book', with [+back] spreading from the stem to the suffix vowel.
- **Dissimilation:** In Latin, sequences of adjacent aspirates were dispreferred, leading to forms like *futtilis* from **futthlis*, where one aspirated stop loses aspiration to avoid repetition.
- **Devoicing:** In German, obstruents are obligatorily devoiced in word-final position, e.g. /Bund/ → [bunt], neutralizing the [voice] contrast in coda position.

- **Tone spreading:** In Yoruba, a high tone can spread rightward onto a following toneless syllable, e.g. /ó + ra/ → [órá] 'he bought', where the floating high tone from the subject prefix attaches to the verb root.

These processes show that rules can either *spread* features (assimilation, harmony, tone), *delete or neutralize* features (devoicing), or *actively avoid* feature co-occurrence (dissimilation). In each case, the causal relation between underlying features and surface outputs can be captured with rule schemata of the form $A \rightarrow B/C\_D$.

Opacity arises when the surface realization obscures the operation of such rules. As Kiparsky (1973: 79) defined it, a rule is opaque when (i) its predicted environment appears on the surface without triggering the expected alternation, or (ii) its output appears on the surface in an unexpected environment [13]. In both cases, the surface form hides or obscures the causal force of the rule. Kiparsky's substantive claim was that opaque rules pose greater learnability challenges and thus tend to simplify over time through diachronic change.

A classic example of opacity in Turkish illustrates this. Turkish exhibits both **vowel harmony** (suffix vowels copy the [back] feature of the stem) and **final devoicing** (voiced obstruents become voiceless in coda position). Consider the underlying form /kitap+da/ 'in the book'. Harmony predicts a back suffix vowel, yielding */kitapda/. Devoicing, however, applies after harmony, turning the /d/ into [t], so the surface form is [kitapta]. Crucially, the devoicing rule obscures the harmony alternation: the suffix vowel appears with [+back] even though the expected trigger (/d/) is neutralized on the surface. This is a case of **counterbleeding**: harmony has already applied before devoicing removes the conditioning environment, leaving the harmony effect "opaque" in the surface form.

The mechanics of opacity depend on *rule ordering*. Bromberger & Halle (1989: 58–59) note that rules do not all apply directly to the underlying representation [6]. Instead, each rule applies in sequence, taking as its input the partially transformed representation produced by earlier rules. Thus, the order of rules determines whether their causal effects feed into one another or bleed each other's applicability. Four logically possible relations emerge:

1. Feeding: Rule A creates new inputs for Rule B.
2. Bleeding: Rule A removes potential inputs for Rule B.
3. Counterfeeding: Rule B would have created new inputs for Rule A, but its later position prevents this.
4. Counterbleeding: Rule B would have removed inputs to Rule A, but its later position prevents this.

Feeding and bleeding represent transparent interactions: they align surface outcomes with underlying rules. Counterfeeding and counterbleeding are the hallmarks of opacity: they produce surface patterns in which the causal force of rules is hidden. For example, in the Turkish case above, harmony and devoicing interact in a counterbleeding order, such that the harmony rule applies even though its conditioning environment is eliminated by the later devoicing rule. The counterfactual logic here is crucial: opacity is always about comparing what the surface *should have looked like* if rules applied differently against what it actually looks like.

Opacity thus provides a natural locus for causal analysis. Rules can be thought of as interventions on hidden representations, with ordering constraints determining how their outputs percolate to the surface. The mismatch between predicted and observed forms reveals precisely the kind of hidden-state mediation that causal abstraction seeks to capture [4].

A.1.2 Phonological Templates

Phonological rules such as assimilation, harmony, dissimilation, devoicing, or tone spreading define compact causal templates, which we denote A_{phon}. These templates specify how underlying features (e.g. [nasal], [voice], [back], [tone]) spread, neutralize, or interact. An alignment map ϕ_{phon} projects distributed hidden states in a model into interpretable causal variables Z_{phon}, such as feature values or abstract segments. Once defined, such mappings make it possible to probe necessity and sufficiency by performing interventions on these causal variables and predicting counterfactual outputs.

The practical key to this approach is the construction of *minimal pairs*. A minimal pair consists of two input forms that differ only with respect to the structural environment relevant for a rule. For example, nasal assimilation in Turkish distinguishes /sen+de/ → [sende] from /sen+ge/ → [seŋge]. By creating such pairs, we can test whether lesioning the causal variable [place] in the assimilation circuit prevents velarization, or whether stimulating it causes overapplication. The same logic applies to opacity: /kitap+da/ surfaces as [kitapta], with harmony obscured by devoicing. Here, lesioning harmony should break the suffix alternation ([kitapde]), while stimulating harmony should preserve hidden [+back] values even when devoicing erases them on the surface.

More generally, the interpretability logic extends across a wide typological range of phonological processes:

- **Assimilation:** feature spreading can be directly tested by probing whether intervention on a feature node (e.g. [nasal]) induces changes in neighboring segments.
- **Dissimilation:** feature avoidance patterns (e.g. Latin *hominem* → *hominem* vs. **homonem*) can be probed by counterfactual suppression of duplicated features.
- **Final devoicing:** intervention on [voice] in German or Turkish word-final position tests whether the model encodes neutralization in a positionally constrained circuit.
- **Tone:** in autosegmental languages (e.g. Yoruba, Igbo, Mandarin), tone can be modeled as a separate causal tier. Minimal pairs involving tone spreading, downstep, or floating tones allow probes of whether hidden tonal features are preserved even when not overtly realized.
- **Harmony and Vowel Reduction:** cross-linguistic examples (Finnish, Turkish, Hungarian) permit interventions on [back], [round], or [ATR] features, testing whether the model encodes long-distance dependencies as modular circuits.

From these worked cases we can abstract a general feature template for comparing models:

$$A_{\text{phon}}: \ \langle \text{Underlying Form},\ \text{Feature Set},\ \text{Rule},\ \text{Surface Form},\ \text{Intervention Predictions} \rangle$$

This template captures the causal structure of phonological processes by linking:

1. **Underlying form** — the morphophonological input (e.g., /sen+ge/).
2. **Feature set** — relevant distinctive features (e.g., [+nasal], [+velar]).
3. **Rule** — the process applied (assimilation, harmony, devoicing).
4. **Surface form** — the attested output (e.g., [seNge]).
5. **Intervention predictions** — necessity/sufficiency counterfactuals (e.g., lesioning assimilation → [senge]).

By instantiating A_{phon} across multiple languages and processes, we obtain a principled set of causal probes for S3Ms and omni-models. These templates let us test whether models implement rules transparently, exhibit opacity, or deviate in systematic ways, thereby aligning model interpretability research with decades of theoretical insight in phonology.

Figure 1 illustrates the general architecture of a phonological causal template. Hidden states h are mapped by ϕ_{phon} into causal variables Z_{phon}, such as [nasal] or [voice]. Interventions on Z_{phon} yield counterfactual predictions over the output phoneme sequence. If the alignment map is faithful, then interventions will yield outputs consistent with the linguistic rule, and necessity/sufficiency criteria will be satisfied. If not, the encoding is entangled and fails to provide a genuine causal abstraction.

Minimal pairs thus instantiate a controlled experimental paradigm within the symbolic domain of phonology. By systematically intervening on alignment maps derived from such pairs, we can reveal whether the model's internal states encode phonological rules as modular causal circuits or as entangled distributed patterns. This provides a linguistically rigorous test for universality in symbolic emergence: if the same templates and intervention logic apply across unrelated processes such as assimilation, dissimilation, devoicing, and tone spreading, we obtain evidence for cross-linguistic invariants in causal abstraction. Moreover, when applied to self-supervised spoken models, tonal and segmental contrasts allow direct testing of whether speech representations capture phonological causality beyond orthographic input, thereby broadening the empirical reach of causal interpretability.

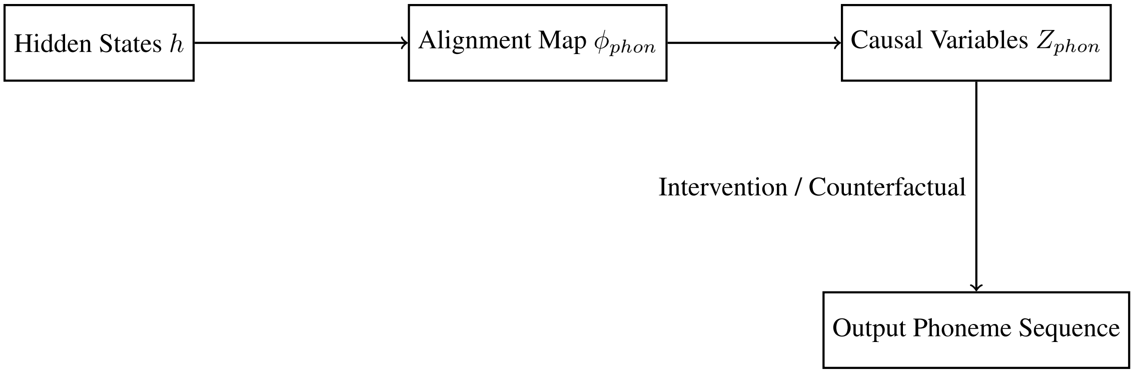
Hidden States h
Alignment Map ϕ_{phon}
Causal Variables Z_{phon}
Intervention / Counterfactual
Output Phoneme Sequence

Figure 1: Causal abstraction for a phonological assimilation rule. Hidden states are mapped to interpretable causal variables such as [nasal] or [voice], which can be directly intervened upon to predict counterfactual surface forms.

A.1.3 Case Study for Cross-Model Comparison: Vowel Harmony and Tone Spreading in Yoruba

Self-supervised speech models (S3Ms) provide a natural testbed for examining how multi-level linguistic constraints are encoded across representational layers. Prior work by Shen et al (2024) found S3Ms encode lexical tone to a significant degree even when they are trained on data from non-tonal languages [21]. Yoruba phonology is especially revealing, as it combines Advanced Tongue Root (ATR) vowel harmony with tone spreading [1], producing natural cases of opacity that allow us to test necessity and sufficiency within an S3M's representational circuits. Unlike text-only LMs, S3Ms integrate acoustic correlates (formant structure for [±ATR], F0 for tone), phonological categories (harmony and tone), and morphological concatenation (suffix alternations).

ATR harmony partitions the vowel inventory into two sets, [+ATR] vowels produced with the tongue root advanced and [-ATR] vowels produced with it retracted, and requires suffix vowels to match the ATR specification of the root. Thus, /ba + -ra/ → [bárá] (harmonic, [+ATR]) contrasts with ɛ̀+ -ra/ → [ɛ̀ɹɛ̀] (harmonic, [-ATR]). Tone spreading operates independently, allowing a high tone to extend rightward (/ó + ra/ → [órá]), but blocking in some contexts due to downstep (/ó + ra/ → [ó!rá]). Opacity arises when harmony applies at the suffixal level but is masked by tone spreading, such that an ATR feature persists in latent space while its surface realization is obscured by a compressed F0 contour.

We capture this interaction with a compact causal template, $A_{\text{phon}} = \{[\text{ATR}] \rightarrow \text{suffix vowel}, [\text{Tone}] \rightarrow \text{rightward spread}, [\text{ATR} \times \text{Tone}] \rightarrow \text{opacity masking}\}$, which constrains alignment maps ϕ between latent S3M layers and symbolic phonological features.

Comparing S3Ms with phoneme-based and text-based LMs highlights their distinct representational scopes. S3Ms such as wav2vec and HuBERT jointly encode acoustics, phonological categories, morphology, and higher-order structure, allowing interventions at the acoustic–phonological interface (e.g., lesioning F0 tracking to probe tone spreading). In such cases, latent ATR features may survive even when surface forms are neutralized by tone, revealing genuine opacity. Phoneme-trained LMs [11, 14], in contrast, directly encode symbolic strings without acoustic grounding; they provide transparency in rule learning, but any harmony or opacity effect must be derived from symbolic patterns alone, without evidence of under-determined acoustic cues. Text-based LMs are not able to learn distributional information about phonological features, making knowledge of opacity less likely.

Predictions under intervention sharpen these contrasts. For ATR harmony, lesioning an S3M may yield disharmonic forms such as [bɛ̀ra], while stimulation can induce over-application ([bára]); phoneme-LMs are expected to show similar symbolic outcomes, but without residual acoustic traces. For tone spreading, lesioning may block spread ([óra]), while stimulation induces over-spreading ([óóra]); again, phoneme-LMs track the symbolic rule but cannot reveal acoustic persistence. In cases of opacity, such as /ó + r`ɛ/ → [ór`ɛ], an S3M may preserve hidden [+ATR] features in its internal space even when they are masked at the surface, whereas a phoneme-LM has no mechanism to represent the latent feature beyond the surface symbolic outcome.

Table 2: **Predicted linguistic interpretability differences between different model organisms**: Comparison between S3Ms, phoneme-trained LMs, and text-based LMs in probing Yoruba opacity. S3Ms allow causal interventions on acoustic correlates, while phoneme-LMs offer symbolic transparency and text-LMs provide little access to phonological constraints.

Model Type	**Representation**	**Hypothesised Interpretability**	**Predicted Behavior on Yoruba Opacity**
S3Ms (wav2vec, HuBERT, omni-models)	Encode acoustics, phonological categories, morphology, syntax, and semantics jointly	Allow intervention at the acoustic–phonological interface (e.g., lesion F0 tracking to probe tone spreading)	Latent ATR features may survive even when surface forms are neutralized by tone, making true opacity detectable
Phoneme-LMs (trained on symbolic transcriptions)	Directly encode discrete phoneme strings, no acoustic grounding	Transparency in rule learning: harmony/opacity effects must be induced from symbolic patterns only	Harmony and tone interaction observable at the symbolic level, with opacity "cleaner" but lacking evidence for acoustic underdetermination
Text-LMs (orthography)	Encode morphology and syntax, but phonological features are indirect or absent	Low interpretability for phonological rules	Harmony and tone absent unless orthographically reflected; opacity untestable

This comparative design underscores that S3Ms provide access to both symbolic and sub-symbolic levels of phonological computation, while phoneme-LMs serve as a clean symbolic baseline. If S3Ms preserve opaque features like ATR harmony even when acoustically masked by tone, they align with symbolic phonological theories that posit hidden structure. If instead their behavior converges with phoneme-LMs, this suggests that S3Ms abstract away from acoustics into symbolic representations earlier than expected. In either case, Yoruba phonology illustrates how theoretical constructs from linguistics can serve as probes for mechanistic interpretability, distinguishing surface correlation-tracking from genuine rule-based generalization.

A.2 Can Linguistic Templates be Theory-Neutral?

The causal abstraction and feature-template methodology we propose is largely theory-neutral in the sense that it does not presuppose a particular representational framework for phonological computation. Our minimal pair interventions, alignment maps, and necessity/sufficiency probes rely only on observable input-output mappings and hypothesized intermediate features (e.g., [nasal], [voice], [back], [tone]). This allows us to operationalize a general notion of "causal rule" in a model-agnostic fashion: any latent variable or subspace that systematically mediates between underlying and surface forms can be tested, regardless of whether the linguistic theory is rule-based, constraint-based, or emergent.

In a rule-based phonology (RBP) setting, templates correspond naturally to ordered rewrite rules. Necessity and sufficiency tests map directly onto interventions in the corresponding causal circuits: lesioning a circuit should block the rule's effect (necessity), while stimulating it should over-apply the rule (sufficiency). Opacity arises naturally as a consequence of rule ordering: counterfeeding and counterbleeding interactions are straightforward to encode in the intervention framework.

If we were to assume Optimality Theory (OT) instead, the structure of the templates would appear differently. OT posits parallel, violable constraints rather than sequential rules [18]. Surface forms arise as the optimal candidate that best satisfies a ranked constraint hierarchy. In this view:

- Minimal pairs and interventions could still be defined, but instead of probing an "ordered rule circuit," one would probe whether a set of latent variables encodes constraint satisfaction preferences.
- Necessity could correspond to lesioning the representation of a constraint's ranking or weighting and observing whether the optimal surface form changes.

- Sufficiency would involve artificially boosting the latent representation of a constraint to see if it forces candidate selection in line with that constraint.
- Opacity effects are handled differently: rather than emerging from rule ordering, they result from constraint interaction and ranking. A lower-ranked constraint might be satisfied in some contexts but overridden in others, creating apparent opacity at the surface. The latent representation would need to encode these ranking relations rather than sequential application steps.

Thus, the same experimental apparatus—alignment maps, minimal pairs, interventions, and counterfactuals—remains applicable, but its interpretive lens shifts. Under RBP, interventions probe sequential causal dependencies; under OT, they probe latent rankings and the combinatorial effect of violable constraints. In this sense, our framework is flexible enough to accommodate different phonological theories, while still producing falsifiable predictions about the mechanistic implementation of phonological generalizations in neural models.

B Localization Complexity Metric

In our linguistically-constrained framework, we propose to only consider alignment maps ϕ and translations τ that respect phonological, morphological, or syntactic templates. We position that we might be interested in measuring a *localization complexity metric* $C(\phi)$ to capture how "spread out" a causal variable Z_i is across a network's hidden states.

In our linguistically-constrained framework, we propose to consider only alignment maps ϕ and translations τ that respect phonological, morphological, or syntactic templates. This restriction ensures that candidate mappings are interpretable and consistent with known linguistic generalizations. Measuring the *localization complexity metric* $C(\phi)$ allows us to quantify how "spread out" a causal variable Z_i is across the network's hidden states. We care about this because a low localization complexity indicates that Z_i is encoded in a compact, modular set of hidden units, which makes the representation easier to interpret, test, and manipulate. Conversely, a high complexity suggests that the variable is distributed across many units, potentially obscuring causal structure and making it harder to verify whether the network genuinely implements the abstract linguistic rules. By linking $C(\phi)$ to linguistically-constrained mappings, we gain a principled, falsifiable measure of how efficiently and faithfully neural networks represent symbolic causal variables.

We outline the conceptual benefits of a localization complexity metric which can provide an intuitive, operational handle on the interpretability of a candidate alignment, linking the abstract symbolic rules in linguistic theory to measurable structure in the network.

$C(\phi)$ measures whether a variable is encoded in a compact, interpretable way or dispersed across many hidden units. Low $C(\phi)$ values indicate that a causal variable is implemented in a sparse, modular subspace of the network, consistent with symbolic, linguistically-informed representations. High values indicate that the causal effect is distributed, which can signal entangled representations that are harder to interpret.

Formally, consider an alignment map $\phi : H \to Z$ from hidden states H to causal variables Z. For a given causal variable Z_i, $C(\phi)$ is defined as the minimal number of disjoint hidden subspaces $H_1, \ldots, H_k \subseteq H$ required to implement the causal effect of Z_i under interventions. That is, for an intervention I on Z_i, the effect on the network's output $f(h \leftarrow \phi^{-1}(I(Z_i)))$ can be reproduced by manipulating only these subspaces. Lower k corresponds to a more localized, interpretable representation. By combining $C(\phi)$ with our linguistically-constrained alignments, we obtain a quantitative, falsifiable measure of whether neural representations implement the hypothesized symbolic rules in a structured and discoverable manner (see Appendix B for further details).

The *localization complexity metric* $C(\phi)$ quantifies how distributed a causal variable Z_i is across the hidden state space H of a neural network. Formally, consider an alignment map $\phi : H \to Z$ linking hidden representations to causal variables. $C(\phi)$ is defined as the minimal number of hidden subspaces required to implement a causal effect associated with Z_i under interventions:

$$C(\phi; Z_i) = \min\{k \mid Z_i \text{ can be reconstructed from } k \text{ disjoint subspaces of } H\}. \tag{1}$$

A lower $C(\phi)$ indicates that the representation of Z_i is sparse and localized, consistent with emergent symbolic structure. Conversely, a higher $C(\phi)$ indicates that Z_i is distributed across many hidden units, suggesting a more entangled or non-symbolic encoding.

Operationally, $C(\phi)$ can be estimated using modern interpretability tools:

- **Circuit localization**: Identify minimal sets of neurons or attention heads whose activations correlate with changes in Z_i.
- **Attribution patching**: Intervene on subsets of hidden states and measure the effect on predicted outcomes W, pruning irrelevant subspaces.
- **Distributed alignment search (DAS)**: Search for combinations of hidden state subspaces that maximally recover Z_i's causal effects.

In practice, $C(\phi)$ provides a principled measure to compare candidate alignments: mappings that achieve low $C(\phi)$ while preserving predicted causal effects are considered more interpretable and consistent with symbolic templates. This allows researchers to reject alignments that, although mathematically valid, are overly complex or distributed to yield meaningful mechanistic explanations.

Finally, $C(\phi)$ can be generalized across all causal variables Z by averaging over the individual complexities:

$$C(\phi) = \frac{1}{|Z|} \sum_{i=1}^{|Z|} C(\phi; Z_i), \tag{2}$$

providing a global metric for the sparsity and modularity of a network's causal representations.