All or None: Identifiable Linear Properties of Next-Token Predictors in Language Modeling

Emanuele Marconato[♣] University of Trento
Sébastien Lachapelle SAIT AI Lab, Montreal
Sebastian Weichwald[♠] University of Copenhagen
Luigi Gresele[♠] University of Copenhagen

Abstract

We analyze identifiability as a possible explanation for the ubiquity of linear properties across language models, such as the vector difference between the representations of "easy" and "easiest" being parallel to that between "lucky" and "luckiest". For this, we ask whether finding a linear property in one model implies that any model that induces the same distribution has that property, too. To answer that, we first prove an identifiability result to characterize distribution-equivalent next-token predictors, lifting a diversity requirement of previous results. Second, based on a refinement of relational linearity [Paccanaro and Hinton, 2001; Hernandez et al., 2024], we show how many notions of linearity are amenable to our analysis. Finally, we show that under suitable conditions, these linear properties either hold in all or none distribution-equivalent next-token predictors.

1 Introduction

In natural language processing, it is well-established that linear relationships between high-dimensional, real-valued vector representations of textual inputs reflect semantic and syntactic patterns. This was motivated in seminal works [Rumelhart and Abrahamson, 1973, Hinton et al., 1986a,b, Rumelhart et al., 1986, Bengio et al., 2000] and extensively validated in word embedding models [Mikolov et al., 2013a,b, Pennington et al., 2014] as well as modern large language models trained for next-token prediction [Burns et al., 2022, Merullo et al., 2023, Tigges et al., 2023, Pal et al., 2023, Gurnee and Tegmark, 2023, Bricken et al., 2023].

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

This ubiquity is puzzling, as different internal representations can produce identical next-token distributions, resulting in distribution-equivalent but internally distinct models. This raises a key question: **Are the observed linear properties shared across all models with the same next-token distribution?** Our **main result** is a mathematical proof that, under suitable conditions, certain linear properties hold for either all or none of the equivalent models generating a given next-token distribution. We demonstrate this through three main contributions.

The **first main contribution** (Section 3) is an identifiability result characterizing distribution-equivalent next-token predictors. Our result is a generalization of the main theorems by Roeder et al. [2021] and Khemakhem et al. [2020a], relaxing the assumptions of diversity and equal representation dimensionality. This result is of independent interest for research on identifiable representation learning since our analysis is applicable to several discriminative models beyond next-token prediction [Roeder et al., 2021].

Our **second main contribution** (Section 4) is to subsume several linear properties in a common framework. We start by defining an analogue to *relational linearity* [Paccanaro and Hinton, 2001], where the definition only relies on terms appearing in our identifiability result. The key idea is to represent entities as vectors, binary relations as matrices, and to model the operation of applying a relation to an entity through matrix-vector multiplication, which yields the vector corresponding to the related entity. For example, in the sentence *"Jimi Hendrix plays the guitar"*, the relation between the entities *"Jimi Hendrix"* and *"the guitar"* is signified by the word *"plays"* and encoded as a matrix-vector multiplication in representation space. We then define relational counterparts to linearity properties described and analyzed in previous works [Arora et al., 2016, Allen and Hospedales, 2019, Park et al., 2024a, Heinzerling and Inui, 2024], thus making them amenable to our analysis.

[♣] Work done while at the University of Copenhagen.
[♠] Shared last author.

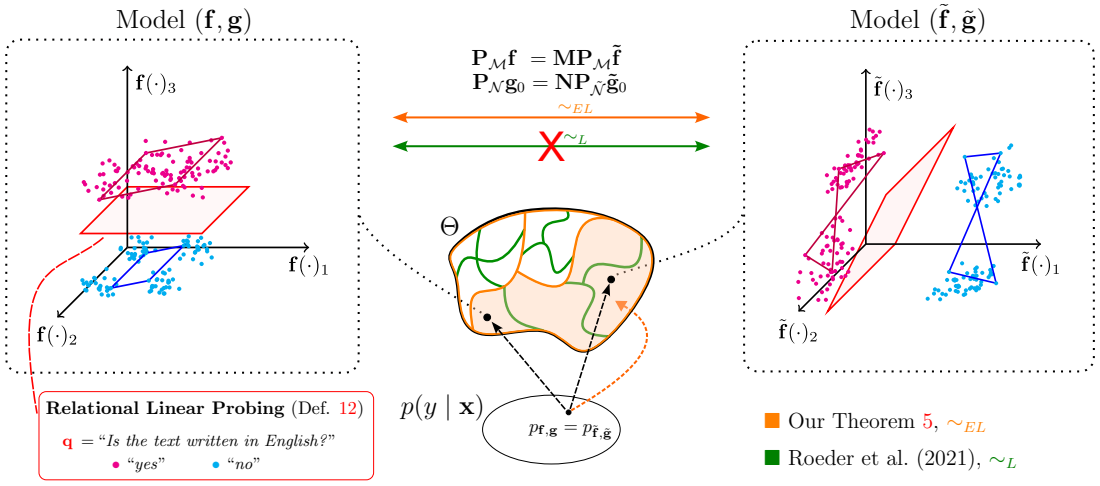

Figure 1: **Identifiability of linear properties.** Plots in the left and right dotted squares show the embeddings of two next-token predictors $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$ that generate the same distribution $p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}$ within a set of conditionals distributions $p(y \mid \mathbf{x})$. Theorem 5 proves a one-to-one correspondence (the dashed **orange** arrow) between conditional distributions and $\sim_{EL}$-equivalent models (the **orange** partitions of Θ). This extends a result by Roeder et al. [2021] characterizing $\sim_L$-equivalent models (**green** partitions of Θ). Here $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ while the embedding representations are not equal up to a linear transformation (thus $(\mathbf{f}, \mathbf{g}) \not\sim_L (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$), as shown by how the **purple** and **blue** parallelograms in the embeddings of the left model $(\mathbf{f}, \mathbf{g})$ get distorted in those of the right model $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$. Both models display relational linear probing for the query $\mathbf{q}$="*Is the text written in English?*": one can linearly separate the embeddings of textual inputs which, when concatenated with $\mathbf{q}$, have "*yes*" as the likeliest next token, from those that yield "*no*". In Theorem 14, we provide conditions under which all or none of the models in the $\sim_{EL}$ equivalence class share the same linear property.

Our **third main contribution** (Section 5) is to show that under suitable conditions, these linear properties either hold in all or none of the models generating a given distribution. For this, we combine the definitions in Section 4 and our characterization of distribution-equivalent next-token prediction models in Section 3. Identifiability theory thus enables us to explain what linearity properties are shared across language models which are equivalent next-token predictors. We illustrate this result in Figure 1.

Lastly, in Section 6 we discuss implications of our findings and in Section 7 we discuss connections to related works and future research directions.

2 Preliminaries

Notation. Italic font letters denote scalars, *e.g.*, a; bold font lower-case letters denote vectors and sequences, *e.g.*, $\mathbf{x}$; and bold font upper-case letters denote matrices, *e.g.*, $\mathbf{M}$. We use $\mathbf{M}^+$ to denote the pseudo-inverse of $\mathbf{M}$. We use the short-hand $[k] = \{1, \ldots, k\}$. Given a finite dictionary of tokens $\mathcal{A}$, the space of all possible finite sequences (or sentences) is denoted by $\mathsf{Seq}(\mathcal{A})$, which is the power set of $\mathcal{A}$. With $\mathbf{x}_{1:t}$, we denote the sub-sequence $(\mathbf{x}_1, ..., \mathbf{x}_t)$ of a sequence $\mathbf{x}$, *i.e.*, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T) \in \mathsf{Seq}(\mathcal{A})$. We use the symbol $\frown$ to denote the concatenation of two elements of $\mathsf{Seq}(\mathcal{A})$, *e.g.*, $\mathbf{x}_1 \frown \mathbf{x}_2 = (\mathbf{x}_{1,1}, ..., \mathbf{x}_{1,k}) \frown (\mathbf{x}_{2,1}, ..., \mathbf{x}_{2,l}) = (\mathbf{x}_{1,1}, ..., \mathbf{x}_{1,k}, \mathbf{x}_{2,1}, ..., \mathbf{x}_{2,l}) \in \mathsf{Seq}(\mathcal{A})$. For a function $\mathbf{h}$, we denote its image by $\mathrm{Im}(\mathbf{h})$. For a k-dimensional subspace $\mathcal{H} \subseteq \mathbb{R}^d$ spanned by an orthonormal basis $\{\mathbf{s}_1, ..., \mathbf{s}_k\}$, we use $\mathbf{P}_{\mathcal{H}} = \sum_{l=1}^{k} \mathbf{s}_l \mathbf{s}_l^\top \in \mathbb{R}^{d \times d}$ to denote the orthogonal projector onto that space.

Next-token predictors. Here, we introduce the general form of next-token predictors used in our analysis.[1] We consider models which take text sequences $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ of tokens $\mathcal{A}$ as input. A next-token predictor $(\mathbf{f}, \mathbf{g})$ consists of two functions: $\mathbf{f} : \mathsf{Seq}(\mathcal{A}) \to \mathbb{R}^d$ maps sequences to their representations, called *embeddings*, and $\mathbf{g} : \mathcal{A} \to \mathbb{R}^d$ maps tokens to their representations, called *unembeddings*. Let Θ_d be the set of all tuples $(\mathbf{f}, \mathbf{g})$ with representation dimensionality d and let $\Theta := \bigcup_{d=1}^{\infty} \Theta_d$ be the set of all tuples with any dimensionality d. A next-token predictor models the conditional distribution of the next-token x_{t+1} given the context $\mathbf{x}_{1:t}$ as

$$p_{\mathbf{f},\mathbf{g}}(x_{t+1} \mid \mathbf{x}_{1:t}) := \frac{\exp(\mathbf{f}(\mathbf{x}_{1:t})^\top \mathbf{g}(x_{t+1}))}{Z(\mathbf{x}_{1:t})}, \qquad (1)$$

where $Z(\mathbf{x}_{1:t}) := \sum_{y \in \mathcal{A}} \exp[\mathbf{f}(\mathbf{x}_{1:t})^\top \mathbf{g}(y)]$ is a normal-

[1] In Appendix A we show that decoder-only transformer models can be expressed in this form [Roeder et al., 2021].

izing constant. Models of the form in Equation (1) are trained to maximize the conditional log-likelihood of the data. For a data distribution $p_{\mathcal{D}}$ over sequences $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$, its log-likelihood is given by:

$$\mathcal{L}(\mathbf{f}, \mathbf{g}) := \mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}} \left[\sum_{t=1}^{T(\mathbf{x})-1} \log p_{\mathbf{f},\mathbf{g}}(x_{t+1}|\mathbf{x}_{1:t}) \right], \quad (2)$$

where $T(\mathbf{x})$ indicates the length of the sequence $\mathbf{x}$. For a fixed representation dimensionality d, the next-token prediction objective can be written as

$$\max_{(\mathbf{f},\mathbf{g})\in\Theta_d} \mathcal{L}(\mathbf{f}, \mathbf{g}) . \quad (3)$$

We consider a setting where both $\mathbf{f}$ and $\mathbf{g}$ are non-parametric functions, so the model's expressivity is determined solely by the parameter d, corresponding to the dimensionality of the representation space.

Representation dimensionality and approximation capacity. In theory and for real-valued inputs, models of the form in Equation (1) have been proven to be universal approximators, *i.e.*, they can approximate any conditional distribution $p(x_{t+1} \mid \mathbf{x}_{1:t})$ to arbitrary precision, given a sufficiently large representation dimensionality d [Khemakhem et al., 2020a]; similar results may apply to next-token predictors. In practice, even if a representation dimensionality of d may be sufficient to represent a given distribution well, different practitioners may choose models with representations of different dimensionality. The linear identifiability results by Roeder et al. [2021], Khemakhem et al. [2020a] cannot be applied in this setting, since they consider models with equal representation dimensionality. Our Theorem 5 alleviates this tension between theory and practice: It characterizes identifiability of next-token predictors modeling the same conditional distribution irrespective of their representation dimensionality.

3 Identifiability of next-token predictors

We introduce a novel identifiability analysis for the model in Equation (1). In general, a statistical model $p_{\boldsymbol{\theta}}(\mathbf{x})$ parameterized by $\boldsymbol{\theta}$ is said to be *identifiable up to an equivalence relation* $\sim$ in the model class Θ if $p_{\boldsymbol{\theta}} = p_{\tilde{\boldsymbol{\theta}}} \implies \boldsymbol{\theta} \sim \tilde{\boldsymbol{\theta}}$. In other words, if two parametrizations $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}$ yield the same distribution, then they coincide under the equivalence relation $\sim$. The precise notion of equivalence depends on the problem setting. Although it is less commonly discussed, this implication can often be shown to hold also in the other direction so that $\sim$ is effectively a characterization of distribution-equivalent models, *i.e.*, $p_{\boldsymbol{\theta}} = p_{\tilde{\boldsymbol{\theta}}} \iff \boldsymbol{\theta} \sim \tilde{\boldsymbol{\theta}}$. In this section, we define an equivalence relation over tuples $(\mathbf{f}, \mathbf{g}) \in \Theta$ that characterizes models that entail the same next-token distribution. In other words, we want to define an equivalence relation $\sim$ over Θ such that $p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \iff (\mathbf{f}, \mathbf{g}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$; we then say $(\mathbf{f}, \mathbf{g})$, $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ are $\sim$-equivalent. Our characterization applies to pairs of models having different dimensionalities, i.e., $d \neq \tilde{d}$, as opposed to previous works only considering $d = \tilde{d}$ [Roeder et al., 2021, Khemakhem et al., 2020a,b, Lachapelle et al., 2023].

Previous works have shown that, under an assumption known as *variability* [Khemakhem et al., 2020a,b, Lachapelle et al., 2023] or *diversity condition* [Roeder et al., 2021], the representations extracted by distributionally-equivalent models are equal up to a linear invertible transformation. Intuitively, requires that at least one model $(\mathbf{f}, \mathbf{g})$ "spans" the whole representation space. To formally state the condition, we define the linear space spanned by the image $\mathrm{Im}(\mathbf{h}) \subseteq \mathbb{R}^d$ of a function $\mathbf{h}$ as $\mathrm{SIm}(\mathbf{h}) := \mathrm{span}\{\mathbf{v} \mid \mathbf{v} \in \mathrm{Im}(\mathbf{h})\}$. Additionally, for the unembeddings, we choose an arbitrary token $y_0 \in \mathcal{A}$ as a *pivot* for the remainder of the paper and define:

$$\mathbf{g}_0(y) := \mathbf{g}(y) - \mathbf{g}(y_0) \quad (4)$$

for all tokens $y \in \mathcal{A}$. The diversity condition can then be defined as follows:

Definition 1 (Diversity condition)**.** *We say that a model* $(\mathbf{f}, \mathbf{g})$ *with representation dimensionality* d *satisfies the diversity condition if* $\mathrm{SIm}(\mathbf{f}) = \mathrm{SIm}(\mathbf{g}_0) = \mathbb{R}^d$.

Intuitively, the diversity condition states that the dimension of the spaces spanned by the output of $\mathbf{f}$ and $\mathbf{g}_0$ match the dimension of the representation space. When both the diversity condition and $d = \tilde{d}$ hold, existing identifiability results for models of the form in Equation (1) (presented in Corollary 6) guarantee equivalence of representations up to an invertible linear transformation [Roeder et al., 2021, Khemakhem et al., 2020a]. Next, we show how to relax these two conditions via a more permissive equivalence relation.

3.1 Effective complexity of the model

To generalize previous results to settings where the diversity condition may not hold, we introduce the notion of *effective complexity* of a model. Starting from the conditional distribution captured by the model $(\mathbf{f}, \mathbf{g})$, we have that for every $y_0 \in \mathcal{A}$,

$$\begin{aligned} p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) &\propto \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y)) \\ &\propto \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y)) \exp(-\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y_0)) \\ &= \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y))) , \end{aligned}$$

indicating that the conditional distribution $p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x})$ is fully determined by the dot product $\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y)$. Denote by $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$ the orthogonal projectors onto $\mathcal{F} := \mathrm{SIm}(\mathbf{f})$ and $\mathcal{G} := \mathrm{SIm}(\mathbf{g}_0)$, respectively. Since

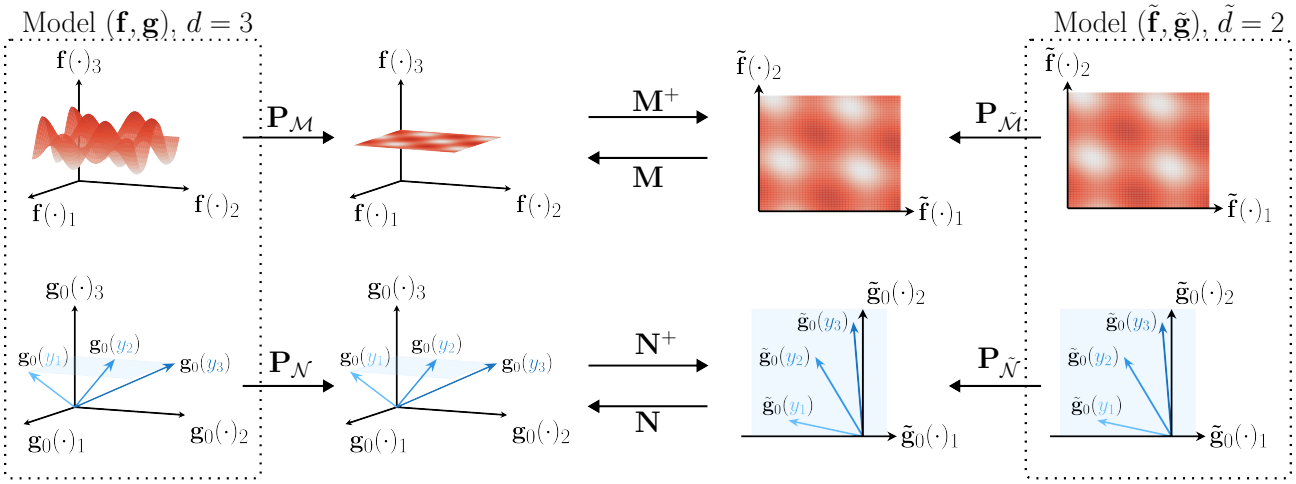

Figure 2: **Illustration of the $\sim_{EL}$ equivalence relation.** (**Left**) In the leftmost model, $(\mathbf{f}, \mathbf{g})$, the embeddings lie on a manifold $\mathrm{Im}(\mathbf{f}) \subsetneq \mathbb{R}^3$, yielding $\mathrm{SIm}(\mathbf{f}) = \mathbb{R}^3$. To ease visualization, $\mathrm{Im}(\mathbf{f})$ is plotted as a continuous manifold in the figure, although in practice textual inputs are discrete. The unembeddings lie on a two-dimensional space, $\mathrm{SIm}(\mathbf{g}_0) \cong \mathbb{R}^2$, drawn in light blue. Consequently the projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ map onto a two-dimensional subspace, *i.e.*, $\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{G}}$. (**Right**) The rightmost model, $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, represents both the embeddings and the unembeddings in a two-dimensional space. We therefore have $\mathrm{SIm}(\tilde{\mathbf{f}}) = \mathrm{SIm}(\tilde{\mathbf{g}}_0) = \mathbb{R}^2$, which implies that $\mathbf{P}_{\tilde{\mathcal{M}}} = \mathbf{P}_{\tilde{\mathcal{N}}} = \mathbf{I}$. Thus applying these projection matrices to embeddings and unembeddings leaves them unchaged (top-right and bottom-right grids). **(Center)** The equivalence relation $\sim_{EL}$ specifies that both $\mathbf{P}_{\mathcal{M}}\mathbf{f}$ and $\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}$, as well as $\mathbf{P}_{\mathcal{N}}\mathbf{g}$ and $\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}$, are related by linear invertible transformations defined by the matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{3\times 2}$.

both $\mathbf{f}(\mathbf{x}) = \mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x})$ and $\mathbf{g}_0(y) = \mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y)$, the dot product between the two can be evaluated as:

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = (\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}))^\top \mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) \tag{5}$$

$$= \mathbf{f}(\mathbf{x})^\top \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y)\,, \tag{6}$$

where we used the fact that $\mathbf{P}_{\mathcal{F}}^\top = \mathbf{P}_{\mathcal{F}}$, as a property of orthogonal projectors. In general, $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$ may not commute [Rehder, 1980]. We consider the subspaces:

$$\mathcal{M} := \mathrm{Im}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}), \quad \mathcal{N} := \ker(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^\perp\,, \tag{7}$$

which will be central to our characterization. Only when $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$ commute, we have that $\mathcal{M} = \mathcal{N} = \mathrm{SIm}(\mathbf{f}) \cap \mathrm{SIm}(\mathbf{g}_0)$ [Rehder, 1980]. In general, $\mathbf{P}_{\mathcal{M}}\mathbf{f} \neq \mathbf{f}$ and $\mathbf{P}_{\mathcal{N}}\mathbf{g}_0 \neq \mathbf{g}_0$, as shown in the following example.

Example 1. *Let $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ be an orthogonal basis of $\mathbb{R}^3$. Take $\mathcal{F} = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_2)$ and $\mathcal{G} = \mathrm{span}(\mathbf{e}_1, \mathbf{e}_3)$. Then it follows that $\mathbf{P}_{\mathcal{F}} = \mathbf{e}_1\mathbf{e}_1^\top + \mathbf{e}_2\mathbf{e}_2^\top$ and $\mathbf{P}_{\mathcal{G}} = \mathbf{e}_2\mathbf{e}_2^\top + \mathbf{e}_3\mathbf{e}_3^\top$. In this case, $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$ commute, so $\mathcal{M} = \mathcal{N} = \mathcal{F} \cap \mathcal{G} = \mathrm{span}(\mathbf{e}_1)$.*

Also, we have the following properties:

Lemma 2. *Given the orthogonal projectors $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$, and the orthogonal projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ onto, respectively, the spaces $\mathcal{M}$ and $\mathcal{N}$, defined as in Equation (7), the following holds: (i) $\dim(\mathcal{M}) = \dim(\mathcal{N}) = \dim(\mathrm{SIm}(\mathbf{f})) - \dim(\mathrm{SIm}(\mathbf{f}) - \mathrm{SIm}(\mathbf{g}_0)^\perp)$; (ii) $\mathcal{M} \subseteq \mathrm{SIm}(\mathbf{f})$ and $\mathcal{N} \subseteq \mathrm{SIm}(\mathbf{g}_0)$; and (iii)*

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = \big(\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x})\big)^\top \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y)\,. \tag{8}$$

All proofs can be found in Appendix B. As a consequence of this lemma, we can view the projections $\mathbf{P}_{\mathcal{M}}\mathbf{f}$ and $\mathbf{P}_{\mathcal{N}}\mathbf{g}$ as the parts of $\mathbf{f}$ and $\mathbf{g}$ that are effectively retained when evaluating the dot product on the left-hand side of Equation (8). This also means the dot product depends solely on the projection of the embeddings onto $\mathcal{M}$ and the unembeddings onto $\mathcal{N}$; components of the embeddings and unembeddings which are orthogonal to these subspaces do not contribute. The dimensionality of $\mathcal{M}$ (and $\mathcal{N}$) can be viewed as a measure of model complexity since, intuitively, the larger $\dim(\mathcal{M})$ is, the more expressive the resulting model, and thus the more complex the relationship between y and $\mathbf{x}$, captured by $p(y \mid \mathbf{x})$, can be. For this reason, we call $\dim(\mathcal{M})$ the *effective complexity* of the model $(\mathbf{f}, \mathbf{g})$. Note that the effective complexity of $(\mathbf{f}, \mathbf{g})$ is less than or equal to the representation dimensionality d.

3.2 Extended linear equivalence relation

We now introduce an equivalence relation among models with potentially different representation dimensionality, building on our notion of effective complexity. We consider next-token prediction models $(\mathbf{f}, \mathbf{g})$ and $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, and spaces $\tilde{\mathcal{F}} := \mathrm{SIm}(\tilde{\mathbf{f}})$, $\tilde{\mathcal{G}} := \mathrm{SIm}(\tilde{\mathbf{g}}_0)$, $\tilde{\mathcal{M}} := \mathrm{Im}(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}})$ and $\tilde{\mathcal{N}} := \ker(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}})^\perp$, as introduced in Section 3.1.

Definition 3 (Extended linear equivalence)**.** *Two models* $(\mathbf{f}, \mathbf{g})$ *and* $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ *are extended-linearly equivalent, if both (i)* $\dim(\mathcal{M}) = \dim(\tilde{\mathcal{M}})$ *and (ii) there exist two full-rank matrices* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d \times \tilde{d}}$ *defining, respectively, invertible transformations from* $\mathcal{M}$ *to* $\tilde{\mathcal{M}}$*, and from* $\mathcal{N}$ *to* $\tilde{\mathcal{N}}$*, such that* $\mathbf{M}^\top \mathbf{N} = \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{N}}}$ *and*

$$\mathbf{P}_{\mathcal{M}} \mathbf{f}(\mathbf{x}) = \mathbf{M} \mathbf{P}_{\tilde{\mathcal{M}}} \tilde{\mathbf{f}}(\mathbf{x}) \tag{9}$$

$$\mathbf{P}_{\mathcal{N}} \mathbf{g}_0(y) = \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y)\,, \tag{10}$$

for all $y \in \mathcal{A}, \mathbf{x} \in \mathsf{Seq}(\mathcal{A})$*. We denote this relation by* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*.*

The above equivalence relation generalizes the linear equivalence already known in the literature [Roeder et al., 2021, Khemakhem et al., 2020a] to that of two models which can be related to each other on subspaces of dimension $\dim(\mathcal{M}) \leq \min\{d, \tilde{d}\}$. It shows that, after projecting the representations to suitable equal-dimensional subspaces, namely $\mathcal{M}, \mathcal{N}, \tilde{\mathcal{M}}$, and $\tilde{\mathcal{N}}$, we can find an invertible linear transformation relating them. Here, the dimensions of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ are equal, requiring that two equivalent models share the same *effective complexity*. Furthermore, models that are $\sim_{EL}$-equivalent encode the same dot-product:

Proposition 4. *If* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*, then*

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}_0(y)\,. \tag{11}$$

As a consequence, models in the $\sim_{EL}$ equivalence class also satisfy $p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}$. In Appendix B.3, we prove that Definition 3 is an equivalence relation and we provide the explicit form of the matrix $\mathbf{N}$. The extended linear equivalence relation is illustrated in Figure 2.

3.3 Identifiability of next-token predictors

The following theorem provides a characterization of models generating the same conditional probability distribution (i.e., distributionally-equivalent next-token predictors):

Theorem 5. *For all* $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$*, with representation dimensions* d *and* $\tilde{d}$ *(not necessarily equal),*

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \iff (\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})\,. \tag{12}$$

In words, there is a one-to-one correspondence between the set of conditional probability distributions expressed in Equation (1) and the set of equivalence classes entailed by $\sim_{EL}$ (cf. Figure 1): models $(\mathbf{f}, \mathbf{g})$, $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ which are $\sim_{EL}$-equivalent can be mapped to a single conditional probability distribution $p_{\mathbf{f},\mathbf{g}}$.

Interestingly, Theorem 5 highlights the fact that our notion of effective complexity, defined in Section 3.1 for next-token prediction models, is a well-defined complexity measure for a distribution $p_{\mathbf{f},\mathbf{g}}$, in the sense that it does not depend on the specific choice of embedding and unembedding functions $(\mathbf{f}, \mathbf{g})$. Indeed, the result implies that

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \implies \dim(\mathcal{M}) = \dim(\tilde{\mathcal{M}})\,. \tag{13}$$

Furthermore, as a special case of Theorem 5, when the diversity condition holds and $d = \tilde{d}$ we recover known results on linear identifiability:

Corollary 6 (Adapted from [Roeder et al., 2021])**.** *For all* $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta_d$ *such that* $(\mathbf{f}, \mathbf{g})$ *satisfies the diversity condition (Definition 1), we have*

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \implies (\mathbf{f}, \mathbf{g}) \sim_{L} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})\,, \tag{14}$$

where, by definition, $(\mathbf{f}, \mathbf{g}) \sim_{L} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ *if and only if there exists an invertible matrix* $\mathbf{M} \in \mathbb{R}^{d \times d}$ *such that for all* $y \in \mathcal{A}, \mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ *we have*

$$\mathbf{f}(\mathbf{x}) = \mathbf{M} \tilde{\mathbf{f}}(\mathbf{x}) \text{ and } \mathbf{g}_0(y) = \mathbf{M}^{-\top} \tilde{\mathbf{g}}_0(y)\,. \tag{15}$$

In Appendix B.5, we provide an example about nonlinear distortions that can arise in models that are $\sim_{EL}$-equivalent but not diverse (Definition 1).

3.4 Implications for empirical practice

Implications for trained models. Suppose that $(\mathbf{f}, \mathbf{g}) \in \Theta_d$ and $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta_{\tilde{d}}$ are both global maximizers (in their respective model classes Θ_d and $\Theta_{\tilde{d}}$) of the objective in Equation (3). If both models have enough capacity to represent the ground-truth data distribution $p_{\mathcal{D}}$, then they necessarily represent the same distribution, i.e., $p_{\mathbf{f},\mathbf{g}} = p_{\mathcal{D}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}$. By Theorem 5, we can thus conclude that $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$. If these assumptions held in practice, this would imply that all models trained with sufficient capacity on a given data distribution $p_{\mathcal{D}}$ will be $\sim_{EL}$-equivalent.[2] This analysis relies crucially on the assumption that Θ has enough capacity to represent $p_{\mathcal{D}}$. In practice, this might not hold for at least two reasons. First, we typically train with a fixed representation dimension d which limits the expressivity of the model; moreover, despite the universal approximation guarantee, there might not even exist a sufficiently large d such that the model can express $p_{\mathcal{D}}$ *exactly*. Secondly, all distributions that can be represented by Θ put nonzero probability mass on all text sequences (because of the exponential on the RHS of Equation (1)), whereas under the ground-truth distribution $p_{\mathcal{D}}$ describing, e.g., text on the internet, several sequences will have zero

[2] We neglect optimiziation issues such as local minima for ease of exposition.

probability. Models of the form in Equation (1) are thus inherently misspecified in such cases. Modeling these settings might thus require an extension of the current theoretical framework [Nielsen et al., 2024].

Different token vocabularies. Our analysis is also restricted to next-token predictors that share the same token vocabulary $\mathcal{A}$. We hypothesize that, for two models with different token vocabularies $\mathcal{A}$ and $\tilde{\mathcal{A}}$, our results may be extended to prove a $\sim_{EL}$-equivalence relation restricted to the shared tokens $\mathcal{A} \cap \tilde{\mathcal{A}}$, under suitable conditions on the next-token probabilities.

4 Linear properties

In the previous section, we established identifiability results for next-token predictors. Here, we turn to precisely defining the linear properties we will focus on. These characterize how a given model $(\mathbf{f}, \mathbf{g})$ represents different inputs—as in our opening example, describing a geometric relationship (parallelism) among the vector differences between the embeddings of two different inputs (*"lucky"* and *"luckiest"*) and that between two further inputs (*"easy"* and *"easiest"*). Importantly, these linear properties are not to be confused with the linear equivalence class $\sim_L$ in Definition 3, which instead describes how *different* models represent the *same* data distribution. In Section 5, we will combine the linear properties defined here with the identifiability results of Section 3 to determine which linear properties hold for all models in a given equivalence class. See also Figure 1 for an illustration.

Our analysis focuses on embeddings $\mathbf{f}(\mathbf{s}) \in \mathrm{SIm}(\mathbf{f})$ and unembeddings $\mathbf{g}_0(y) \in \mathrm{SIm}(\mathbf{g}_0)$, allowing us to define relational linear properties solely in terms of the quantities described in our identifiability result: our analysis is thus agnostic to assumptions on the data-generating process underlying natural language and it does not require positing unobserved variables. In principle, the linear properties we will define can apply to any collection of strings—for example, the difference between the unembeddings of *"1fv0sywi"* and *"eg2op3te"* could be parallel to the difference between those of *"tgsqil2h"* and *"khdo5zof"*. As we will argue, such parallel structures imply certain symmetries in a model's conditional next-token probabilities. Our work is motivated by the commonly observed instances of linear properties involving collections of semantically meaningful strings, where symmetries in next-token probabilities likely reflect regularities in human-produced text.

4.1 Parallel vectors

We begin with a definition of vector parallelism. This is motivated by recent empirical findings that differences in semantically or syntactically related token unembeddings often exhibit parallelism, such as $\mathbf{g}($*"easy"*$) - \mathbf{g}($*"easiest"*$)$ being parallel to $\mathbf{g}($*"lucky"*$) - \mathbf{g}($*"luckiest"*$)$.[3] Central to our theory will be the following definition of parallelism in a subspace $\Gamma \subseteq \mathbb{R}^d$:

Definition 7 (Parallelism in Γ)**.** *We say that two vectors $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathbb{R}^d$ are parallel in Γ if there exists $\beta \neq 0$ (see Remark 20) such that $\mathbf{P}_\Gamma \boldsymbol{\gamma} = \beta \cdot \mathbf{P}_\Gamma \boldsymbol{\gamma}'$.*

We next show that parallel vectors induce similar log ratios of conditional probabilities, as noted in [Park et al., 2024a, Jiang et al., 2024]:

Lemma 8. *Consider a model $(\mathbf{f}, \mathbf{g}) \in \Theta$. For $y_0, y_1, y_2, y_3 \in \mathcal{A}$, the difference vectors $\mathbf{g}(y_1) - \mathbf{g}(y_0)$ and $\mathbf{g}(y_3) - \mathbf{g}(y_2)$ are parallel in $\mathcal{N}$ if and only if there exists $\beta \neq 0$, s.t. $\forall \mathbf{s} \in \mathsf{Seq}(\mathcal{A})$*

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s})} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(y_2 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_3 \mid \mathbf{s})}\,. \tag{16}$$

That is, parallel difference vectors for token pairs correspond to proportional likelihood ratios between the tokens in each token pair. Notice that, as in Definition 7, the difference vectors are parallel only in the space $\mathcal{N}$. This implies that the components outside $\mathcal{N}$ for two $\mathcal{N}$-parallel vectors are allowed to not be parallel. All proofs for the results presented in this section are provided in Appendix C.

4.2 Relational linear property

Beyond parallelism, the first property we define is relational linearity, introduced by Paccanaro and Hinton [2001] and recently studied by Hernandez et al. [2024], who found empirical evidence that hidden transformer layers in language models display this property.

Context-query-reply sequences. We consider sequences $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ that can be decomposed as $\mathbf{x} = \mathbf{s} \frown \mathbf{q} \frown y$, where $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$ is termed *context* (or *subject*), $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ is termed *query* (or *relation*), and $y \in \mathcal{A}$ is termed *reply* (or *object*). The following example illustrates a semantically meaningful context-query-reply sequence.

Example 2. *Consider a sequence $\mathbf{x} = \mathbf{s} \frown \mathbf{q} \frown y$ where $\mathbf{s} =$ "All roads lead to Rome", $\mathbf{q} =$ "What is the written language?", and $y =$ "English". We deliberately pick $y =$ "English" as the most likely next-token prediction following $\mathbf{s} \frown \mathbf{q}$ made by English speakers. The string $\mathbf{s} \frown \mathbf{q}$ could also be provided as input to a language model to test whether it can recognize English language. Another example of context-query-reply sequence is $\mathbf{s} =$ "Rome", $\mathbf{q} =$ "is the capital of" and $y =$ "Italy".*

[3] Similar properties had previously been observed in word embedding models [Park et al., 2024a, Mikolov, 2013].

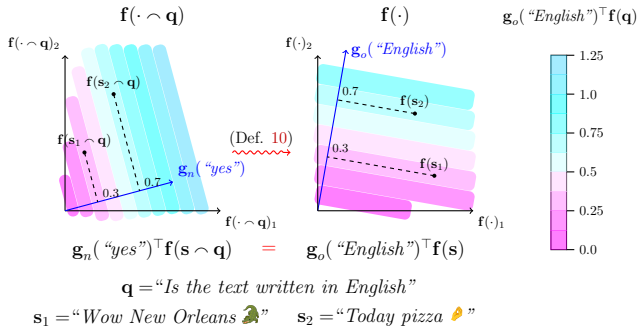

Figure 3: **Relational linear subspaces**. The figure depicts the embedding function $\mathbf{f}$ of a model $(\mathbf{f}, \mathbf{g}) \in \Theta$ with representation dimension $d = 2$. Let $\mathbf{g}_o(\text{“English”}) := \mathbf{g}(\text{“English”}) - \mathbf{g}(\text{“other language”})$ and $\mathbf{g}_n(\text{“yes”}) := \mathbf{g}(\text{“yes”}) - \mathbf{g}(\text{“no”})$. Here, $(\mathbf{f}, \mathbf{g})$ linearly represents (Definition 10) the subspace spanned by $\mathbf{g}_o(\text{“English”})$ for the query $\mathbf{q}$ =“*Is the text written in English?*”. Accordingly, there exists a vector, here $\mathbf{g}_n(\text{“yes”})$, such that the dot product $\mathbf{g}_o(\text{“English”})^\top \mathbf{f}(\mathbf{s})$, whose magnitude is represented through the color map on the right, matches the dot product $\mathbf{g}_n(\text{“yes”})^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})$, on the left. For ease of visualization, we set $\mathbf{g}_n(\text{“yes”})^\top \mathbf{a}_\mathbf{q} = 0$ and we display the values of the dot products for two input contexts $\mathbf{s}_1, \mathbf{s}_2$. Intuitively, the dot product of a context's embedding $\mathbf{f}(\mathbf{s})$ with $\mathbf{g}_o(\text{“English”})$ captures the log-probability ratio of “*yes*” vs. “*no*” as next tokens for the same context $\mathbf{s}$ concatenated with the query $\mathbf{q}$.

As shown in the examples above, it is often possible to parse natural language expressions into strings $\mathbf{x} = \mathbf{s} \frown \mathbf{q} \frown y$ which capture relational aspects encoded in substrings $\mathbf{s}$ and y through a substring $\mathbf{q}$. In principle, one could also consider strings $\mathbf{s} \frown \mathbf{q}$ involving queries whose expected reply is independent of the context, such as $\mathbf{q}$ =“*Whatever follows reply with* 42”, or paraphrases of the query, *e.g.*, $\mathbf{q}'$ =“*Now reply with* 42”. In Appendix E, we disscuss how our analysis can capture these corner cases. Intuitively, relational linearity entails the property that all the information relevant for next-token prediction carried by the embeddings of the joint string $\mathbf{s} \frown \mathbf{q}$ (*i.e.*, $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$) can be retrieved by considering the embeddings of $\mathbf{s}$ (*i.e.*, $\mathbf{f}(\mathbf{s})$) via an affine transformation. To formalize this, we focus on the embeddings $\mathbf{f}(\mathbf{s})$ of the model and on subspaces $\Gamma \subseteq \text{SIm}(\mathbf{g}_0)$ of the unembeddings, which contain the relevant tokens for $\mathbf{q}$.[4]

Definition 9 (ΓLR: Relational linearity of $\mathbf{q}$ in Γ)**.** *For a model $(\mathbf{f}, \mathbf{g}) \in \Theta$, let $\Gamma \subseteq \text{SIm}(\mathbf{g}_0)$ be a subspace. We say that $(\mathbf{f}, \mathbf{g})$ linearly represents the query $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ on Γ, if there exist a matrix $\mathbf{A}_\mathbf{q} \in \mathbb{R}^{d \times d}$ and a vector $\mathbf{a}_\mathbf{q} \in \mathbb{R}^d$ such that, for all $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$,*

$$\mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_\Gamma \big(\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q}\big). \tag{17}$$

When this holds, we define $\Gamma_\mathbf{q} := \text{Im}(\mathbf{A}_\mathbf{q}^\top \mathbf{P}_\Gamma)$.

Intuitively, all the information within $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$ which is relevant to compute the probability of next-tokens in Γ is captured, up to an affine transformation, by $\mathbf{f}(\mathbf{s})$ in the subspace $\Gamma_\mathbf{q}$. Indeed, one can show that, if $\mathbf{g}_0(\mathbf{y}) \in \Gamma$, then necessarily $\mathbf{f}(\mathbf{s} \frown \mathbf{q})^\top \mathbf{g}_0(y) = (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})^\top \mathbf{g}_0(y)$. The spaces Γ and $\Gamma_\mathbf{q}$ are central to proving whether relational linearity holds for all $\sim_{EL}$-equivalent models, as we will show in Section 5.

Connection to other linear properties. In the following, we show how to capture three additional linear properties building on the definition of relational linearity. We follow the taxonomy by Park et al. [2024a].

4.2.1 Linear subspaces (LS)

Parallel vectors naturally define a one-dimensional subspace $\Gamma \in \mathbb{R}^d$ that contains all of those vectors. In language model representations, several such subspaces have been identified that appear to encode semantic and syntactic properties, for example translation across languages or the transformation of an adjective into its comparative or superlative form [Mikolov et al., 2013a, Park et al., 2024a]. Our relational formulation of this linear property focuses on whether these subspaces contain the information in the embeddings $\mathbf{f}(\mathbf{s})$ which is relevant to predict the reply token to a query $\mathbf{q}$ when appended to the context $\mathbf{s}$. E.g., this could happen if the embeddings projected in the subspace $\Gamma_{eng,ita}$, containing the vector $\mathbf{g}(\text{“Rome”}) - \mathbf{g}(\text{“Roma”})$, encode whether the replies to the query $\mathbf{q}$ = *“Is written in English or Italian?”* are more likely to be y = *“English”* or y' = *“Italian”*. We can capture this through the following definition:

Definition 10 (LS: Relational Linear Subspaces)**.** *A model $(\mathbf{f}, \mathbf{g}) \in \Theta$ linearly represents a subspace $\Gamma \subseteq \text{SIm}(\mathbf{g}_0)$ relative to $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ if for all pairs of tokens $y_i, y_j \in \mathcal{A}$ such that $\mathbf{g}_i(y_j) := \mathbf{g}(y_j) - \mathbf{g}(y_i) \in \Gamma$, there exists a vector $\boldsymbol{\gamma} \in \text{SIm}(\mathbf{g}_0)$ such that $\forall \mathbf{s} \in \mathsf{Seq}(\mathcal{A})$*

$$\mathbf{g}_i(y_j)^\top \mathbf{f}(\mathbf{s}) = \boldsymbol{\gamma}^\top (\mathbf{f}(\mathbf{s} \frown \mathbf{q}) - \mathbf{a}_\mathbf{q}). \tag{18}$$

We provide one example of this property in Figure 3. The LS property is implied by relational linearity (Definition 9) in the following sense:

Proposition 11 (ΓLR $\implies$ LS)**.** *Suppose that a model $(\mathbf{f}, \mathbf{g}) \in \Theta$ (i) linearly represents $\mathbf{q}$ on $\Gamma \subseteq \text{SIm}(\mathbf{g}_0)$, and (ii) $\Gamma_\mathbf{q} \subseteq \text{SIm}(\mathbf{g}_0)$, then the model $(\mathbf{f}, \mathbf{g})$ linearly represents $\Gamma_\mathbf{q}$ relative to $\mathbf{q}$ (Definition 10).*

[4] E.g., for a query $\mathbf{q}$ = *“What is the written language?”*, next-tokens corresponding to different languages may be more probable and interesting to look at, thought it's ultimately a modeler's choice what subspace Γ to focus on.

4.2.2 Linear probing (LP)

There is empirical evidence that, in language models, sentence embeddings can be linearly separated with good accuracy based on the language of the corresponding sentences [Park et al., 2024a, Chang et al., 2022].This property is also termed *linear probing* [Alain, 2016, Kim et al., 2018]. Below, we redefine linear probing as a relational property, based on Definition 9:

Definition 12 (LP: Relational Linear Probing)**.** *We say that a model* $(\mathbf{f}, \mathbf{g}) \in \Theta$ *can be linearly probed for a query* $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ *and a collection* $\mathcal{Y}_P \subseteq \mathcal{A}$ *of* ℓ *elements if there exist* $\mathbf{W} \in \mathbb{R}^{\ell \times d}$ *and* $\mathbf{b} \in \mathbb{R}^{\ell}$ *such that for all* $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$ *and* $\forall i \in [\ell]$

$$\operatorname{softmax}\big(\mathbf{W}\mathbf{f}(\mathbf{s}) + \mathbf{b}\big)_i = p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P), \quad (19)$$

where $p(y \mid \cdot\,; \mathcal{Y}_P) = p(y \mid \cdot)/(\sum_{y' \in \mathcal{Y}_P} p(y' \mid \cdot))$ *is the conditional distribution restricted to the set* $\mathcal{Y}_P$*.*

To illustrate why this is termed *linear probing*, suppose a model given the query $\mathbf{q}$ ="*Is the text written in English?*" discriminates input sequences $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$ between positive y_0 ="*yes*" and negative examples y_1 ="*no*"—that is, it assigns high probability $p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s} \frown \mathbf{q})$ to sequences $\mathbf{s}$ corresponding to English sentences, and high probability $p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s} \frown \mathbf{q})$ to non-English sentences. Then, these conditional distributions can be evaluated directly from $\mathbf{f}(\mathbf{s})$ via a linear probe. Figure 1 includes an illustration of LP. Below, we relate LP (Definition 12) to ΓLR (Definition 9):

Proposition 13 (ΓLR $\implies$ LP)**.** *If a model* $(\mathbf{f}, \mathbf{g}) \in \Theta$ *(i) linearly represents* $\mathbf{q}$ *on* Γ*, and (ii)* $\mathbf{g}(y_i) - \mathbf{g}(y_j) \in \Gamma$ *for all* $y_i \in \mathcal{Y}_P$*, then the model can be linear probed (Definition 12) for* $\mathbf{q}$ *and* $\mathcal{Y}_P$*, with parameters given by* $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_\ell)^\top$ *and* $\mathbf{b} = (b_1, \ldots, b_\ell)^\top$*, where* $\mathbf{w}_i := \mathbf{A}_{\mathbf{q}}^\top \mathbf{g}(y_i)$ *and* $b_i := (\mathbf{a}_{\mathbf{q}})^\top \mathbf{g}(y_i)$*.*

4.2.3 Linear Steering

Another property that has attracted considerable attention is the *linear steering property* [Stolfo et al., 2024], also termed *linear intervening property* by Park et al. [2024a]. By knowing what queries are linearly represented by the model (as per Definition 9), this property allows us to *steer* the model embeddings such that the most-likely reply to a given query changes, while the replies to other queries remain unaffected. In Appendix C.1, we define a relational version of this property, and show under what conditions it is implied by the relational linear property (Proposition 19).

5 Linear properties shared by all distribution-equivalent models

Based on Theorem 5, we can now analyze which linear properties are shared across models expressing the same next-token distribution. We start from relational linearity (Definition 9). To this end, pick a model $(\mathbf{f}, \mathbf{g})$ that linearly represents $\mathbf{q}$ on Γ, and consider the space $\Gamma_{\mathbf{q}} := \mathrm{Im}(\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)$. We show that under an additional condition on Γ and $\Gamma_{\mathbf{q}}$, two models that are $\sim_{EL}$-equivalent share the same linear properties (results from this section are proved in Appendix D):

Theorem 14. *For two models* $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$ *s.t.* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*, if* $\mathbf{f}$ *linearly represents* $\mathbf{q}$ *on* $\Gamma \subseteq \mathcal{N}$*, and* $\Gamma_{\mathbf{q}} \subseteq \mathcal{M}$*, then* $\tilde{\mathbf{f}}$ *linearly represents* $\mathbf{q}$ *on* $\tilde{\Gamma} \subseteq \tilde{\mathcal{N}}$*, where* $\tilde{\Gamma} = \mathrm{Im}(\mathbf{N}^+ \mathbf{P}_\Gamma)$ *and* $\mathbf{N}$ *is the matrix relating* $\mathbf{g}_0$ *and* $\tilde{\mathbf{g}}_0$ *by the equivalence relation in Definition 3.*

This shows how, under the the condition that $\Gamma \subseteq \mathcal{N}$ and $\Gamma_{\mathbf{q}} \subseteq \mathcal{M}$, relational linearity (ΓLR) *is a property of all or none next-token predictors modeling the same conditional distribution.* As a consequence, the same holds for LS (by Proposition 11) and LP (by Proposition 13). Intuitively, the extra condition underlies that relational linearity of $(\mathbf{f}, \mathbf{g})$ is displayed by the components of $\mathbf{f}$ that contribute to the dot product with $\mathbf{g}_0$. A $\sim_{EL}$-equivalent model $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ would linearly transform these components, thus preserving relational linearity. Vice versa, since all components of $\mathbf{f}$ outside $\mathcal{M}$ can be arbitrarily distorted, any property of $(\mathbf{f}, \mathbf{g})$ that depends on those components may not hold for $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$. The extra condition precisely avoids that.[5] Notice that the special case where the diversity condition (Definition 1) holds implies a similar conclusion because the condition that $\Gamma \subseteq \mathcal{N}$ and $\Gamma_{\mathbf{q}} \subseteq \mathcal{M}$ is then always satisfied (as $\mathcal{M} = \mathcal{N} = \mathbb{R}^d$). This testifies that (a special case of) relational linearity is shared among $\sim_{EL}$ models. In contrast, vector parallelism may not be preserved: Two parallel vectors in one model $(\mathbf{f}, \mathbf{g})$ may not be parallel in another model $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ with the same conditional distribution. They remain parallel only within the subspaces $\mathcal{N}$ and $\tilde{\mathcal{N}}$, respectively:

Theorem 15. *For two models* $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$*, such that* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*, the vectors* $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathrm{SIm}(\mathbf{g}_0)$ *are parallel within* $\mathcal{N}$ *if and only if the corresponding vectors* $\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\gamma}}' \in \mathrm{SIm}(\tilde{\mathbf{g}}_0)$ *are parallel in* $\tilde{\mathcal{N}}$*.*

6 Discussion

Theorem 14 is an example of a property that all distribution-equivalent next-token predictors, as characterized by our identifiability result (Theorem 5), must share. One may then ask whether the widely observed linear properties of language models are indeed examples of shared properties akin to the one in Theorem 14. Tautologically, claims about the ubiquity of linear properties cannot solely be based on observations of linearity

[5] If $\Gamma \not\subseteq \mathcal{N}$, then relational linearity (Definition 9) would be trivially satisfied for $(\mathbf{f}, \mathbf{g})$ and, in turn, also for $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, because $\boldsymbol{\gamma} \in \Gamma$ would give $\boldsymbol{\gamma}^\top \mathbf{f}(\mathbf{x}) = 0$, and so $\mathbf{P}_\Gamma \mathbf{f}(\mathbf{x}) = \mathbf{0}$.

in individual model instances. One might thus hypothesize that **only those properties shared across all equivalent models should be ubiquitously observable**.[6] In the following, we critically examine this hypothesis in light of empirical evidence.

Do we observe properties which are not shared across all $\sim_{EL}$-equivalent models? Theorem 15 shows that vector parallelism is only preserved within a linear subspace $\mathcal{N}$ of the unembedding space. Surprisingly, a different kind of parallelism, which according to our theory is not shared by all $\sim_{EL}$-equivalent models, appears to be consistently observed in language models. In fact, several empirical studies apply dimensionality reduction through PCA to the embeddings and unembeddings to reveal and visualize linear properties including parallelism (*e.g.*, [Mikolov et al., 2013a, Figure 2]; [Marks and Tegmark, 2023, Figure 1]). Note that, across $\sim_{EL}$-equivalent models, the embeddings and unembeddings may not be completely contained within $\mathcal{M}$ and $\mathcal{N}$, respectively: that is the case when both $\mathcal{M} \subsetneq \mathrm{SIm}(\mathbf{f})$ and $\mathcal{N} \subsetneq \mathrm{SIm}(\mathbf{g}_0)$. As a consequence, unbounded distortions within the orthogonal complements of $\mathcal{M}$ and of $\mathcal{N}$ are inconsequential for the dot product between the embedding and the unembedding vectors (see, *e.g.*, Figure 2, top left, for an embedding manifold not contained within any proper linear subspace). If these distortions were sufficiently large, they would prevent the visualization of vectors parallel in the sense of Theorem 15 through PCA, as the distortions would dominate the covariance matrix on which the PCA of the representations is performed, and thus the first principal components would mostly reflect those. This suggests that, in models where PCA reveals parallelism, these distortions are small, and the representations live close to a proper linear subspace.

How can we explain this? These observations suggest that something other than the assumptions of our Theorem 5 determines what models are learned in practice. One possible explanation is that some additional assumptions and constraints are at play which imply that only models in a subset of the $\sim_{EL}$ equivalence class are observed empirically. For parallelism, this could occur, for example, if the modeler chooses a fixed d for which the diversity condition happens to hold (Definition 1): in which case, the resulting equivalence class would be $\sim_L$, and parallelism in $\mathbb{R}^d$ (Definition 7) is a shared property across $\sim_L$-equivalent models with representation dimensionality d. An alternative possibility is that other inductive biases, not captured by the identifiability result, are influencing the learned representations. These biases could stem from the training algorithm or architecture, steering the model toward a subset of the $\sim_{EL}$-equivalent models. Our contribution is to provide a mathematical framework that enables a clear articulation of these questions, guiding future empirical investigation.

[6] An analogy could be made with the principle of covariance in physics [Einstein, 1920, Thorne et al., 2000], which asserts that physical laws should be expressible as coordinate-independent and reference-frame-independent geometric relationships between objects that represent physical entities [Thorne and Blandford, 2017]. Recently, Villar et al. [2023] suggested that this principle could inspire future developments in machine learning.

7 Related work and future directions

Linear properties of next-token predictors have attracted widespread attention, also beyond language modeling [Li et al., 2022, Nanda et al., 2023, Elhage et al., 2022]. More complex, non-linear properties have also been observed, such as circular token representations [Engels et al., 2024]. Formalizing these properties and investigating whether all distribution-equivalent next-token predictors share them, in the sense we studied for linear properties, is an interesting open venue.

Theoretical studies on linear properties. Park et al. [2024a] introduce binary latent concepts to describe several linear properties (though not relational linearity) in a unified framework. This was also applied to study categorical and hierarchical concepts [Park et al., 2024b]. Jiang et al. [2024] explain linear properties of language models based on assumptions on the data-generating process and latent variables underlying natural text. This allows them to reason about the *origins* of linearity; in this work, we instead focus on the *ubiquity* of linear properties, with an agnostic stance on latent concept variables. An exciting direction for future work is to prove, within our framework, why and how linear properties emerge, if they do at all.

Identifiability of representations is a central theme in generative modeling [Moran et al., 2022, Xi and Bloem-Reddy, 2023], particularly in non-linear ICA [Hyvarinen et al., 2019, Gresele et al., 2020, Hälvä and Hyvarinen, 2020, Buchholz et al., 2022, Hyttinen et al., 2022] and causal representation learning [Lippe et al., 2022, Ahuja et al., 2023, Liang et al., 2024, von Kügelgen et al., 2024, Varici et al., 2024, Zhang et al., 2024a, Rajendran et al., 2024, Li et al., 2024, Bortolotti et al., 2025]. Buchholz [2024] studied when token partitions can be identified from their interactions; Reizinger et al. [2024] discussed what role identifiability may play in explaining several aspects of large language models [Zhang et al., 2024b]. Our work highlights the role of identifiability in explaining the ubiquity of linear properties in language models.

Acknowledgments

We thank Beatrix Miranda Nielsen, Antonio Vergari, Frederik Hytting Jørgensen, Filippo Camilloni, Adrián Javaloy, and Julius von Kügelgen for useful discussions. We acknowledge positive feedback by Stefano Teso and interesting conversations with the participants of the 2024 Bellairs Workshop on Causality. E.M. acknowledges support from TANGO, Grant Agreement No. 101120763. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. L.G. was supported by Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429).

References

Alberto Paccanaro and Geoffrey E. Hinton. Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):232–244, 2001.

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. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning (ICML)*, pages 9030–9039. PMLR, 2021.

David E. Rumelhart and Adele A. Abrahamson. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, 1973.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 77–109. 1986a.

Geoffrey E. Hinton et al. Learning distributed representations of concepts. In *Proceedings of the eighth Annual Conference of the Cognitive Science Society*, volume 1, page 12. Amherst, MA, 1986b.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in Neural Information Processing Systems (NeurIPS)*, 13, 2000.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 746–751, 2013b.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. *arXiv preprint arXiv:2305.16130*, 2023.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897*, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12768–12778, 2020a.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4: 385–399, 2016.

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning (ICML)*, pages 223–231. PMLR, 2019.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR, 21–27 Jul 2024a.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. *arXiv preprint arXiv:2403.10381*, 2024.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2207–2217. PMLR, 2020b.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning (ICML)*, pages 18171–18206. PMLR, 2023.

W Rehder. When do projections commute? *Zeitschrift für Naturforschung A*, 35(4):437–441, 1980.

Beatrix Miranda Ginn Nielsen, Luigi Gresele, and Andrea Dittadi. Challenges in explaining representational similarity through identifiability. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024.

Tyler A. Chang, Zhuowen Tu, and Benjamin K Bergen. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*, 2022.

Guillaume Alain. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning (ICML)*, pages 2668–2677. PMLR, 2018.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. *arXiv preprint arXiv:2410.12877*, 2024.

Albert Einstein. Fundamental ideas and methods of the theory of relativity, presented in their development. *The collected papers of Albert Einstein*, 7:1918–1921, 1920.

Kip S. Thorne, Charles W. Misner, and John Archibald Wheeler. *Gravitation*. Freeman San Francisco, 2000.

Kip S. Thorne and Roger D. Blandford. *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton University Press, 2017.

Soledad Villar, David W. Hogg, Weichi Yao, George A. Kevrekidis, and Bernhard Schölkopf. Towards fully covariant machine learning. *Transactions on Machine Learning Research*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2022.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024b.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 6912–6939. PMLR, 2023.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 859–868. PMLR, 2019.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence (UAI)*, pages 217–227. PMLR, 2020.

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 939–948. PMLR, 2020.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 16946–16961, 2022.

Antti Hyttinen, Vitória Barin Pacela, and Aapo Hyvärinen. Binary independent component analysis: a non-stationarity-based approach. In *Uncertainty in Artificial Intelligence (UAI)*, pages 874–884. PMLR, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning (ICML)*, pages 13557–13603. PMLR, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning (ICML)*, pages 372–407. PMLR, 2023.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2314–2322. PMLR, 2024.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024a.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled representation learning in non-markovian causal systems. 2024.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. *arXiv preprint arXiv:2502.11245*, 2025.

Simon Buchholz. Learning partitions from context. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Patrik Reizinger, Szilvia Ujváry, Anna Mészáros, Anna Kerekes, Wieland Brendel, and Ferenc Huszár. Position: Understanding LLMs Requires More Than Statistical Generalization. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024b.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. 2021. URL https://github.com/kingoflolz/mesh-transformer-jax.

Sheldon Axler. *Linear Algebra Done Right*. Springer, 2015.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/**Not Applicable**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/**Not Applicable**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/**Not Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

Appendices

Table of Contents

A Functional Form of Decoders-only Transformers

Decoders-only transformer models can be reduced to the form we use in the main text, as already shown in a derivation by Roeder et al. [2021]. We consider autoregressive, GPT-like models [Liu et al., 2018, Radford et al., 2021], focusing on the GPT-J model [Wang and Komatsuzaki, 2021].

We denote with $\mathbf{h}(\mathbf{x};\theta)$ the representation given by the a transformer model $\mathbf{h}$ with trainable parameters $\theta \in \mathbb{R}^q$ for the architecture. We consider the transformer to represent sentences of lenght up to C, and by convention we pick the last C tokens if the sentence has length $t(\mathbf{x}) > C$. Denote with $\tau := \max(t(\mathbf{x}) - C, 1)$ it holds that:

$$\mathbf{h}(\mathbf{x};\theta) = \mathbf{h}(\mathbf{x}_{t(\mathbf{x})-\tau:t(\mathbf{x})};\theta) \in \mathbb{R}^{d\times\tau}\,. \tag{20}$$

We focus on the representation of the last token of sentence, which is used to perform next-token prediction, and denote it with

$$\mathbf{h}(\mathbf{x};\theta)_{-1} \in \mathbb{R}^{d}\,. \tag{21}$$

When predicting the next token among $K := |\mathcal{A}|$ possible options, a common approach is to use a layer (or *head*) with weights $\mathbf{W}_D \in \mathbb{R}^{K\times d}$ and a bias $\mathbf{b} \in \mathbb{R}^K$. The representation of $\mathbf{g}_\varphi$ is determined by the parameters $\varphi = (\mathbf{W}_D, \mathbf{b})$, which are used to compute the logits as follows:

$$\text{logits}(\mathbf{x}) := \mathbf{W}_D\mathbf{h}(\mathbf{x};\theta)_{-1} + \mathbf{b}\,. \tag{22}$$

By extending the representation of $\mathbf{h}_{-1}$ to $\mathbf{f}_\theta(\mathbf{x}) := (1, \mathbf{h}(\mathbf{x};\theta)_{-1}^\top)^\top$, we can incorporate the bias by adding one column to $\mathbf{W}_D$, obtaining:

$$\text{logits}(\mathbf{x}) = \tilde{\mathbf{W}}_D\mathbf{f}_\theta(\mathbf{x}), \quad \text{where } \tilde{\mathbf{W}}_D = \big(\mathbf{b}, \mathbf{W}_D\big)\,. \tag{23}$$

To predict the probability of the next token y it is then sufficient to consider

$$\log p_{\mathbf{f}_\theta,\mathbf{g}_\varphi}(y \mid \mathbf{x}) = \text{logits}(\mathbf{x})_y - \log \textstyle\sum_{y'\in\mathcal{A}} \text{logits}(\mathbf{x}_{y'})\,. \tag{24}$$

Transforming the token $y \in \mathcal{A}$ to its one-hot representation, $\mathbf{y} \in \{0,1\}^K$, we can write

$$\mathbf{g}_\varphi(y) = \tilde{\mathbf{W}}_D^\top\mathbf{y}, \tag{25}$$

thereby leading to the expression:

$$\log p_{\mathbf{f}_\theta,\mathbf{g}_\varphi}(y \mid \mathbf{x}) = \mathbf{g}_\varphi(y)^\top\mathbf{f}_\theta(\mathbf{x}) - \log \sum_{y'\in\mathcal{A}} \mathbf{g}_\varphi(y)^\top\mathbf{f}_\theta(\mathbf{x}), \tag{26}$$

where we used:

$$\mathbf{f}_\theta(\mathbf{x}) = \begin{pmatrix} 1 \\ \mathbf{h}(\mathbf{x}_{t(\mathbf{x})-\tau:t(\mathbf{x})};\theta)_{-1} \end{pmatrix}, \qquad \mathbf{g}_\varphi(y) = \tilde{\mathbf{W}}_D^\top\mathbf{y}. \tag{27}$$

This explicit form of the embedding and unembedding respectively and consistent with Equation (1), as previously detailed by Roeder et al. [2021].

B Proofs of Section 3

B.1 Reminder of useful properties of pseudo-inverses

We will often make use of the pseudo-inverse $\mathbf{A}^+$ of a matrix $\mathbf{A}$ [Axler, 2015, page 221]. We denote with $\mathbf{T}\,|_{\ker(\mathbf{T})^\perp}$ the restriction of $\mathbf{T}$ to its orthogonal complement of the kernel [Axler, 2015].

Definition 16 (Pseudo-inverse)**.** *Let $\mathbf{T} \in \mathbb{R}^{m\times n}$ be a matrix. The pseudo-inverse $\mathbf{T}^+ \in \mathbb{R}^{n\times m}$ of $\mathbf{T}$ is defined as the linear map:*

$$\mathbf{T}^+\mathbf{w} := \left(\mathbf{T}\,|_{\ker(\mathbf{T})^\perp}\right)^{-1}\mathbf{P}_{\mathrm{Im}(\mathbf{T})}\mathbf{w} \tag{28}$$

for all $\mathbf{w} \in \mathbb{R}^n$.

Accordingly, the pseudo-inverse always exists and it is unique. Notice that for any matrix $\mathbf{T} \in \mathbb{R}^{m\times n}$ it holds:

$$\mathbf{T}\mathbf{T}^+ = \mathbf{P}_{\mathrm{Im}(\mathbf{T})} \tag{29}$$
$$\mathbf{T}^+\mathbf{T} = \mathbf{P}_{\ker(\mathbf{T})^\perp} \tag{30}$$
$$\mathbf{T}^\top(\mathbf{T}^\top)^+ = \mathbf{P}_{\ker(\mathbf{T})^\perp} \tag{31}$$
$$(\mathbf{T}^\top)^+\mathbf{T}^\top = \mathbf{P}_{\mathrm{Im}(\mathbf{T})}\,. \tag{32}$$

B.2 Proof of Lemma 2

We provide here a longer version of the Lemma capturing different properties between the projectors.

Lemma (Ref Lemma 2)**.** *Let $(\mathbf{f}, \mathbf{g}) \in \Theta$, and take $\mathcal{F} := \mathrm{SIm}(\mathbf{f})$ and $\mathcal{G} := \mathrm{SIm}(\mathbf{g}_0)$. For the orthogonal projectors $\mathbf{P}_\mathcal{F}$ and $\mathbf{P}_\mathcal{G}$ and the orthogonal projectors $\mathbf{P}_\mathcal{M}$ and $\mathbf{P}_\mathcal{N}$ projecting on the spaces:*

$$\mathcal{M} = \mathrm{Im}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}), \quad \mathcal{N} = \ker(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^\perp \tag{33}$$

it holds:

(i) $\dim(\mathcal{M}) = \dim(\mathcal{N}) = \dim(\mathcal{F}) - \dim\big(\mathcal{F}\cap\mathcal{G}^\perp\big)$*;*

(ii) The orthogonal projectors are also given by:

$$\mathbf{P}_\mathcal{M} = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^+, \quad \mathbf{P}_\mathcal{N} = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^+(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) \tag{34}$$
$$\mathbf{P}_\mathcal{M} = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})^+(\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F}), \quad \mathbf{P}_\mathcal{N} = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})(\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})^+ \tag{35}$$

(iii) We have:

$$\mathbf{P}_\mathcal{M}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})\ \mathbf{P}_\mathcal{N} \tag{36}$$
$$\mathbf{P}_\mathcal{N}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^+ = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^+ = (\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^+\mathbf{P}_\mathcal{M} \tag{37}$$
$$\mathbf{P}_\mathcal{N}(\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F}) = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F}) = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})\ \mathbf{P}_\mathcal{M} \tag{38}$$
$$\mathbf{P}_\mathcal{M}(\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})^+ = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})^+ = (\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{F})^+\mathbf{P}_\mathcal{N} \tag{39}$$

(iv) $\mathcal{M} \subseteq \mathcal{F}$ *and* $\mathcal{N} \subseteq \mathcal{G}$*;*

(v) $\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{M} = \mathbf{P}_\mathcal{M} = \mathbf{P}_\mathcal{M}\mathbf{P}_\mathcal{F}$ *and* $\mathbf{P}_\mathcal{G}\mathbf{P}_\mathcal{N} = \mathbf{P}_\mathcal{N} = \mathbf{P}_\mathcal{N}\mathbf{P}_\mathcal{G}$*;*

(vi) It holds $\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G} = \mathbf{P}_\mathcal{M}\mathbf{P}_\mathcal{N}$ *and in particular:*

$$\mathbf{f}(\mathbf{x})^\top\mathbf{g}_0(y) = (\mathbf{P}_\mathcal{M}\mathbf{f}(\mathbf{x}))^\top\mathbf{P}_\mathcal{N}\mathbf{g}_0(y) \tag{40}$$

Proof. (i) By the rank-nullity theorem [Axler, 2015, page 62], we have that $\dim \mathrm{Im}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = d - \dim\ker(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})$. Notice that $\mathrm{Im}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = \mathcal{M}$ and $\ker(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = \mathbb{R}^d \setminus \mathcal{N}$, therefore

$$\dim(\mathcal{M}) = \dim \mathrm{Im}(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = d - \dim\ker(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G}) = d - d + \dim\ker(\mathbf{P}_\mathcal{F}\mathbf{P}_\mathcal{G})^\perp = \dim(\mathcal{N})\,. \tag{41}$$

Next, we derive the dimensionality of $\mathcal{M}$. Recall that for two matrices $\mathbf{A} \in \mathbb{R}^{m\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$, from the rank-nullity theorem [Axler, 2015, page 62], it follows that:

$$\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}(\mathbf{B}) - \dim(\ker(\mathbf{A}) \cap \operatorname{Im}(\mathbf{B}))\,. \tag{42}$$

Using this for $\mathbf{A} = \mathbf{P}_{\mathcal{G}}$ and $\mathbf{B} = \mathbf{P}_{\mathcal{F}}$, we have that:

$$\begin{aligned}
\dim(\mathcal{M}) &= \operatorname{rank}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}) && (43)\\
&= \operatorname{rank}(\mathbf{P}_{\mathcal{F}}) - \dim(\ker(\mathbf{P}_{\mathcal{G}}) \cap \operatorname{Im}(\mathbf{P}_{\mathcal{F}})) && (44)\\
&= \dim(\mathcal{F}) - \dim(\mathcal{G}^{\perp} \cap \mathcal{F}) && (45)\\
&= \dim(\operatorname{SIm}(\mathbf{f})) - \dim(\operatorname{SIm}(\mathbf{g}_0)^{\perp} \cap \operatorname{SIm}(\mathbf{f}))\,. && (46)
\end{aligned}$$

(ii) From the property of the pseudo-inverse, see Equation (29) and Equation (30), we have that

$$(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+} = \mathbf{P}_{\operatorname{Im}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})} = \mathbf{P}_{\mathcal{M}} \tag{47}$$

and that:

$$(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}) = \mathbf{P}_{\ker(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{\perp}} = \mathbf{P}_{\mathcal{N}}\,. \tag{48}$$

From Equation (47), taking the transpose of $\mathbf{P}_{\mathcal{M}}$ we get:

$$\begin{aligned}
\mathbf{P}_{\mathcal{M}}^{\top} &= ((\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+})^{\top}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{\top} && (49)\\
\mathbf{P}_{\mathcal{M}} &= (\mathbf{P}_{\mathcal{G}}^{\top}\mathbf{P}_{\mathcal{F}}^{\top})^{+}(\mathbf{P}_{\mathcal{G}}^{\top}\mathbf{P}_{\mathcal{F}}^{\top}) && (50)\\
&= (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})\,, && (51)
\end{aligned}$$

and from Equation (48), taking the transpose we obtain:

$$\begin{aligned}
\mathbf{P}_{\mathcal{N}}^{\top} &= (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{\top}((\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+})^{\top} && (52)\\
\mathbf{P}_{\mathcal{N}} &= (\mathbf{P}_{\mathcal{G}}^{\top}\mathbf{P}_{\mathcal{F}}^{\top})(\mathbf{P}_{\mathcal{G}}^{\top}\mathbf{P}_{\mathcal{F}}^{\top})^{+} && (53)\\
&= (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\,. && (54)
\end{aligned}$$

(iii) Denote $\mathbf{A} := \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}$. From the pseudo-inverse definition [Axler, 2015, page 221], it holds

$$(\mathbf{A}\mathbf{A}^{+})\mathbf{A} = \mathbf{A} = \mathbf{A}(\mathbf{A}^{+}\mathbf{A}) \tag{55}$$

and substituting for $\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}$ and the projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ we get:

$$\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}) = \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} = (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})\mathbf{P}_{\mathcal{N}}\,. \tag{56}$$

Similarly, for the pseudo-inverse, consider

$$(\mathbf{A}^{+}\mathbf{A})\mathbf{A}^{+} = \mathbf{A}^{+} = \mathbf{A}^{+}(\mathbf{A}\mathbf{A}^{+}) \tag{57}$$

and substituting we get

$$\mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+} = (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+} = (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}\mathbf{P}_{\mathcal{M}}\,. \tag{58}$$

Similarly to point (ii), taking the transpose of Equation (56), we obtain:

$$(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})\,, \tag{59}$$

and taking the transpose of Equation (58) we obtain:

$$(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\mathbf{P}_{\mathcal{N}} = (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+} = \mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\,. \tag{60}$$

(iv) To show this, notice that $\mathcal{M} := \operatorname{Im}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}) = \ker(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{\perp}$. Therefore, we have that

$$\ker(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{\perp} = \{\mathbf{P}_{\mathcal{F}}\mathbf{v} \notin \mathcal{G}^{\perp} \mid \mathbf{v} \in \mathbb{R}^d\}, \quad \mathcal{F} = \{\mathbf{P}_{\mathcal{F}}\mathbf{v} \neq \mathbf{0} \mid \mathbf{v} \in \mathbb{R}^d\}\,, \tag{61}$$

which shows that $\mathcal{M}$ is a subset of $\mathcal{F}$, and they are equal only when $\mathbf{P}_{\mathcal{G}} = \mathbf{I}$: $\mathcal{M} \subseteq \mathcal{F}$. Similarly, for $\mathcal{N}$, we get the same: $\mathcal{N} \subseteq \mathcal{G}$.

(v) Follows by the property of orthogonal projectors. Since $\mathcal{M} \subseteq \mathcal{F}$, we have that $\mathbf{P}_{\mathcal{M}}$ and$\mathbf{P}_{\mathcal{F}}$ commute, which means that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{F}} = \mathbf{P}_{\mathcal{M}}$. The same conclusion also holds for $\mathcal{N}$ and $\mathcal{G}$, because of $\mathcal{N} \subseteq \mathcal{G}$.

(vi) We have to rework the following expression

$$\begin{aligned}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} &= \mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}} &(62)\\ &= \mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}}\,, &(63)\end{aligned}$$

where in the second line we used that $\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})\mathbf{P}_{\mathcal{N}}$, and the final step is given by using that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}$. In particular:

$$\begin{aligned}\mathbf{f}(\mathbf{x})^\top\mathbf{g}_0(y) &= \mathbf{f}(\mathbf{x})^\top\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &(64)\\ &= (\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}))^\top\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y)\,, &(65)\end{aligned}$$

where in the last line we used that $\mathbf{P}_{\mathcal{M}}^\top = \mathbf{P}_{\mathcal{M}}$, being an orthogonal projector. □

B.3 Extended linear equivalence

We show that the relation defined in Definition 3 is an equivalence relation.

Definition 3 (Extended linear equivalence)**.** *Two models* $(\mathbf{f}, \mathbf{g})$ *and* $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ *are extended-linearly equivalent, if both (i)* $\dim(\mathcal{M}) = \dim(\tilde{\mathcal{M}})$ *and (ii) there exist two full-rank matrices* $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times\tilde{d}}$ *defining, respectively, invertible transformations from* $\mathcal{M}$ *to* $\tilde{\mathcal{M}}$*, and from* $\mathcal{N}$ *to* $\tilde{\mathcal{N}}$*, such that* $\mathbf{M}^\top\mathbf{N} = \mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{N}}}$ *and*

$$\begin{aligned}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) &(9)\\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\,, &(10)\end{aligned}$$

for all $y \in \mathcal{A}, \mathbf{x} \in \mathsf{Seq}(\mathcal{A})$*. We denote this relation by* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*.*

Proof. To prove $\sim_{EL}$ is an equivalence relation we have to show its reflexivity, symmetry, and transitivity.

Reflexivity. Take:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\mathbf{f}, \mathbf{g})\,. \tag{66}$$

This means that there must exist $\mathbf{M}, \mathbf{N}$ of rank $k := \dim(\mathcal{M})$ such that,

$$\begin{cases}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x})\\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y)\end{cases} \tag{67}$$

which are given by $\mathbf{M} = \mathbf{P}_{\mathcal{M}}$, $\mathbf{N} = \mathbf{P}_{\mathcal{N}}$.

Symmetry. We have to show that:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \iff (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \sim_{EL} (\mathbf{f}, \mathbf{g})\,. \tag{68}$$

This can be seen by showing one side of the implication ($\implies$):

$$\begin{cases}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})\\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\end{cases} \implies \begin{cases}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) &= \tilde{\mathbf{M}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x})\\ \mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) &= \tilde{\mathbf{N}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y)\end{cases} \tag{69}$$

Take:

$$\begin{cases}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})\\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\end{cases} \tag{70}$$

and consider the pseudo-inverses $\mathbf{M}^+$ and $\mathbf{N}^+$ obtaining:

$$\begin{cases}\mathbf{M}^+\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}^+\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})\\ \mathbf{N}^+\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}^+\mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\end{cases} \tag{71}$$

Notice that $\mathbf{M}^+\mathbf{M} = \mathbf{P}_{\tilde{\mathcal{M}}}$ and $\mathbf{N}^+\mathbf{N} = \mathbf{P}_{\tilde{\mathcal{N}}}$, which follows by the fact that $\ker(\mathbf{M})^\perp = \tilde{\mathcal{M}}$ and $\ker(\mathbf{N})^\perp = \tilde{\mathcal{N}}$ [Axler, 2015, Page 211]. Using this we have $\mathbf{M}^+\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}} = \mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{M}}} = \mathbf{P}_{\tilde{\mathcal{M}}}$, and similarly $\mathbf{N}^+\mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}} = \mathbf{P}_{\tilde{\mathcal{N}}}$. Therefore, on the right-hand side only the projectors $\mathbf{P}_{\tilde{\mathcal{M}}}$ and $\mathbf{P}_{\tilde{\mathcal{N}}}$ remain, whereas we have to set $\tilde{\mathbf{M}} = \mathbf{M}^+$ and $\tilde{\mathbf{N}} = \mathbf{N}^+$. Both $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{N}}$ have range k, as a consequence of being pseudo-inverses. Therefore, we get $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \sim_{EL} (\mathbf{f}, \mathbf{g})$.

The other side of the implication ($\Longleftarrow$) follows a similar proof.

Transitivity. We have to show that:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \ \wedge \ (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \sim_{EL} (\mathbf{f}^*, \mathbf{g}^*) \implies (\mathbf{f}, \mathbf{g}) \sim_{EL} (\mathbf{f}^*, \mathbf{g}^*) \,. \tag{72}$$

This can be verified by substitution:

$$\begin{cases} \mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) = \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) \\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) = \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) \end{cases} \tag{73}$$

$$\begin{cases} \mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) = \mathbf{M}\tilde{\mathbf{M}}\mathbf{P}_{\mathcal{M}^*}\mathbf{f}^*(\mathbf{x}) \\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) = \mathbf{N}\tilde{\mathbf{N}}\mathbf{P}_{\mathcal{N}^*}{\mathbf{g}^*}_0(y) \end{cases} \tag{74}$$

and by setting $\overline{\mathbf{M}} = \mathbf{M}\tilde{\mathbf{M}}$ and $\overline{\mathbf{N}} = \mathbf{N}\tilde{\mathbf{N}}$, it holds:

$$\begin{cases} \mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) = \overline{\mathbf{M}}\mathbf{P}_{\mathcal{M}^*}\mathbf{f}^*(\mathbf{x}) \\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) = \overline{\mathbf{N}}\mathbf{P}_{\mathcal{N}^*}{\mathbf{g}^*}_0(y) \end{cases} \tag{75}$$

Notice that, since $\mathbf{M} : \tilde{\mathcal{M}} \to \mathcal{M}$ is a linear invertible transformation, and similarly $\tilde{\mathbf{M}} : \mathcal{M}^* \to \tilde{\mathcal{M}}$ is also a linear invertible transformation, the composition $\overline{\mathbf{M}} = \mathbf{M}\tilde{\mathbf{M}}$ is a linear invertible transformation from $\mathcal{M}^*$ to $\mathcal{M}$, with $\operatorname{rank}(\overline{\mathbf{M}}) = k$. A similar observation also applies to $\overline{\mathbf{N}} : \mathcal{N}^* \to \mathcal{N}$. Therefore, we have shown that:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\mathbf{f}^*, \mathbf{g}^*) \,. \tag{76}$$

This concludes the proof. □

Explicit form on N. Based on the requirement that matrices $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d \times \tilde{d}}$ obey $\mathbf{M}^\top\mathbf{N} = \mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{N}}}$, we have that:

$$\mathbf{N} := (\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}})^+(\mathbf{M}^\top)^+(\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{N}}}) \tag{77}$$

Proof. Notice that the following holds:

- $\mathbf{M}^\top\mathbf{P}_{\mathcal{F}} = \mathbf{M}^\top$, by the fact that $\mathbf{M}^\top = \mathbf{M}^\top\mathbf{P}_{\mathcal{M}} = \mathbf{M}^\top\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}}$ by Lemma 2 (v), and
- $\mathbf{P}_{\mathcal{G}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^+ = (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^+$, since we have $(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})\mathbf{P}_{\mathcal{G}} = (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})$.

We use these identities to obtain:

$$\begin{aligned} \mathbf{M}^\top(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^+(\mathbf{M}^\top)^+(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) &= \mathbf{M}^\top\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^+(\mathbf{M}^\top)^+(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) && (78) \\ &= \mathbf{M}^\top\mathbf{P}_{\mathcal{M}}(\mathbf{M}^\top)^+(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) && \text{(Using Lemma 2 (ii))} \\ &= \mathbf{M}^\top(\mathbf{M}^\top)^+(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) && (\text{Using } \mathbf{M}^\top\mathbf{P}_{\mathcal{M}} = \mathbf{M}^\top) \\ &= \mathbf{P}_{\tilde{\mathcal{M}}}(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) && (\mathbf{M}^\top(\mathbf{M}^\top)^+ = \mathbf{P}_{\ker(\mathbf{M})^\perp} = \mathbf{P}_{\mathcal{M}}) \\ &= \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}} \,. && (79) \end{aligned}$$

where we used that $\mathbf{P}_{\tilde{\mathcal{M}}}(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}) = \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}$ by Lemma 2 (iii). Notice that by Lemma 2 (vi) we have that $\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}} = \mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{N}}}$. □

B.4 Proof of Proposition 4

Proposition 4. *If* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*, then*

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}_0(y)\,. \tag{11}$$

Proof. Starting from Lemma 2 (vi), we have that:

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = \mathbf{f}(\mathbf{x})^\top \mathbf{P}_{\mathcal{M}} \mathbf{P}_{\mathcal{N}} \mathbf{g}_0(y)\,. \tag{80}$$

Considering the expression of $\mathbf{f}$ and $\mathbf{g}_0$ given by the $\sim_{EL}$ equivalence relation (Definition 3):

$$\mathbf{P}_{\mathcal{M}} \mathbf{f}(\mathbf{x}) = \mathbf{M} \mathbf{P}_{\tilde{\mathcal{M}}} \tilde{\mathbf{f}}(\mathbf{x}) \tag{81}$$
$$\mathbf{P}_{\mathcal{N}} \mathbf{g}_0(y) = \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y) \tag{82}$$

we use also the condition that $\mathbf{M}^\top \mathbf{N} = \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{N}}}$. We get

$$\begin{aligned}
\mathbf{f}(\mathbf{x})^\top \mathbf{P}_{\mathcal{M}} \mathbf{P}_{\mathcal{N}} \mathbf{g}_0(y) &= \tilde{\mathbf{f}}(\mathbf{x})^\top \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{M}^\top \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y) \\
&= \tilde{\mathbf{f}}(\mathbf{x})^\top \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{N}}} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y) \\
&= \tilde{\mathbf{f}}(\mathbf{x})^\top \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y)\,,
\end{aligned}$$

where we used the idempotency of projectors, i.e., that $\mathbf{P}^2_{\tilde{\mathcal{M}}} = \mathbf{P}_{\tilde{\mathcal{M}}}$ and $\mathbf{P}^2_{\tilde{\mathcal{N}}} = \mathbf{P}_{\tilde{\mathcal{N}}}$. To prove the claim, it is sufficient to use Lemma 2 (vi) again, obtaining:

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) = \tilde{\mathbf{f}}(\mathbf{x})^\top \mathbf{P}_{\tilde{\mathcal{M}}} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\mathbf{g}}_0(y) = \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}_0(y)\,. \tag{83}$$

□

B.5 Counterexample when diversity condition does not hold

We detail here a counter-example to linear identifiability (Corollary 6) when the diversity condition does not hold. Let $\mathbf{f} : \mathsf{Seq}(\mathcal{A}) \to \mathbb{R}^2$ and $\mathbf{g} : \mathcal{A} \to \mathbb{R}^2$, and $\mathcal{A} = \{y_0, y_1, y_2\}$. Let $\mathbf{g}(y_0) = (1, 0)^\top$, $\mathbf{g}(y_1) = (1, 1)^\top$, and $\mathbf{g}(y_2) = (1, -1)^\top$ be unembeddings, which do not fulfill the diversity condition (Definition 1): In fact, these unembeddings give $\mathcal{G} = \mathrm{span}(\mathbf{e}_2)$, and $\dim(\mathcal{G}) = 1$ which is less than the dimensionality of the representation space. The vector $\mathbf{e}_2 := (0, 1)^\top$ is drawn as a **blue** arrow in Figure 4. We can construct another model where $\tilde{\mathbf{g}} = \mathbf{g}$ and choose $\tilde{\mathbf{f}}(\mathbf{x}) = (\mathbf{f}_1(\mathbf{x}) + 0.2 \cos(40 a_1/\pi), \mathbf{f}_2(\mathbf{x}))^\top$. Figure 4 shows this transformation. By construction, this model generates the same next-token distribution of the first one; however, the two model representations are not equal up to a linear transformation.

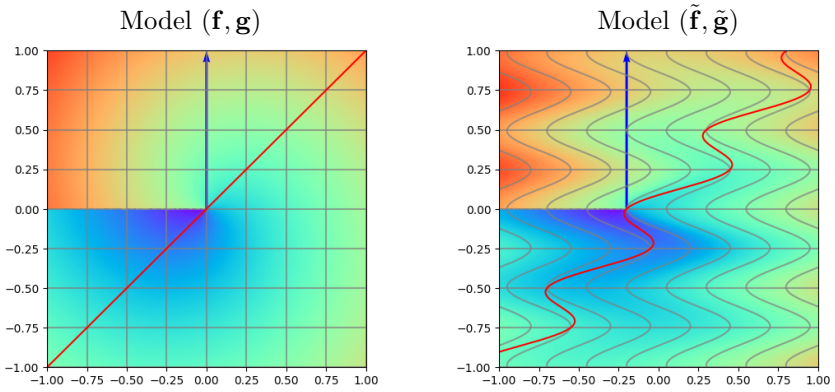

Figure 4: **Allowed distorsions among $\sim_{EL}$-equivalent models**. From the left, model embeddings $\mathbf{f}$ are given different colors. The red segment is non-linearly transformed on the right along $\tilde{\mathbf{f}}_1$, whereas they remain equal to the left on the component $\tilde{\mathbf{f}}_2$. This shows that the models are not $\sim_L$-equivalent.

B.6 Proof of Theorem 5

We begin with the following lemma, which will be used in the proof of Theorem 5.

Lemma 17. *Let $\mathcal{F}$ be a subspace of $\mathbb{R}^d$ and $\mathcal{M} \subseteq \mathcal{F}$ a subspace of $\mathcal{F}$. Consider D elements $\mathbf{v}_i \in \mathbb{R}^d$, such that the matrix*

$$\mathbf{F} := \big(\mathbf{v}_1, \dots, \mathbf{v}_D\big) \tag{84}$$

has range $\mathrm{Im}(\mathbf{F}) = \mathcal{F}$. *Then, it holds* $\mathrm{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{F}) = \dim(\mathcal{M})$.

Proof. For two matrices $\mathbf{A} \in \mathbb{R}^{m\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times n}$, from the rank-nullity theorem [Axler, 2015, page 62], it follows that:

$$\mathrm{rank}(\mathbf{AB}) = \mathrm{rank}(\mathbf{B}) - \dim(\ker(\mathbf{A}) \cap \mathrm{range}(\mathbf{B}))\,. \tag{85}$$

We use this formula to obtain:

$$\begin{aligned}
\mathrm{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{F}) &= \mathrm{rank}(\mathbf{F}) - \dim(\ker(\mathbf{P}_{\mathcal{M}}) \cap \mathrm{range}(\mathbf{F})) && (86)\\
&= \dim(\mathcal{F}) - \dim\big(\mathcal{M}^{\perp} \cap \mathcal{F}\big) && (87)\\
&= \dim(\mathcal{F}) - \dim\big(\mathbb{R}^d \setminus \mathcal{M} \cap \mathcal{F}\big) && (88)\\
&= \dim(\mathcal{F}) - \dim\big(\mathcal{F} \setminus \mathcal{M}\big) && (89)\\
&= \dim(\mathcal{F}) - \dim(\mathcal{F}) + \dim(\mathcal{M}) && (90)\\
&= \dim(\mathcal{M}) && (91)
\end{aligned}$$

where the equality on the second-last line follows from the fact that $\mathcal{M} \subseteq \mathcal{F}$. □

We now proceed to prove Theorem 5.

Theorem 5. *For all $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$, with representation dimensions d and $\tilde{d}$ (not necessarily equal),*

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \iff (\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})\,. \tag{12}$$

Proof sketch. To prove the implication, we divide into five parts:

1. Starting from log-equality of probabilities, we adopt a pivoting strategy to get rid of normalizing constants[7];
2. We derive an explicit expression of $\mathbf{M}$ and $\tilde{\mathbf{M}}$ such that

   $$\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) = \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})$$
   $$\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) = \tilde{\mathbf{M}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x})$$

   hold for every $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$. We achieve that by considering q points, or tokens, of $\mathcal{A}$, such that the matrices $\mathbf{G} := (\mathbf{g}_0(y_1), \dots, \mathbf{g}_0(y_q))$ and $\tilde{\mathbf{G}} := (\tilde{\mathbf{g}}_0(y_1), \dots, \tilde{\mathbf{g}}(y_q))$ have images corresponding to $\mathrm{SIm}(\mathbf{g}_0)$ and $\mathrm{SIm}(\tilde{\mathbf{g}}_0)$, respectively;
3. From the linear relation obtained between $\mathbf{f}$ and $\tilde{\mathbf{f}}$, we show that, having $\dim(\mathcal{M}) = k$, also $\dim(\tilde{\mathcal{M}}) = k$ and $\mathrm{rank}(\mathbf{M}) = k$. By considering ℓ points, or sequences, $x_i \in \mathsf{Seq}(\mathcal{A})$, such that the matrices $\mathbf{F} := (\mathbf{f}(\mathbf{x}_q), \dots, \mathbf{f}(\mathbf{x}_\ell))$ and $\tilde{\mathbf{F}} := (\tilde{\mathbf{f}}(\mathbf{x}_q), \dots, \tilde{\mathbf{f}}(\mathbf{x}_\ell))$ have images corresponding to $\mathrm{SIm}(\mathbf{f})$ and $\mathrm{SIm}(\tilde{\mathbf{f}})$, respectively, we will obtain that

   $$k \leq m \leq k, \quad k \leq k' \leq k\,;$$

4. We derive an explicit expression of $\mathbf{N}$ such that

   $$\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) = \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)$$

   hold for every $y \in \mathcal{A}$. We achieve that by considering again ℓ points such that $\mathbf{F}$ and $\tilde{\mathbf{F}}$ have images corresponding to $\mathrm{SIm}(\mathbf{f})$ and $\mathrm{SIm}(\tilde{\mathbf{f}})$, and using the linear relation obtained between $\mathbf{f}$ and $\tilde{\mathbf{f}}$. This allows us to derive an expression of $\mathbf{N}$ that depends on $\mathbf{M}$;

[7] A similar "pivoting" strategy is the starting point of several identifiability proofs in nonlinear ICA with auxiliary variables, e.g., Roeder et al. [2021], Khemakhem et al. [2020a,b], Hyvarinen et al. [2019].

5. Finally, we show that it follows that $\text{rank}(\mathbf{N})$ is equal to $k := \dim(\mathcal{M})$. To achieve that we use the relation between $\mathbf{g}_0$ and $\tilde{\mathbf{g}}_0$ to show that:
$$k \leq \text{rank}(\mathbf{N}) \leq k\,.$$

Recall that we are indicating with:

$$\mathcal{F} := \text{SIm}(\mathbf{f}), \quad \tilde{\mathcal{F}} := \text{SIm}(\tilde{\mathbf{f}}), \quad \mathcal{G} := \text{SIm}(\mathbf{g}_0), \quad \tilde{\mathcal{G}} := \text{SIm}(\tilde{\mathbf{g}}_0)\,,$$

and with:

$$\mathcal{M} := \text{Im}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}), \quad \tilde{\mathcal{M}} := \text{Im}(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}), \quad \mathcal{N} := \ker(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{\perp}, \quad \tilde{\mathcal{N}} := \ker(\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}})^{\perp}\,. \tag{92}$$

Proof. We start by proving the implication ($\implies$).

Step 1: Using pivoting to get rid of normalizing constants. We start from the equality between log probabilities:

$$\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y) - \log Z(\mathbf{x}) = \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y) - \log \tilde{Z}(\mathbf{x}) \tag{93}$$

Subtracting the pivot $y_0 \in \mathcal{A}$ to all remaining elements $y \in \mathcal{A}$, and defining $\mathbf{g}_0 := \mathbf{g}(y) - \mathbf{g}(y_0)$ and $\tilde{\mathbf{g}}_0(y) := \tilde{\mathbf{g}}(y) - \tilde{\mathbf{g}}(y_0)$, we get rid of the terms containing the log of the normalizing constant, *i.e.*,

$$\begin{aligned} \mathbf{f}(\mathbf{x})^\top(\mathbf{g}(y) - \mathbf{g}(y_0)) &= \tilde{\mathbf{f}}(\mathbf{x})^\top(\tilde{\mathbf{g}}(y) - \tilde{\mathbf{g}}(y_0)) && \text{(Subtract by } \mathbf{f}(\mathbf{x})^\top\mathbf{g}(y_0) \text{ and by } \tilde{\mathbf{f}}(\mathbf{x})^\top\tilde{\mathbf{g}}(y_0)\text{)} \\ \mathbf{f}(\mathbf{x})^\top \mathbf{g}_0(y) &= \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}_0(y)\,. && (94) \end{aligned}$$

Step 2: Obtaining the relation between embeddings $\mathbf{f}$ and $\tilde{\mathbf{f}}$. We will now show that $\mathbf{M} := (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^+(\mathbf{G}^\top)^+\tilde{\mathbf{G}}^\top(\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}})$ and $\tilde{\mathbf{M}} := (\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}})^+(\tilde{\mathbf{G}}^\top)^+\mathbf{G}^\top\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}$ satisfies Equation (9), showing that $\mathbf{f}$ and $\tilde{\mathbf{f}}$ are linearly related. For this, consider q elements, or tokens, $y \in \mathcal{A}$ such that the matrix

$$\mathbf{G} = \big(\mathbf{g}_0(y_1), \quad \cdots, \quad \mathbf{g}_0(y_q)\big)\,, \quad \tilde{\mathbf{G}} = \big(\tilde{\mathbf{g}}_0(y_1), \quad \cdots, \quad \tilde{\mathbf{g}}_0(y_q)\big) \tag{95}$$

span $\mathcal{G}$ and $\tilde{\mathcal{G}}$, respectively, *i.e.*, $\text{Im}(\mathbf{G}) = \mathcal{G}$ and $\text{Im}(\tilde{\mathbf{G}}) = \tilde{\mathcal{G}}$. Taking the transpose of Equation (94) and considering these q elements, we can write

$$\mathbf{G}^\top \mathbf{f}(\mathbf{x}) = \tilde{\mathbf{G}}^\top \tilde{\mathbf{f}}(\mathbf{x})\,. \tag{96}$$

Next, we consider the projectors on the subspace where all functions' images are contained. Indicate with $\mathbf{P}_{\mathcal{F}}$ and $\mathbf{P}_{\mathcal{G}}$ the orthogonal projectors on $\mathcal{F}$ and $\mathcal{G}$, respectively, and with $\mathbf{P}_{\tilde{\mathcal{F}}}$ and $\mathbf{P}_{\tilde{\mathcal{G}}}$ the orthogonal projectors on $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$, respectively. It holds

$$\mathbf{f}(\mathbf{x}) = \mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}), \quad \mathbf{G} = \mathbf{P}_{\mathcal{G}}\mathbf{G}, \quad \tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{x}), \quad \tilde{\mathbf{G}} = \mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{G}}\,, \tag{97}$$

which can be inserted in Equation (96), leading to

$$\mathbf{G}^\top\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}) = \tilde{\mathbf{G}}^\top\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{x})\,. \tag{98}$$

We make use of the pseudo-inverse of $\mathbf{G}^\top$ to obtain:

$$\begin{aligned} (\mathbf{G}^\top)^+\mathbf{G}^\top\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}) &= (\mathbf{G}^\top)^+\tilde{\mathbf{G}}^\top\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{x}) && \text{(Multiply on the left by } (\mathbf{G}^\top)^+\text{)} \\ \mathbf{P}_{\text{Im}(\mathbf{G})}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}) &= (\mathbf{G}^\top)^+\tilde{\mathbf{G}}^\top\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{x}) && \text{(From Equation (32), } (\mathbf{G}^\top)^+\mathbf{G}^\top = \mathbf{P}_{\text{Im}(\mathbf{G})}\text{)} \\ \mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{x}) &= \mathbf{A}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{x})\,, && (99) \end{aligned}$$

where in the last line we used $\mathbf{P}_{\text{Im}(\mathbf{G})}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{G}}$ since $\text{Im}(\mathbf{G}) = \mathcal{G}$, and we denoted with $\mathbf{A} := (\mathbf{G}^\top)^+\tilde{\mathbf{G}}^\top$.

Next, we insert the projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\tilde{\mathcal{M}}}$ in Equation (99), using Lemma 2 (iii) we get:

$$\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})\,. \tag{100}$$

We now consider the left pseudo-inverse of $\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}$ to obtain

$$
\begin{aligned}
(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\mathbf{A}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) && \text{(Multiply on the left by } (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\text{)} \\
\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}\mathbf{A}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) && \text{(By Lemma 2 (ii) } \mathbf{P}_{\mathcal{M}} = (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})\text{)} \\
\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x}) &= \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x})\,, && (101)
\end{aligned}
$$

where we used the idempotency of the $\mathbf{P}_{\mathcal{M}}$ projector, and we defined

$$
\mathbf{M} := (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}(\mathbf{G}^{\top})^{+}\tilde{\mathbf{G}}^{\top}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}}\,. \tag{102}
$$

Following similar steps but starting from $\tilde{\mathbf{G}}^{\top}\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{G}^{\top}\mathbf{f}(\mathbf{x})$, we arrive at a similar expression for $\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}$:

$$
\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{x}) = \tilde{\mathbf{M}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{x})\,, \tag{103}
$$

where $\tilde{\mathbf{M}} := (\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}})^{+}(\tilde{\mathbf{G}}^{\top})^{+}\mathbf{G}^{\top}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}}$.

Step 3: Showing that $\dim(\tilde{\mathcal{M}}) = \operatorname{rank}(\mathbf{M}) = \dim(\mathcal{M})$. Let $k := \dim(\mathcal{M})$ and $k' := \dim(\tilde{\mathcal{M}})$. Also, let $m := \operatorname{rank}(\mathbf{M})$. We want to show that $k' = m = k$ and to this end we will obtain that $k \leq m \leq k$ and that $k' \leq k \leq k$. This is done in three points:

(**I**) We show that $m \leq k$ from the definition of $\mathbf{M}$ in Equation (102);

(**II**) We show that necessarily $m \geq k$ and $k' \geq k$ from Equation (101);

(**III**) We show that $k' \leq k$ from Equation (103).

(**I**) By equation (102), we have that $\mathbf{M} = \mathbf{P}_{\mathcal{M}}\mathbf{M}$, since by Equation (102), in $\mathbf{M}$ we have on the left $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+} = (\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}})^{+}$ by Lemma 2 (ii). Taking the rank of $\mathbf{M}$ and using the fact that $\operatorname{rank}(\mathbf{AB}) \leq \min\big(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\big)$ [Axler, 2015], we obtain:

$$
\begin{aligned}
\operatorname{rank}(\mathbf{M}) &= \operatorname{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{M}) && \text{(Take the rank of } \mathbf{M} = \mathbf{P}_{\mathcal{M}}\mathbf{M}\text{)} \\
m &\leq \min\big(\operatorname{rank}(\mathbf{P}_{\mathcal{M}}), \operatorname{rank}(\mathbf{M})\big) && \text{(Using } \operatorname{rank}(\mathbf{AB}) \leq \min\big(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\big)\text{)} \\
m &\leq \min\big(m, k\big) && (104) \\
\implies m &\leq k && (105)
\end{aligned}
$$

Next, we consider ℓ sequences $\mathbf{x}_i \in \mathsf{Seq}(\mathcal{A})$ such that the matrices:

$$
\mathbf{F} = \big(\mathbf{f}(\mathbf{x}_1), \quad \cdots, \quad \mathbf{f}(\mathbf{x}_\ell)\big)\,, \quad \tilde{\mathbf{F}} = \big(\tilde{\mathbf{f}}(\mathbf{x}_1), \quad \cdots, \quad \tilde{\mathbf{f}}(\mathbf{x}_\ell)\big) \tag{106}
$$

span the whole $\mathcal{F}$ and $\tilde{\mathcal{F}}$, respectively, *i.e.*, $\operatorname{Im}(\mathbf{F}) = \mathcal{F}$ and $\operatorname{Im}(\tilde{\mathbf{F}}) = \tilde{\mathcal{F}}$. Moreover, since $\mathcal{M} \subseteq \mathcal{F}$ we have that by Lemma 17 that $\operatorname{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{F}) = k$. Similarly, we have $\operatorname{rank}(\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}}) = k'$.

(**II**) We consider from the Equation (101) the condition for ℓ points:

$$
\mathbf{P}_{\mathcal{M}}\mathbf{F} = \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}}\,. \tag{107}
$$

We evaluate the rank from Equation (107) to obtain:

$$
\begin{aligned}
\operatorname{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{F}) &= \operatorname{rank}(\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}}) && (108) \\
k &\leq \min\big(\operatorname{rank}(\mathbf{M}), \operatorname{rank}(\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}})\big) && \text{(Using } \operatorname{rank}(\mathbf{AB}) \leq \min\big(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\big)\text{)} \\
k &\leq \min\big(m, k'\big) && (109) \\
\implies m &\geq k, \quad k' \geq k && (110)
\end{aligned}
$$

Together with (105), it shows that $k \leq m \leq k$, so it must be that $m = k$ and so $\operatorname{rank}(\mathbf{M}) = \dim(\mathcal{M})$.

(**III**) From the Equation (103), we get similarly:

$$\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}} = \tilde{\mathbf{M}}\mathbf{P}_{\mathcal{M}}\mathbf{F}\,. \tag{111}$$

Following a similar proof to (**II**), we evaluate the rank from Equation (111) to obtain:

$$\begin{aligned} \operatorname{rank}(\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}}) &\leq \min\big(\operatorname{rank}(\tilde{\mathbf{M}}), \operatorname{rank}(\mathbf{P}_{\mathcal{M}}\mathbf{F})\big) && \text{(Using } \operatorname{rank}(\mathbf{AB}) \leq \min\big(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\big)\text{)} \\ &\implies k' \leq k && (112) \end{aligned}$$

which, together with (110), it holds only as long as $k' \leq k$. This shows that $k \leq k' \leq k$, so it must be that $k' = k$.

Hence, we have shown that $\dim(\tilde{\mathcal{M}}) = \dim(\mathcal{M}) = \operatorname{rank}(\mathbf{M})$. In particular, it holds that $\mathbf{M}$ is an invertible map from $\tilde{\mathcal{M}}$ to $\mathcal{M}$, with pseudo-inverse $\mathbf{M}^{+}$, such that:

$$\operatorname{Im}(\mathbf{M}) = \mathcal{M}, \quad \ker(\mathbf{M})^{\perp} = \tilde{\mathcal{M}}\,. \tag{113}$$

Step 4. Obtaining the relation between unembeddings $\mathbf{g}$ and $\tilde{\mathbf{g}}$. We will now show that $\mathbf{N} := (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}(\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}$, using the matrix $\mathbf{M}$ in Equation (102) from **Step 2** satisfies Equation (10).

Similar to **Step 2**, we take ℓ points, or sequences, $\mathbf{x}_i \in \mathsf{Seq}(\mathcal{A})$ such that $\mathbf{F}$ and $\tilde{\mathbf{F}}$ in Equation (106) span $\mathcal{F}$ and $\tilde{\mathcal{F}}$, respectively. We then have:

$$\begin{aligned} \mathbf{F}^{\top}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\tilde{\mathbf{g}}_0(y) && \text{(Considering } \ell \text{ points for } \mathbf{F} \text{ and } \tilde{\mathbf{F}}\text{)} \\ \mathbf{F}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(Using orthogonal projectors Equation (97))} \\ \mathbf{F}^{\top}\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(Using Lemma 2 (iii))} \\ (\mathbf{P}_{\mathcal{M}}\mathbf{F})^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(Taking the transpose } \mathbf{F}^{\top}\mathbf{P}_{\mathcal{M}} = (\mathbf{P}_{\mathcal{M}}\mathbf{F})^{\top}\text{)} \\ (\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{F}})^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && (114) \end{aligned}$$

where in the last line we substituted the expression for $\mathbf{P}_{\mathcal{M}}\mathbf{F}$ given by Equation (102). Thus, restarting from (114), and reworking the expression we get:

$$\begin{aligned} \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(Expanding the transpose on the left)} \\ (\tilde{\mathbf{F}}^{\top})^{+}\tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= (\tilde{\mathbf{F}}^{\top})^{+}\tilde{\mathbf{F}}^{\top}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(Multiply by pseudo-inverse } (\tilde{\mathbf{F}}^{\top})^{+}\text{)} \\ \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y) && \text{(From Equation (32) we get } (\tilde{\mathbf{F}}^{\top})^{+}(\tilde{\mathbf{F}}^{\top}) = \mathbf{P}_{\operatorname{Im}(\tilde{\mathbf{F}})} = \mathbf{P}_{\tilde{\mathcal{F}}}\text{)} \\ \mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{g}_0(y) &= \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\tilde{\mathbf{g}}_0(y)\,, && (115) \end{aligned}$$

where we used in the last line that $\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top} = \mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top} = \mathbf{M}^{\top}$, and $\mathbf{P}_{\tilde{\mathcal{M}}}\mathbf{M}^{\top} = \mathbf{M}^{\top}$ follows by the definition of $\mathbf{M}$ (Equation (102)), containing on the right $(\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}})$. Recall that by Lemma 2 (iii), it holds that

$$\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}, \quad \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}} = \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\,. \tag{116}$$

Using this in Equation (115), we obtain

$$\begin{aligned} \mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) && (117) \\ (\mathbf{M}^{\top})^{+}\mathbf{M}^{\top}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= (\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) && \text{(Multiply on the left for the pseudo-inverse } (\mathbf{M}^{\top})^{+}\text{)} \\ \mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= (\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\,, && (118) \end{aligned}$$

where in the last line we use $(\mathbf{M}^{\top})^{+}\mathbf{M}^{\top} = \mathbf{P}_{\operatorname{Im}(\mathbf{M})} = \mathbf{P}_{\mathcal{M}}$, where the first equality follows by Equation (32) and the second one is given by $\operatorname{Im}(\mathbf{M}) = \mathcal{M}$ from Equation (113). We now use that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}$ to obtain:

$$\begin{aligned} \mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= (\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) && (119) \\ (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}(\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) && \text{(Multiply by pseudo-inverse } (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}\text{)} \\ \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}(\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y) && ((\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}} = \mathbf{P}_{\mathcal{N}}\text{, by Lemma 2 (ii)}) \\ \mathbf{P}_{\mathcal{N}}\mathbf{g}_0(y) &= \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{g}}_0(y)\,, && (120) \end{aligned}$$

where we set $\mathbf{N} := (\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}(\mathbf{M}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\mathbf{P}_{\tilde{\mathcal{G}}}$. This expression is in line with that of Definition 3, showing the equivalence relation.

Step 5. Showing that $\text{rank}(\mathbf{N}) = \dim(\mathcal{N})$. It remains to show that $\mathbf{N}$ has rank equal to $k := \dim(\mathcal{N}) = \dim(\mathcal{M})$. Similarly to **Step 3**, we will show that $k \leq \text{rank}(\mathbf{N}) \leq k$ and we proceed with two points:

(**I**) We show that by the form of $\mathbf{N}$, we obtain $\text{rank}(\mathbf{N}) \leq k$;

(**II**) We use Equation (120) to obtain that $\text{rank}(\mathbf{N}) \geq k$.

(**I**) Notice that $\mathbf{N}$ is left-invariant by multiplication to $\mathbf{P}_{\mathcal{N}}$, because of the term $(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+} = \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{F}}\mathbf{P}_{\mathcal{G}})^{+}$, by Lemma 2 (iii). Therefore:

$$\begin{aligned}
\mathbf{N} &= \mathbf{P}_{\mathcal{N}}\mathbf{N} && (121)\\
\text{rank}(\mathbf{N}) &= \text{rank}(\mathbf{P}_{\mathcal{N}}\mathbf{N}) && (122)\\
\text{rank}(\mathbf{N}) &\leq \min\big(\text{rank}(\mathbf{P}_{\mathcal{N}}), \text{rank}(\mathbf{N})\big) && (\text{Using } \text{rank}(\mathbf{AB}) \leq \min\big(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\big))\\
\text{rank}(\mathbf{N}) &\leq \min\big(k, \text{rank}(\mathbf{N})\big) && (123)\\
\implies \text{rank}(\mathbf{N}) &\leq k && (124)
\end{aligned}$$

(**II**) Next, consider q elements of $\mathcal{A}$ such that the matrices $\mathbf{G}$ and $\tilde{\mathbf{G}}$ in Equation (95) have rank equal to $\dim(G)$ and $\dim(\tilde{\mathcal{G}})$, respectively. From Equation (120), it holds:

$$\mathbf{P}_{\mathcal{N}}\mathbf{G} = \mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{G}} \tag{125}$$

Notice that $\text{rank}(\mathbf{P}_{\mathcal{N}}\mathbf{G}) = \dim(\mathcal{N})$ and $\text{rank}(\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{G}}) = \dim(\tilde{\mathcal{N}})$ by Lemma 17, and it also holds $\dim(\mathcal{N}) = \dim(\mathcal{M}) = \dim(\tilde{\mathcal{M}}) = \dim(\tilde{\mathcal{N}})$ by Lemma 2 (i) and **Step 3**. Let $k := \dim(\mathcal{N})$. Using this in Equation (125), we obtain:

$$\begin{aligned}
\text{rank}(\mathbf{P}_{\mathcal{N}}\mathbf{G}) &= \text{rank}(\mathbf{N}\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{G}}) && (126)\\
k &\leq \min\big(\text{rank}(\mathbf{N}), \text{rank}(\mathbf{P}_{\tilde{\mathcal{N}}}\tilde{\mathbf{G}})\big) && (127)\\
k &\leq \min\big(\text{rank}(\mathbf{N}), k\big) && (128)\\
\implies \text{rank}(\mathbf{N}) &\geq k && (129)
\end{aligned}$$

This shows that, combined with Equation (124) we have $k \leq \text{rank}(\mathbf{N}) \leq k$, which means that $\text{rank}(\mathbf{N}) = k$. Taking **Steps 2, 3, 4, 5** together, we have that:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})\,. \tag{130}$$

This shows the implication.

($\Longleftarrow$) To prove the other direction show that also $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \implies p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}(y \mid \mathbf{x})$, for all $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ and all $y \in \mathcal{A}$. We start from Proposition 4, which gives:

$$(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \implies \mathbf{f}(\mathbf{x})^{\top}\mathbf{g}_0(y) = \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}_0(y) \tag{131}$$

for all $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ and all $y \in \mathcal{A}$. We continue from the right-hand side to obtain:

$$\begin{aligned}
\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y) - \mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y_0) &= \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y) - \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y_0) && (\text{Use explicit expression for } \mathbf{g}_0 \text{ and } \tilde{\mathbf{g}}_0)\\
\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y) &= \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y) + \mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y_0) - \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y_0) && (\text{Reordering all } y_0 \text{ terms on the right})\\
\exp(\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y)) &= \exp(\tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y)) \cdot \exp\big(\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y_0) - \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y_0)\big)\\
& && (\text{Taking the exponential on both sides})\\
\frac{\exp(\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y))}{Z(\mathbf{x})} &= \exp(\tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y)) \cdot \frac{\exp\big(\mathbf{f}(\mathbf{x})^{\top}\mathbf{g}(y_0) - \tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y_0)\big)}{Z(\mathbf{x})}\\
& && (\text{Dividing by the normalizing constant } Z(\mathbf{x}))\\
p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) &= \exp(\tilde{\mathbf{f}}(\mathbf{x})^{\top}\tilde{\mathbf{g}}(y)) \cdot \frac{1}{\tilde{Z}(\mathbf{x})}\,, && (132)
\end{aligned}$$

where in the last line we included the expression for the conditional probability $p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y))/Z(\mathbf{x})$ from Equation (1), and we denoted $\tilde{Z}(\mathbf{x}) := Z(\mathbf{x})/\exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y_0) - \tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y_0)\right)$. To obtain the value of $\tilde{Z}$ we consider the sum over all $y \in \mathcal{A}$ for Equation (132), giving:

$$\sum_{y\in\mathcal{A}} p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = \sum_{y\in\mathcal{A}} \exp(\tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y)) \cdot \frac{1}{\tilde{Z}(\mathbf{x})} \tag{133}$$

$$1 = \sum_{y\in\mathcal{A}} \exp(\tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y)) \cdot \frac{1}{\tilde{Z}(\mathbf{x})} \tag{134}$$

$$\tilde{Z}(\mathbf{x}) = \sum_{y\in\mathcal{A}} \exp(\tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y)) \tag{135}$$

which means that, from Equation (132) we have:

$$p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = \frac{\exp(\tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y))}{\sum_{y\in\mathcal{A}} \exp(\tilde{\mathbf{f}}(\mathbf{x})^\top \tilde{\mathbf{g}}(y))} \tag{136}$$

$$= p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}(y \mid \mathbf{x}) \tag{137}$$

showing the claim. This concludes the proof. □

B.7 Proof of Corollary 6

The following corollary constitutes a special case of Theorem 5, which can be easily proven by setting $d = \tilde{d}$ and requiring that $\mathcal{M} = \mathcal{N} = \mathbb{R}^d$.Here, we provide an alternative proof expanding previous results by Roeder et al. [2021], relaxing two assumptions that were used in that context. For comparison, we report the statement by Roeder et al. [2021]. To this end, fix a pivot $\mathbf{x}_0 \in \mathsf{Seq}(\mathcal{A})$ and indicate with

$$\mathbf{f}_0(\mathbf{x}) := \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0) \tag{138}$$

the difference between embeddings and the pivot.

Theorem (Roeder et al. [2021])**.** *Given two models* $(\mathbf{f},\mathbf{g}), (\mathbf{f},\mathbf{g}) \in \Theta$*, under the assumption that:*

1. $\mathrm{SIm}(\mathbf{f}_0) = \mathrm{SIm}(\mathbf{g}_0) = \mathbb{R}^d$*;*
2. $\mathrm{SIm}(\tilde{\mathbf{f}}_0) = \mathrm{SIm}(\tilde{\mathbf{g}}_0) = \mathbb{R}^d$*;*

it holds:

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \implies (\mathbf{f},\mathbf{g}) \sim_L (\tilde{\mathbf{f}},\tilde{\mathbf{g}}) \tag{139}$$

where the linear equivalence relation is given by:

$$(\mathbf{f},\mathbf{g}) \sim_L (\tilde{\mathbf{f}},\tilde{\mathbf{g}}) \iff \begin{cases} \mathbf{f}(\mathbf{x}) & = \mathbf{M}\tilde{\mathbf{f}}(\mathbf{x}) \\ \mathbf{g}_0(y) & = \mathbf{N}\tilde{\mathbf{g}}_0(y) \end{cases} \tag{140}$$

$\forall \mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ *and* $\forall y \in \mathcal{A}$*, where* $\mathbf{M}^\top\mathbf{N} = \mathbf{I}$ *and in particular* $\mathbf{N} = \mathbf{M}^{-\top}$*.*

To highlight deviations, we present a proof that follows a somewhat analogous argument to the proof of Roeder et al. [2021]; a direct proof may show $\mathbf{N} = \mathbf{M}^{-1}$ relying on Theorem 5. We relax condition *2* and use the fact the assumption that $\mathrm{SIm}(\mathbf{f}) = \mathbb{R}^d$, which is a milder condition to requiring that $\mathrm{SIm}(\mathbf{f}_0) = \mathbb{R}^d$. We prove the following:

Corollary 6 (Adapted from [Roeder et al., 2021])**.** *For all* $(\mathbf{f},\mathbf{g}), (\tilde{\mathbf{f}},\tilde{\mathbf{g}}) \in \Theta_d$ *such that* $(\mathbf{f},\mathbf{g})$ *satisfies the diversity condition (Definition 1), we have*

$$p_{\mathbf{f},\mathbf{g}} = p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}} \implies (\mathbf{f},\mathbf{g}) \sim_L (\tilde{\mathbf{f}},\tilde{\mathbf{g}})\,, \tag{14}$$

where, by definition, $(\mathbf{f},\mathbf{g}) \sim_L (\tilde{\mathbf{f}},\tilde{\mathbf{g}})$ *if and only if there exists an invertible matrix* $\mathbf{M} \in \mathbb{R}^{d\times d}$ *such that for all* $y \in \mathcal{A}, \mathbf{x} \in \mathsf{Seq}(\mathcal{A})$ *we have*

$$\mathbf{f}(\mathbf{x}) = \mathbf{M}\tilde{\mathbf{f}}(\mathbf{x}) \text{ and } \mathbf{g}_0(y) = \mathbf{M}^{-\top}\tilde{\mathbf{g}}_0(y)\,. \tag{15}$$

Proof. Our proof follows a similar technique to Lachapelle et al. [2023, Theorem B.4].

Identifiability of $\mathbf{f}$. First notice that from the equivalnce of log-likelihood we can write:

$$\log p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = \log p_{\tilde{\mathbf{f}},\tilde{\mathbf{g}}}(y \mid \mathbf{x}) \tag{141}$$

$$\mathbf{g}(y)^\top \mathbf{f}(\mathbf{x}) - \log Z(\mathbf{x}) = \tilde{\mathbf{g}}(y)^\top \tilde{\mathbf{f}}(\mathbf{x}) - \log \tilde{Z}(\mathbf{x}) \tag{142}$$

$$(\mathbf{g}(y) - \mathbf{g}(y_0))^\top \mathbf{f}(\mathbf{x}) = (\tilde{\mathbf{g}}(y) - \tilde{\mathbf{g}}(y_0))^\top \tilde{\mathbf{f}}(\mathbf{x}) \tag{143}$$

$$\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{x}) = \tilde{\mathbf{g}}_0(y)^\top \tilde{\mathbf{f}}(\mathbf{x}) \tag{144}$$

where in the last line we subtracted the pivot for different log-probabilities on the points y and y_0. We now consider the matrix $\mathbf{G}$ constructed to contain d differences:

$$\mathbf{G} = \big(\mathbf{g}_0(y_1), \quad \cdots \quad \mathbf{g}_0(y_d)\big) \tag{145}$$

such that it is invertible. Since by the diversity condition $\mathrm{SIm}(\mathbf{g}_0) = \mathbb{R}^d$, such a marix always exists. Let $\tilde{\mathbf{G}}$ be the corresponding matrix of differences for $\tilde{\mathbf{g}}$, we obtain:

$$\mathbf{G}^\top \mathbf{f}(\mathbf{x}) = \tilde{\mathbf{G}}^\top \tilde{\mathbf{f}}(\mathbf{x}) \tag{146}$$

Then, we obtain:

$$\mathbf{f}(\mathbf{x}) = \mathbf{G}^{-\top} \tilde{\mathbf{G}}^\top \tilde{\mathbf{f}}(\mathbf{x}) \tag{147}$$

$$\mathbf{f}(\mathbf{x}) = \mathbf{M} \tilde{\mathbf{f}}(\mathbf{x}) \tag{148}$$

where we denoted as $\mathbf{M} = \mathbf{G}^{-\top} \tilde{\mathbf{G}}^\top$. Next, since $\mathrm{SIm}(\mathbf{f}) = \mathbb{R}^d$ we can consider d elements $\mathbf{x}_i \in \mathsf{Seq}(\mathcal{A})$ such that

$$\mathbf{F} = \big(\mathbf{f}(x_1), \quad \cdots, \quad \mathbf{f}(\mathbf{x}_d)\big) \tag{149}$$

is invertible. Let $\tilde{\mathbf{F}}$ be the corresponding matrix for $\tilde{\mathbf{f}}$. In this way we obtain the following:

$$\mathbf{F} = \mathbf{M} \tilde{\mathbf{F}}\,, \tag{150}$$

and since $\mathbf{F}$ is invertible, it must be that also $\mathbf{M}$ and $\mathbf{F}$ are invertible matrices of rank d. This shows that

$$\mathbf{f}(\mathbf{x}) = \mathbf{M} \tilde{\mathbf{f}}(\mathbf{x}) \tag{151}$$

$\mathbf{M} \in \mathbb{R}^{d \times d}$ is invertible.

Identifiability of $\mathbf{g}_0$. Next, we consider the implication for $\mathbf{g}_0$. We start again from the pivot difference of Equation Equation (144):

$$\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{x}) = \tilde{\mathbf{g}}_0(y)^\top \tilde{\mathbf{f}}(\mathbf{x}) \tag{152}$$

$$\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{x}) = \tilde{\mathbf{g}}_0(y)^\top \mathbf{M}^{-1} \mathbf{f}(\mathbf{x}) \tag{153}$$

where we substituted $\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{M}^{-1} \mathbf{f}(\mathbf{x})$ from Equation (99). Therefore, taking d points $\mathbf{x} \in \mathsf{Seq}(\mathcal{A})$, such that the matrix $\mathbf{F}$ is invertible, restarting from the transpose of Equation (153) we obtain:

$$\mathbf{F}^\top \mathbf{g}_0(y) = \mathbf{F}^\top \mathbf{M}^{-\top} \tilde{\mathbf{g}}_0(y) \tag{154}$$

$$\mathbf{g}_0(y) = \mathbf{M}^{-\top} \tilde{\mathbf{g}}_0(y) \qquad \text{(Multiplying for the inverse of } \mathbf{F}^\top\text{)}$$

$$\mathbf{g}_0(y) = \mathbf{M}^{-\top} \tilde{\mathbf{g}}_0(y)\,, \tag{155}$$

This means that we have:

$$\mathbf{f}(\mathbf{x}) \;\; = \mathbf{M} \tilde{\mathbf{f}}(\mathbf{x}) \tag{156}$$

$$\mathbf{g}_0(y) = \mathbf{N} \tilde{\mathbf{g}}_0(y)\,, \tag{157}$$

where we have defined $\mathbf{N} := \mathbf{M}^{-\top}$, such that $\mathbf{M}^\top \mathbf{N} = \mathbf{I}$, proving the claim. □

C Additional Results and Proofs of Section 4

C.1 Relational Linear Steering Property

We here want to discuss an additional linear property besides those presented in Section 4, termed *linear steering property*. This behavior is also referred to as the *linear intervening property* by Park et al. [2024a].

It has been empirically observed that there exist steering vectors that influence next-token predictions [Hernandez et al., 2024, Park et al., 2024a, Hase et al., 2024, Arditi et al., 2024], in the following sense: If $\mathbf{v}$ encode the average difference between English to Italian embeddings, adding $\mathbf{v}$ to $\mathbf{f}(\mathbf{s})$ for the sentence $\mathbf{s}$ ="*The king sits on the*" would change the most-likely next token prediction y ="*throne*" to $\hat{y}$ ="*trono*", and similarly this applies for other sentences, affecting the most-likely next-token prediction to move from the English token to the Italian counterpart.

We define this property as follows:

Definition 18 (Relational Linear Steering)**.** *We say that a model $(\mathbf{f}, \mathbf{g}) \in \Theta$ possess linear relational steering for $\mathbf{q}_0$ and the set of $\{\mathbf{q}_1, \dots, \mathbf{q}_m\}$, for $m \geq 1$ queries $\mathbf{q}_j \neq \mathbf{q}_0$, if (1) it linearly represents $\mathbf{q}_0$ in Γ_0 and all $\mathbf{q}_j$ on Γ_j, and (2) there exists a vector $\mathbf{v} \in \mathbb{R}^d$ such that:*

$$\mathbf{P}_{\Gamma_0} \mathbf{A}_{\mathbf{q}_0} \mathbf{v} \neq \mathbf{0}, \quad \mathbf{P}_{\Gamma_j} \mathbf{A}_{\mathbf{q}_j} \mathbf{v} = \mathbf{0}, \quad \forall j \in [m] \tag{158}$$

We prove that relational linearity allows for this property:

Proposition 19. *If (1) $(\mathbf{f}, \mathbf{g})$ linearly represents $\mathbf{q}_0$ on Γ_0 (Definition 9) and (2) $(\mathbf{f}, \mathbf{g})$ linearly represents $m \geq 1$ queries $\mathbf{q}_j \neq \mathbf{q}_0$ on Γ_j, such that $\left(\bigcup_j \Gamma_{\mathbf{q}_j}\right) \cap \Gamma_{\mathbf{q}_0} \subsetneq \Gamma_{\mathbf{q}_0}$, then the model $(\mathbf{f}, \mathbf{g})$ satisfies linear relational steering for $\mathbf{q}_0$ and the set of queries $\{\mathbf{q}_1, \dots, \mathbf{q}_m\}$.*

Proof. From the assumptions (1) and (2) we have that relational linearity as per Definition 9 implies that:

$$\begin{cases} \mathbf{P}_{\Gamma_0} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_0) & = \mathbf{P}_{\Gamma_0} \mathbf{A}_{\mathbf{q}_0} \mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma_0} \mathbf{a}_{\mathbf{q}_0} \\ \mathbf{P}_{\Gamma_j} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_j) & = \mathbf{P}_{\Gamma_j} \mathbf{A}_{\mathbf{q}_j} \mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma_j} \mathbf{a}_{\mathbf{q}_j}, \quad \forall j \in [\ell] \end{cases} \tag{159}$$

Let $\Gamma_{0,int} = \bigcup_{j=1}^{\ell} \Gamma_{\mathbf{q}_j} \cap \Gamma_{\mathbf{q}_0}$. By assumption (2), $\Gamma_{0,int} \subsetneq \Gamma_0$, meaning that it exists a non-empty $\Gamma_{0,\perp} = \Gamma_{\mathbf{q}_0} \setminus \Gamma_{0,int}$. Let $\mathbf{v} \in \Gamma_{0,\perp}$. It holds:

$$\mathbf{P}_{\Gamma_0} \mathbf{A}_{\mathbf{q}_0} \mathbf{v} = \mathbf{v} \tag{160}$$

because $\Gamma_{0,\perp} \subseteq \Gamma_{\mathbf{q}_0}$. Proceeding similarly, it holds $\Gamma_{0,\perp} \cap \Gamma_{\mathbf{q}_j} = \varnothing \ \forall j \in [\ell]$, by assumption (2). Therefore, we have:

$$\mathbf{P}_{\Gamma_j} \mathbf{A}_{\mathbf{q}_j} \mathbf{v} = \mathbf{0} \tag{161}$$

showing the claim. □

This means that adding $\mathbf{v}$ to $\mathbf{f}(\mathbf{s})$ would alter only the value of $\mathbf{P}_{\Gamma_0} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_0)$ without changing that of $\mathbf{P}_{\Gamma_j} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_j)$, for $j \in [m]$. As a result, the modified representation $\mathbf{f}(\mathbf{s}) + \mathbf{v}$ would not affect the next-token prediction on other queries $\mathbf{q}_j$. For example, take $\mathbf{q}_0$ ="*Is the previous sentence written in English?*", $\mathbf{q}_1$ ="*Is the previous sentence written in Italian?*", and $\mathbf{q}_2$ = *"Does the previous sentence contain the symbol "+"?"*, and consider $\Gamma_0 = \Gamma_1 = \Gamma_2 = \text{span}(\mathbf{g}(\text{"}no\text{"}) - \mathbf{g}(\text{"}yes\text{"}))$. When assumptions of Proposition 19 hold, we can alter the reply to the question $\mathbf{q}_0$ by adding a vector proportional to $\mathbf{v}$, *e.g.*, moving from English to another language, without affecting the representation on $\mathbf{q}_1$, *i.e.*, moving to another language but not Italian, and the representation on $\mathbf{q}_2$, *i.e.*, leaving the symbol *"+"* in the sentence if present.

C.2 Proof of Lemma 8

Lemma 8. *Consider a model $(\mathbf{f}, \mathbf{g}) \in \Theta$. For $y_0, y_1, y_2, y_3 \in \mathcal{A}$, the difference vectors $\mathbf{g}(y_1) - \mathbf{g}(y_0)$ and $\mathbf{g}(y_3) - \mathbf{g}(y_2)$ are parallel in $\mathcal{N}$ if and only if there exists $\beta \neq 0$, s.t. $\forall \mathbf{s} \in \mathsf{Seq}(\mathcal{A})$*

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s})} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(y_2 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_3 \mid \mathbf{s})}. \tag{16}$$

Proof. We start by considering the equality between log-ratios appearing as the ($\Longleftarrow$) condition. Writing it down we obtain:

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s})} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(y_2 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_3 \mid \mathbf{s})} \tag{162}$$

$$\log \frac{\exp(\mathbf{f}(\mathbf{s})^\top \mathbf{g}(y_0))}{\exp(\mathbf{f}(\mathbf{s})^\top \mathbf{g}(y_1))} = \beta \cdot \log \frac{\exp(\mathbf{f}(\mathbf{s})^\top \mathbf{g}(y_2))}{\exp(\mathbf{f}(\mathbf{s})^\top \mathbf{g}(y_3))} \tag{163}$$

$$\log \exp\left(\mathbf{f}(\mathbf{s})^\top (\mathbf{g}(y_0) - \mathbf{g}(y_1))\right) = \beta \cdot \log \exp\left(\mathbf{f}(\mathbf{s})^\top (\mathbf{g}(y_2) - \mathbf{g}(y_3))\right) \tag{164}$$

$$\mathbf{f}(\mathbf{s})^\top (\mathbf{g}(y_0) - \mathbf{g}(y_1)) = \beta \cdot \mathbf{f}(\mathbf{s})^\top (\mathbf{g}(y_2) - \mathbf{g}(y_3)) \tag{165}$$

And substituting $\mathbf{g}_1(y_0) := \mathbf{g}(y_0) - \mathbf{g}(y_1)$ and $\mathbf{g}_3(y_2) := \mathbf{g}(y_2) - \mathbf{g}(y_3)$ we obtain:

$$\mathbf{f}(\mathbf{s})^\top \mathbf{g}_1(y_0) = \beta \cdot \mathbf{f}(\mathbf{s})^\top \mathbf{g}_3(y_2) \tag{166}$$

Consider ℓ elements $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$, such that:

$$\mathbf{F} = (\mathbf{f}(\mathbf{s}_1), \dots, \mathbf{f}(\mathbf{s}_\ell)) \tag{167}$$

spans $\mathcal{F} := \mathrm{SIm}(\mathbf{f})$. We then obtain:

$$\mathbf{F}^\top \mathbf{g}_1(y_0) = \beta \mathbf{F}^\top \mathbf{g}_3(y_2) \tag{168}$$

and multiplying both sides of Equation (168) from the left with the pseudo-inverse of $\mathbf{F}^\top$ we get:

$$\mathbf{P}_{\mathcal{F}} \mathbf{g}_1(y_0) = \beta \mathbf{P}_{\mathcal{F}} \mathbf{g}_3(y_2). \tag{169}$$

Notice that, both $\mathbf{g}_1(y_0), \mathbf{g}_3(y_2) \in \mathcal{G} := \mathrm{SIm}(\mathbf{g}_0)$, then it holds $\mathbf{g}_1(y_0) = \mathbf{P}_{\mathcal{G}} \mathbf{g}_1(y_0)$ $\mathbf{g}_3(y_2) = \mathbf{P}_{\mathcal{G}} \mathbf{g}_3(y_2)$. Using this we obtain:

$$\mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}} \mathbf{g}_1(y_0) = \beta \mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}} \mathbf{g}_3(y_2) \tag{170}$$

$$(\mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}})^+ \mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}} \mathbf{g}_1(y_0) = \beta (\mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}})^+ \mathbf{P}_{\mathcal{F}} \mathbf{P}_{\mathcal{G}} \mathbf{g}_3(y_2) \tag{171}$$

$$\mathbf{P}_{\mathcal{N}} \mathbf{g}_1(y_0) = \beta \mathbf{P}_{\mathcal{N}} \mathbf{g}_3(y_2)\,. \tag{172}$$

Regrouping the two terms on one side we can see that to have parallelism in $\mathcal{N}$ (Definition 7), we must have the following:

$$\mathbf{P}_{\mathcal{N}}(\mathbf{g}_1(y_0) - \beta \cdot \mathbf{g}_3(y_2)) = \mathbf{0} \tag{173}$$

i.e., $\mathbf{g}_1(y_0) - \beta \cdot \mathbf{g}_3(y_2) \in \mathcal{N}^\perp$. Therefore, $\mathbf{g}_1(y_0)$ and $\mathbf{g}_3(y_2)$ are parallel in $\mathcal{N}$.

The implication ($\Longrightarrow$) is given by a similar proof by starting from Definition 7, *i.e.*, that $\mathbf{P}_{\mathcal{N}} \mathbf{g}_1(y_0) = \beta \mathbf{P}_{\mathcal{N}} \mathbf{g}_3(y_2)$. Therefore by multiplying the two for any embedding $\mathbf{f}(\mathbf{s})$ we get

$$\mathbf{g}_1(y_0)^\top \mathbf{P}_{\mathcal{N}} \mathbf{f}(\mathbf{s}) = \beta \mathbf{g}_3(y_2)^\top \mathbf{P}_{\mathcal{N}} \mathbf{f}(\mathbf{s}) \tag{174}$$

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s})} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(y_2 \mid \mathbf{s})}{p_{\mathbf{f},\mathbf{g}}(y_3 \mid \mathbf{s})}\,. \tag{175}$$

This shows the claim. □

Remark 20. *We exclude the case $\beta = 0$ because, for $\mathbf{a} = 0$, a contradiction arises. Specifically, any vector $\mathbf{b} \in \mathbb{R}^d$ would be trivially parallel to $\mathbf{a}$ (i.e., $\mathbf{a} = \beta \mathbf{b}$ with $\beta = 0$). However, conversely, we would obtain that no scalar $\beta \in \mathbb{R}$ exists such that $\mathbf{b} = \beta \mathbf{a}$.*

C.3 Proof of Proposition 11

Proposition 11 (ΓLR $\Longrightarrow$ LS)**.** *Suppose that a model $(\mathbf{f}, \mathbf{g}) \in \Theta$ (i) linearly represents $\mathbf{q}$ on $\Gamma \subseteq \mathrm{SIm}(\mathbf{g}_0)$, and (ii) $\Gamma_{\mathbf{q}} \subseteq \mathrm{SIm}(\mathbf{g}_0)$, then the model $(\mathbf{f}, \mathbf{g})$ linearly represents $\Gamma_{\mathbf{q}}$ relative to $\mathbf{q}$ (Definition 10).*

Proof. We start from the relational linearity as per Definition 9 for $\mathbf{q}$ on $\Gamma \subset \mathbb{R}^d$:

$$\mathbf{P}_{\mathcal{G}} \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_{\mathcal{G}} \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \mathbf{P}_{\mathcal{G}} \mathbf{a}_{\mathbf{q}}\,, \tag{176}$$

and recall that $\Gamma_{\mathbf{q}} := \mathrm{Im}(\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma) = \{\mathbf{A}_{\mathbf{q}}^\top \mathbf{v} \mid \mathbf{v} \in \Gamma\}$. By assumption (2) $\Gamma_{\mathbf{q}} \subseteq \mathrm{SIm}(\mathbf{g}_0)$. Then, for any pair $y_i, y_j \in \mathcal{A}$ such that $\mathbf{g}_i(y) := \mathbf{g}(y_j) - \mathbf{g}(y_i) \in \Gamma_{\mathbf{q}}$, we can find a vector $\boldsymbol{\gamma} \in \Gamma$ such that

$$\mathbf{A}_{\mathbf{q}}^\top \boldsymbol{\gamma} = \mathbf{g}_i(y_j)\,. \tag{177}$$

Notice that from this expression we can also write:

$$\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma \boldsymbol{\gamma} = \mathbf{g}_i(y_j)\,, \tag{178}$$

since $\mathbf{P}_\Gamma \boldsymbol{\gamma} = \boldsymbol{\gamma}$. Hence, the vector $\boldsymbol{\gamma}$ can be obtained taking the pseudo-inverse of $\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma$:

$$(\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+ \mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma \boldsymbol{\gamma} = (\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+ \mathbf{g}_i(y_j) \tag{179}$$

$$\mathbf{P}_{\mathrm{Im}(\mathbf{P}_\Gamma \mathbf{A}_{\mathbf{q}})} \boldsymbol{\gamma} = (\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+ \mathbf{g}_i(y_j) \tag{180}$$

Notice that, since $(\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+ \mathbf{P}_{\Gamma_{\mathbf{q}}} = (\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+$ and $\mathbf{P}_{\Gamma_{\mathbf{q}}} \mathbf{g}_i(y_j) = \mathbf{g}_i(y_j)$, we have that $(\mathbf{A}_{\mathbf{q}}^\top \mathbf{P}_\Gamma)^+ \mathbf{g}_i(y_j) = \mathbf{0}$ only when $\mathbf{g}_i(y_j) = \mathbf{0}$. As a consequence, when $\mathbf{g}_i(y_j) \neq \mathbf{0}$, we have that also $\mathbf{P}_{\mathrm{Im}(\mathbf{P}_\Gamma \mathbf{A}_{\mathbf{q}})} \boldsymbol{\gamma} \neq \mathbf{0}$. Fix this $\boldsymbol{\gamma}$ and consider:

$$\boldsymbol{\gamma}^\top \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) = \boldsymbol{\gamma}^\top \big(\mathbf{f}(\mathbf{s} \frown \mathbf{q}) - \mathbf{a}_{\mathbf{q}}\big) \tag{181}$$

$$(\mathbf{A}_{\mathbf{q}}^\top \boldsymbol{\gamma})^\top \mathbf{f}(\mathbf{s}) = \boldsymbol{\gamma}^\top \big(\mathbf{f}(\mathbf{s} \frown \mathbf{q}) - \mathbf{a}_{\mathbf{q}}\big) \tag{182}$$

$$\mathbf{g}_i(y_j)^\top \mathbf{f}(\mathbf{s}) = \boldsymbol{\gamma}^\top \big(\mathbf{f}(\mathbf{s} \frown \mathbf{q}) - \mathbf{a}_{\mathbf{q}}\big)\,, \tag{183}$$

which shows that $(\mathbf{f}, \mathbf{g})$ linearly represents $\Gamma_{\mathbf{q}}$ related to $\mathbf{q}$, showing the claim. □

C.4 Proof of Proposition 13

Proposition 13 (ΓLR $\implies$ LP)**.** *If a model $(\mathbf{f}, \mathbf{g}) \in \Theta$ (i) linearly represents $\mathbf{q}$ on Γ, and (ii) $\mathbf{g}(y_i) - \mathbf{g}(y_j) \in \Gamma$ for all $y_i \in \mathcal{Y}_P$, then the model can be linear probed (Definition 12) for $\mathbf{q}$ and $\mathcal{Y}_P$, with parameters given by $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_\ell)^\top$ and $\mathbf{b} = (b_1, \ldots, b_\ell)^\top$, where $\mathbf{w}_i := \mathbf{A}_{\mathbf{q}}^\top \mathbf{g}(y_i)$ and $b_i := (\mathbf{a}_{\mathbf{q}})^\top \mathbf{g}(y_i)$.*

Proof. Under the assumption (2), take a pivot $y_j \in \mathcal{Y}_P$, then for all remaining $y_i \in \mathcal{Y}_P$ denote with $\mathbf{g}_j(y_i) := \mathbf{g}(y_i) - \mathbf{g}(y_j) \in \Gamma$. Taking the log-ratios between the conditional probabilities

$$p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P), \quad \text{and } p_{\mathbf{f},\mathbf{g}}(y_j \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)\,, \tag{184}$$

for conditional probabilities restricted to $\mathcal{Y}_P$, as in Definition 12, we obtain:

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)}{p_{\mathbf{f},\mathbf{g}}(y_j \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} = \log \frac{\exp\big(\mathbf{g}(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})\big)}{\exp\big(\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})\big)} \tag{185}$$

$$= \log \exp\big((\mathbf{g}(y_i) - \mathbf{g}(y_j))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})\big) \tag{186}$$

$$= \big((\mathbf{g}(y_i) - \mathbf{g}(y_j))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})\big) \tag{187}$$

$$= \mathbf{g}_j(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})\,. \tag{188}$$

Due to relational linearity of $\mathbf{q}$ onto Γ (Assumption (1)), we can write following (see Definition 9):

$$\mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_\Gamma \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \mathbf{P}_\Gamma \mathbf{a}_{\mathbf{q}}\,, \tag{189}$$

we can then take any $\mathbf{g}_j(y_i) \in \Gamma$, and multiply their transpose times both sides of Equation (189) from the right. We then get

$$\mathbf{g}_j(y_i)^\top \mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{g}_j(y_i)^\top \mathbf{P}_\Gamma \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \mathbf{g}_j(y_i)^\top \mathbf{P}_\Gamma \mathbf{a}_{\mathbf{q}} \tag{190}$$

$$(\mathbf{P}_\Gamma \mathbf{g}_j(y_i))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = (\mathbf{P}_\Gamma \mathbf{g}_j(y_i))^\top \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + (\mathbf{P}_\Gamma \mathbf{g}_j(y_i))^\top \mathbf{a}_{\mathbf{q}} \tag{191}$$

$$\mathbf{g}_j(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{g}_j(y_i)^\top \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \mathbf{g}_j(y_i)^\top \mathbf{a}_{\mathbf{q}}\,. \tag{192}$$

We can then substitute the expression on the RHS in Equation (192) to Equation (188) to obtain:

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)}{p_{\mathbf{f},\mathbf{g}}(y_j \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} = \mathbf{g}_j(y_i)^\top \mathbf{A}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \mathbf{g}_j(y_i)^\top \mathbf{a}_{\mathbf{q}}\,. \tag{193}$$

Now take the conditional probability $p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)$ which can be written as

$$
\begin{aligned}
p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) &= \frac{e^{\mathbf{g}(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} && (194)\\
&= \frac{e^{\mathbf{g}(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} \frac{e^{\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{e^{\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}} && \text{(Multiply and divide by the same term)}\\
&= \frac{e^{(\mathbf{g}(y_i) - \mathbf{g}(y_j))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} e^{\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})} && \text{(Rearrange terms in the exponential)}\\
&= \frac{e^{\mathbf{g}_j(y_i)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} e^{\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})} && \text{(Substitute } \mathbf{g}_j(y_i)\text{)}\\
&= \exp\left(\mathbf{g}_j(y_i)^\top \mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{g}_j(y_i)^\top \mathbf{a}_\mathbf{q}\right) \frac{e^{(\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)} && \text{(Use Equation (192))}\\
&= \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) \exp\left(-\mathbf{g}(y_j)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) \frac{e^{(\mathbf{g}(y_j)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)}\\
& && \text{(Separate the term depending on } y_i \text{ to those that do not)}\\
&= \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) \frac{e^{\mathbf{g}(y_j)^\top \left(\mathbf{f}(\mathbf{s} \frown \mathbf{q}) - \mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q}\right)}}{Z(\mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P)}\\
& && \text{(Rearrange exponential on the right)}\\
&= \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) C\,, && (195)
\end{aligned}
$$

where we denoted with C the scaling factor applied to the exponential, which we will treat as a constant since it does not depend on $y_i \in \mathcal{A}$. From this expression, take the sum on $\mathcal{Y}_P$ to obtain that:

$$
\begin{aligned}
\sum_{y_i \in \mathcal{Y}_P} \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) C &= \sum_{y_i \in \mathcal{Y}_P} p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) && (196)\\
\sum_{y_i \in \mathcal{Y}_P} \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) C &= 1 && \text{(The sum on the right equals to 1)}\\
C &= 1 \Big/ \sum_{y_i \in \mathcal{Y}_P} \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) && (197)
\end{aligned}
$$

Denote with $\mathbf{w}_i := \mathbf{A}_\mathbf{q}^\top \mathbf{g}(y_i)$ and with $b_i := \mathbf{g}(y_i)^\top \mathbf{a}_\mathbf{q}$. Then using this and Equation (197) inside Equation (195) we get

$$
\begin{aligned}
p_{\mathbf{f},\mathbf{g}}(y_i \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) &= \exp\left(\mathbf{g}(y_i)^\top (\mathbf{A}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{a}_\mathbf{q})\right) C && (198)\\
&= \frac{\exp\left(\mathbf{w}_i^\top \mathbf{f}(\mathbf{s}) + b_i\right)}{\sum_{y_i \in \mathcal{Y}_P} \exp\left(\mathbf{w}_i \mathbf{f}(\mathbf{s}) + b_i\right)} && \text{(Substitute for } \mathbf{w}_i \text{ and } b_i\text{, and Equation (197))}\\
&= \operatorname{softmax}\big(\mathbf{W}\mathbf{f}(\mathbf{s}) + \mathbf{b}\big)_i\,, && (199)
\end{aligned}
$$

where we defined $\mathbf{W} := (\mathbf{w}_1, \ldots, \mathbf{w}_\ell)^\top$ and $\mathbf{b} := (b_1, \ldots, b_\ell)^\top$. This shows that the model $(\mathbf{f}, \mathbf{g})$ can be linear probed for $\mathbf{q}$ in $\mathcal{Y}_P$ with $\mathbf{W}$ and $\mathbf{b}$. □

D Proof of Section 5

D.1 Proof of Theorem 14

Theorem 14. *For two models* $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$ *s.t.* $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$*, if* $\mathbf{f}$ *linearly represents* $\mathbf{q}$ *on* $\Gamma \subseteq \mathcal{N}$*, and* $\Gamma_{\mathbf{q}} \subseteq \mathcal{M}$*, then* $\tilde{\mathbf{f}}$ *linearly represents* $\mathbf{q}$ *on* $\tilde{\Gamma} \subseteq \tilde{\mathcal{N}}$*, where* $\tilde{\Gamma} = \mathrm{Im}(\mathbf{N}^{+}\mathbf{P}_{\Gamma})$ *and* $\mathbf{N}$ *is the matrix relating* $\mathbf{g}_0$ *and* $\tilde{\mathbf{g}}_0$ *by the equivalence relation in Definition 3.*

Proof Sketch. The proof is divided in two steps:

1. We first prove the implication that $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ linearly represents $\mathbf{q}$ on a subset $\tilde{\Gamma} \subseteq \mathbb{R}^d$;
2. Then we show that $\tilde{\Gamma} \subseteq \mathcal{N}$ and that $\tilde{\Gamma}_{\mathbf{q}} := \mathrm{Im}(\tilde{\mathbf{A}}_{\mathbf{q}}^{\top}\mathbf{P}_{\Gamma}) \subseteq \tilde{\mathcal{N}}$.

Proof. **Step 1**. We begin from the relational linearity definition for model $(\mathbf{f}, \mathbf{g})$. It holds

$$\mathbf{P}_{\Gamma}\mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}}\,, \tag{200}$$

where $\Gamma \subseteq \mathcal{N}$. This also means that $\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\Gamma}$. Denote with $\mathcal{F} := \mathrm{SIm}(\mathbf{f})$ and with $\mathcal{G} := \mathrm{SIm}(\mathbf{g}_0)$. By assumption, it holds that $\Gamma_{\mathbf{q}} := \mathrm{Im}(\mathbf{A}_{\mathbf{q}}^{\top}\mathbf{P}_{\Gamma}) = \{\mathbf{A}_{\mathbf{q}}^{\top}\mathbf{v} \mid \mathbf{v} \in \Gamma\}$ is a subset of $\mathcal{M}$. This implies, in turn, that $\mathbf{P}_{\mathcal{M}}\mathbf{A}_{\mathbf{q}}^{\top}\mathbf{P}_{\Gamma} = \mathbf{A}_{\mathbf{q}}^{\top}\mathbf{P}_{\Gamma}$. We use this to write:

$$\begin{aligned}
\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}} && \text{(Using } \mathbf{P}_{\Gamma} = \mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\text{)} \\
\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{F}}\mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}} && \text{(Using } \mathbf{f} = \mathbf{P}_{\mathcal{F}}\mathbf{f} \text{ on the left and } \mathbf{A}_{\mathbf{q}} = \mathbf{A}_{\mathbf{q}}\mathbf{P}_{\mathcal{M}} \text{ on the right)} \\
\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{P}_{\mathcal{M}}\mathbf{f}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}}\,, && (201)
\end{aligned}$$

where in the last line we used that $\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{G}}\mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}$ by Lemma 2 (vi). We now substitute the expression for $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ based on the RHS of the equivalence relation Equation (9) in Equation (201) to get

$$\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{P}_{\mathcal{N}}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}}\,. \tag{202}$$

Now starting from $\mathbf{N}^{\top}\mathbf{M} = \mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}}$, as specified in Definition 3, we can apply the following steps:

$$\begin{aligned}
\mathbf{N}^{\top}\mathbf{M} &= \mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}} && (203) \\
(\mathbf{N}^{\top})^{+}\mathbf{N}^{\top}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}} && \text{(Multiply on the left by the pseudo-inverse } (\mathbf{N}^{\top})^{+}\text{)} \\
\mathbf{P}_{\mathcal{N}}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}} && \text{(Using } (\mathbf{N}^{\top})^{+}\mathbf{N}^{\top} = \mathbf{P}_{\mathcal{N}}\text{)} \\
\mathbf{P}_{\mathcal{N}}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}} && \text{(Idempotency of the orthogonal projector } \mathbf{P}_{\tilde{\mathcal{N}}}\text{)} \\
\mathbf{P}_{\mathcal{N}}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}} && \text{(Substitute } \mathbf{P}_{\tilde{\mathcal{G}}}\mathbf{P}_{\tilde{\mathcal{F}}} = \mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{M}}} \text{ from Lemma 2 (vi))} \\
\mathbf{P}_{\mathcal{N}}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{F}}} && \text{(Using } \mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\mathcal{G}}} = \mathbf{P}_{\tilde{\mathcal{N}}}\text{)} \\
\mathbf{P}_{\mathcal{N}}\mathbf{M} &= (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}} && \text{(Using } (\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{N}}} = (\mathbf{N}^{\top})^{+}\text{)} \\
\mathbf{P}_{\Gamma}\mathbf{P}_{\mathcal{N}}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}} &= \mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\,. && (204)
\end{aligned}$$

where in the last line we multiplied on the left by the orthogonal projector $\mathbf{P}_{\Gamma}$ and used that $\mathbf{M} = \mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}$. We can now show that we can substitute Equation (204) into the left-hand side of Equation (202):

$$\begin{aligned}
\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+}\mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}} && (205) \\
\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{f}}(\mathbf{s}) + \mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}} && \text{(Use } \mathbf{P}_{\tilde{\mathcal{F}}}\tilde{\mathbf{f}} = \tilde{\mathbf{f}}\text{)} \\
(\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+})^{+}\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) &= (\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+})^{+}\mathbf{P}_{\Gamma}\mathbf{A}_{\mathbf{q}}\mathbf{M}\tilde{\mathbf{f}}(\mathbf{s}) + (\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+})^{+}\mathbf{P}_{\Gamma}\mathbf{a}_{\mathbf{q}} \\
& && \text{(Multiply on the left by } (\mathbf{P}_{\Gamma}(\mathbf{N}^{\top})^{+})^{+}\text{)} \\
\mathbf{P}_{\tilde{\Gamma}}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) &= \tilde{\mathbf{A}}_{\mathbf{q}}\tilde{\mathbf{f}}(\mathbf{s}) + \tilde{\mathbf{a}}_{\mathbf{q}}\,, && (206)
\end{aligned}$$

where we denoted with $\mathbf{P}_{\tilde{\Gamma}}$ the orthogonal projector on $\tilde{\Gamma}$ and we defined:

$$\mathbf{P}_{\tilde{\Gamma}} := (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma(\mathbf{N}^\top)^+ \tag{207}$$
$$\tilde{\mathbf{A}}_\mathbf{q} := (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{M} \tag{208}$$
$$\tilde{\mathbf{a}}_\mathbf{q} := (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma\mathbf{a}_\mathbf{q}\,. \tag{209}$$

Multiplying by $\mathbf{P}_{\tilde{\Gamma}}$ we arrive at the same expression for linearity for the model $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$:

$$\mathbf{P}_{\tilde{\Gamma}}\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_{\tilde{\Gamma}}\tilde{\mathbf{A}}_\mathbf{q}\tilde{\mathbf{f}} + \mathbf{P}_{\tilde{\Gamma}}\tilde{\mathbf{a}}_\mathbf{q}\,. \tag{210}$$

Step 2. We proceed to show that $\tilde{\Gamma} \subseteq \tilde{\mathcal{N}}$. First, we have to show that $\mathbf{P}_{\tilde{\Gamma}}\mathbf{P}_{\tilde{\mathcal{N}}} = \mathbf{P}_{\tilde{\Gamma}}$. To this end, notice that $(\mathbf{N}^\top)^+ : \tilde{\mathcal{N}} \to \mathcal{N}$. Therefore, we have from Equation (207):

$$\begin{aligned}
\mathbf{P}_{\tilde{\Gamma}} &= (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma(\mathbf{N}^\top)^+ && (211)\\
&= (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma(\mathbf{N}^\top)^+\mathbf{P}_{\tilde{\mathcal{N}}} && (\text{Using } (\mathbf{N}^\top)^+ = (\mathbf{N}^\top)^+\mathbf{P}_{\tilde{\mathcal{N}}}) \\
&= \mathbf{P}_{\tilde{\Gamma}}\mathbf{P}_{\tilde{\mathcal{N}}}\,. && (212)
\end{aligned}$$

Taking the transpose of $\mathbf{P}_{\tilde{\Gamma}}$ we obtain:

$$\mathbf{P}_{\tilde{\Gamma}}^\top = (\mathbf{P}_{\tilde{\Gamma}}\mathbf{P}_{\tilde{\mathcal{N}}})^\top \tag{213}$$
$$\mathbf{P}_{\tilde{\Gamma}} = \mathbf{P}_{\tilde{\mathcal{N}}}^\top\mathbf{P}_{\tilde{\Gamma}}^\top \tag{214}$$
$$\mathbf{P}_{\tilde{\Gamma}} = \mathbf{P}_{\tilde{\mathcal{N}}}\mathbf{P}_{\tilde{\Gamma}} \tag{215}$$

where we used that $\mathbf{P}_{\tilde{\mathcal{N}}}^\top = \mathbf{P}_{\tilde{\mathcal{N}}}$ and $\mathbf{P}_{\tilde{\Gamma}}^\top = \mathbf{P}_{\tilde{\Gamma}}$ because both are symmetric matrices. This means, in turn, that $\mathbf{P}_{\tilde{\mathcal{N}}}$ and $\mathbf{P}_{\tilde{\Gamma}}$ commute, and so $\tilde{\Gamma}$ must be a contained in $\tilde{\mathcal{N}}$. Similarly from the expression of $\tilde{\mathbf{A}}_\mathbf{q}$ we get:

$$\begin{aligned}
\tilde{\mathbf{A}}_\mathbf{q} &= (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{M} && (216)\\
&= (\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^+\mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{M}\mathbf{P}_{\tilde{\mathcal{M}}} && (217)\\
&= \tilde{\mathbf{A}}_\mathbf{q}\mathbf{P}_{\tilde{\mathcal{M}}}\,. && (218)
\end{aligned}$$

This means that $\tilde{\mathbf{A}}_\mathbf{q}^\top\mathbf{P}_{\tilde{\Gamma}} = \mathbf{P}_{\tilde{\mathcal{M}}}\tilde{\mathbf{A}}_\mathbf{q}^\top\mathbf{P}_{\tilde{\Gamma}}$, and so $\Gamma_\mathbf{q} := \mathrm{Im}(\tilde{\mathbf{A}}_\mathbf{q}^\top\mathbf{P}_{\tilde{\Gamma}}) \subseteq \tilde{\mathcal{M}}$. To find the expression for $\tilde{\Gamma}$, we consider the following:

$$\begin{aligned}
\tilde{\Gamma} &= \ker(\mathbf{P}_\Gamma(\mathbf{N}^\top)^+)^\perp && (219)\\
&= \mathrm{Im}(\mathbf{N}^+\mathbf{P}_\Gamma^\top) && (\ker(\mathbf{A})^\perp = \mathrm{Im}(\mathbf{A}^\top), \text{ for any matrix } \mathbf{A} \text{ [Axler, 2015]}) \\
&= \mathrm{Im}(\mathbf{N}^+\mathbf{P}_\Gamma)\,. && (220)
\end{aligned}$$

This proves the claim. □

What happens if $\Gamma_\mathbf{q} \not\subseteq \mathcal{M}$?. We discuss the case when the condition in Theorem 14 is not met due to $\Gamma_\mathbf{q} := \mathrm{Im}(\mathbf{A}_\mathbf{q}^\top\mathbf{P}_\Gamma) \not\subseteq \mathcal{M}$. We show that even if a model $(\mathbf{f}, \mathbf{g})$ linearly represents $\mathbf{q}$ on Γ, a $\sim_{EL}$-equivalent model $(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ may not. This is due to the fact that the information contained in $\mathcal{F} \setminus \mathcal{M}$, used to relationally represent $\mathbf{q}$, may not be linearly transformed on another $\sim_{EL}$-equivalent model.

In fact, when $\Gamma_\mathbf{q} \not\subseteq \mathcal{M}$, consider the expression for relational linearity given by Definition 9 where

$$\mathbf{P}_\Gamma\mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{f}(\mathbf{s}) + \mathbf{P}_\Gamma\mathbf{a}_\mathbf{q}\,. \tag{221}$$

We take the first term on the RHS of Equation (221): by inserting the projector $\mathbf{P}_{\Gamma_\mathbf{q}}$, we rewrite it as follows:

$$\mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{P}_{\Gamma_\mathbf{q}}\mathbf{f}(\mathbf{s}) = \mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{P}_{\Gamma_\mathbf{q}}\mathbf{P}_\mathcal{M}\mathbf{f}(\mathbf{s}) + \mathbf{P}_\Gamma\mathbf{A}_\mathbf{q}\mathbf{P}_{\Gamma_\mathbf{q}}(\mathbf{I} - \mathbf{P}_\mathcal{M})\mathbf{f}(\mathbf{s})\,, \tag{222}$$

where we used the identity $\mathbf{I} = \mathbf{P}_{\mathcal{M}} + (\mathbf{I} - \mathbf{P}_{\mathcal{M}})$ to separate the contributions, inside and outside $\mathcal{M}$. We can thus rewrite Equation (221) as:

$$\begin{aligned}
\mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} \mathbf{P}_{\mathcal{M}} \mathbf{f}(\mathbf{s}) + \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{P}_{\mathcal{F}} \mathbf{f}(\mathbf{s}) + \mathbf{P}_\Gamma \mathbf{a}_\mathbf{q} && \text{(Introduce } \mathbf{P}_{\mathcal{F}} \mathbf{f} = \mathbf{f}) \\
&\propto \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{P}_{\mathcal{F}} \mathbf{f}(\mathbf{s}) && (\mathbf{P}_{\Gamma_\mathbf{q}} \mathbf{P}_{\mathcal{M}} = \mathbf{0}) \\
&= \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}} \mathbf{P}_{\mathcal{F}}) \mathbf{f}(\mathbf{s}) && \text{(Multiplying } (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}} \mathbf{P}_{\mathcal{F}}) \\
&= \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \mathbf{f}(\mathbf{s}) && (223)
\end{aligned}$$

where the in the last equation we used that $\mathbf{P}_{\mathcal{M}} \mathbf{P}_{\mathcal{F}} = \mathbf{P}_{\mathcal{M}}$ by Lemma 2 (v). To show that this can lead to non-linearities implying deviations from relational linearity, suppose that:

$$\begin{aligned}
\mathbf{f}(\mathbf{s}) &= \mathbf{P}_{\mathcal{M}} \mathbf{f}(\mathbf{s}) + (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \mathbf{f}(\mathbf{s}) && (224) \\
&= \mathbf{M} \mathbf{P}_{\tilde{\mathcal{M}}} \tilde{\mathbf{f}}(\mathbf{s}) + (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \tilde{\mathbf{f}}^2(\mathbf{x}) && (225)
\end{aligned}$$

where the first term follows from the equivalence relation Equation (9), *i.e.*, $\mathbf{P}_{\mathcal{M}} \mathbf{f}(\mathbf{s}) = \mathbf{M} \mathbf{P}_{\tilde{\mathcal{M}}} \tilde{\mathbf{f}}(\mathbf{s})$, and we used $\tilde{\mathbf{f}}^2(\mathbf{x}) = (\tilde{f}_1(\mathbf{x})^2, \ldots, \tilde{f}_{\tilde{d}}(\mathbf{x})^2)$ to denote the square of the components of $\tilde{\mathbf{f}}$. Notice that, this choice is allowed since the components of $\mathbf{f}$ outside $\mathcal{M}$, *i.e.*, those in $\mathcal{F} \setminus \mathcal{M}$, can be arbitrarily chosen and do not contribute to the dot-product with $\mathbf{g}_0$. Therefore, substituting this expression to Equation (223) we get:

$$\mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) \propto \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \tilde{\mathbf{f}}^2(\mathbf{s}) \tag{226}$$

and substituting $\mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_\Gamma \mathbf{N}^+ \tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q})$, implied by the LHS of Equation (201) and from the equality in Equation (204), we have that:

$$\mathbf{P}_\Gamma \mathbf{N}^+ \tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q}) \propto \mathbf{P}_\Gamma \mathbf{A}_\mathbf{q} \mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \tilde{\mathbf{f}}^2(\mathbf{s}) \,, \tag{227}$$

which shows a non-linear dependence of $\tilde{\mathbf{f}}(\mathbf{s} \frown \mathbf{q})$ on $\tilde{\mathbf{f}}(\mathbf{s})$, invalidating relational linearity when $\mathbf{P}_{\Gamma_\mathbf{q}} (\mathbf{P}_{\mathcal{F}} - \mathbf{P}_{\mathcal{M}}) \neq \mathbf{0}$.

D.2 Proof of Theorem 15

Theorem 15. *For two models $(\mathbf{f}, \mathbf{g}), (\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \in \Theta$, such that $(\mathbf{f}, \mathbf{g}) \sim_{EL} (\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, the vectors $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \mathrm{SIm}(\mathbf{g}_0)$ are parallel within $\mathcal{N}$ if and only if the corresponding vectors $\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\gamma}}' \in \mathrm{SIm}(\tilde{\mathbf{g}}_0)$ are parallel in $\tilde{\mathcal{N}}$.*

Proof. Given that $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ are parallel in $\mathcal{N}$ (Definition 7), we have that:

$$\mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma} = \beta \mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}' \,, \tag{228}$$

where $\beta \neq 0$ is given by $\beta = \|\mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}\| / \|\mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}'\|$ (see also Remark 20. We consider the components of the $\sim_{EL}$-equivalent model, given by

$$\mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma} = \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}, \quad \mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}' = \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' \,. \tag{229}$$

Using this in Equation (228) we get:

$$\mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}} = \beta \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' \tag{230}$$

and multiplying from the left by the pseudoinverse of $\mathbf{N}$ we get:

$$\begin{aligned}
\mathbf{N}^+ \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}} &= \beta \mathbf{N}^+ \mathbf{N} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' && (231) \\
\mathbf{P}_{\tilde{\mathcal{N}}} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}} &= \beta \mathbf{P}_{\tilde{\mathcal{N}}} \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' && (\text{Using } \mathbf{N}^+ \mathbf{N} = \mathbf{P}_{\tilde{\mathcal{N}}}) \\
\mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}} &= \beta \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' && (232)
\end{aligned}$$

which shows that $\tilde{\boldsymbol{\gamma}}$ is parallel to $\tilde{\boldsymbol{\gamma}}'$ in $\tilde{\mathcal{N}}$. To prove the reverse implication, the same steps can be repeated by symmetry, taking:

$$\mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}} = \mathbf{N}^+ \mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}, \quad \mathbf{P}_{\tilde{\mathcal{N}}} \tilde{\boldsymbol{\gamma}}' = \mathbf{N}^+ \mathbf{P}_{\mathcal{N}} \boldsymbol{\gamma}' \,. \tag{233}$$

for two vectors $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}}'$ parallel in $\tilde{\mathcal{N}}$, giving the desired result. □

E Context-query-reply sentences: Corner cases

E.1 Paraphrases

We consider a sentence $\mathbf{q}_2$ to be the paraphrase of $\mathbf{q}_1$ when $\mathbf{q}_2$ repeats what was written in $\mathbf{q}_1$ using different words (for this definition, we only slightly adapted the one from the Cambridge Dictionary). To provide an example, let $\mathbf{q}_1$ = *"Is the text written in English?"*. A paraphrase of $\mathbf{q}_1$ can be $\mathbf{q}_2$ = *"Was the previous text written in English or not?"*. These two equivalent formulations of the same question can nonetheless be treated differently by a next-token predictor: for example, given a string $\mathbf{s}$, it can be that $\mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) \neq \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2)$.

Here, we analyze how paraphrastic aspects of textual data can be described with relational context-query-reply $(\mathbf{s} \frown \mathbf{q} \frown y)$ strings for a model $(\mathbf{f}, \mathbf{g}) \in \Theta$. We start by providing a tentative definition of paraphrastic sentences in terms of their entailed conditional log-probabilities for different pairs of next-tokens.

Definition 21 (Paraphrases)**.** *We say that $\mathbf{q}_2 \in \mathsf{Seq}(\mathcal{A})$ on $\mathcal{Y}_2 \subseteq \mathcal{A}$ is a paraphrase of $\mathbf{q}_1 \in \mathsf{Seq}(\mathcal{A})$ on $\mathcal{Y}_1 \subseteq \mathcal{A}$ for the model $(\mathbf{f}, \mathbf{g}) \in \Theta$ if (1) there exists $\beta \neq 0$ such that, for all $y_0, y_1 \in \mathcal{Y}_1 \subseteq \mathcal{A}$, and (2) there exist $\hat{y}_0, \hat{y}_1 \in \mathcal{Y}_2 \subseteq \mathcal{A}$, for which it holds:*

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s} \frown \mathbf{q}_1)}{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s} \frown \mathbf{q}_1)} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(\hat{y}_0 \mid \mathbf{s} \frown \mathbf{q}_2)}{p_{\mathbf{f},\mathbf{g}}(\hat{y}_1 \mid \mathbf{s} \frown \mathbf{q}_2)} \,. \tag{234}$$

For example, consider the strings $\mathbf{q}_1$ = *"Is the text written in English?"*, with expected replies y_0 = *"yes"* and y_1 = *"no"*, *i.e.*, $y_0, y_1 \in \mathcal{Y}_1$; and a second string $\mathbf{q}_2$ = *"Reply with only A or B. Was the text written in (A) English or (B) not ?"*, with expected replies $\hat{y}_0$ = *"A"* and $\hat{y}_1$ = *"B"*. Definition 21 entails that, for all input-strings $\mathbf{s}$, the concatenation to the query $\mathbf{q}_1$ gives a ratio of the log-probabilities of y_0 and y_1 that matches, up to a constant β, that of $\hat{y}_0$ and $\hat{y}_1$ for the concatenation to $\mathbf{q}_2$. A model that successfully recognizes between English and non-English text and considers $\mathbf{q}_2$ a paraphrase of $\mathbf{q}_1$, then will attribute similar conditional probabilities to both $p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s} \frown \mathbf{q})$ and $p_{\mathbf{f},\mathbf{g}}(\hat{y}_0 \mid \mathbf{s} \frown \mathbf{q})$, for any input-context $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$.

We show that sentences and next-tokens as in Definition 21 induce a specific structure in the model embeddings, linearly relating the representations $\mathbf{f}(\mathbf{s} \frown \mathbf{q}_1)$ and $\mathbf{f}(\mathbf{s} \frown \mathbf{q}_2)$. To this end, we will define $\mathrm{SIm}(\mathbf{g}_0)_{\mathcal{Y}} := \mathrm{span}\{\mathbf{g}(y) - \mathbf{g}(y_0) \mid \; y_0, y \in \mathcal{Y}\}$ and $\mathrm{SIm}(\mathbf{f})_{\mathbf{q}} := \mathrm{span}\{\mathbf{f}(\mathbf{s} \frown \mathbf{q}) \mid \mathbf{s} \in \mathsf{Seq}(\mathcal{A})\}$.

Proposition 22. *If (1) $\mathbf{q}_2 \in \mathsf{Seq}(\mathcal{A})$ on $\mathcal{Y}_2 \subseteq \mathcal{A}$ is a paraphrase of $\mathbf{q}_1 \in \mathsf{Seq}(\mathcal{A})$ on $\mathcal{Y}_1 \subseteq \mathcal{A}$ for the model $(\mathbf{f}, \mathbf{g})$ and (2) for the subspaces $\Gamma_1 := \mathrm{SIm}(\mathbf{g}_0)_{\mathcal{Y}_1}$ and $\Gamma_2 := \mathrm{SIm}(\mathbf{g}_0)_{\mathcal{Y}_2}$, it holds that $\Gamma_1 \subseteq \mathrm{SIm}(\mathbf{f})_{\mathbf{q}_1} =: \mathcal{F}_1$ and $\Gamma_2 \subseteq \mathrm{SIm}(\mathbf{f})_{\mathbf{q}_2} =: \mathcal{F}_2$, then $\dim(\Gamma_1) = \dim(\Gamma_2)$ and there exists a matrix $\mathbf{O} \in \mathbb{R}^{d \times d}$ that defines a linear, invertible transformation from Γ_2 to Γ_1 such that*

$$\mathbf{P}_{\Gamma_1} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) = \beta \mathbf{O} \mathbf{P}_{\Gamma_2} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) \,. \tag{235}$$

Proof. We start with the equality between logs of probabilities given by Definition 21:

$$\log \frac{p_{\mathbf{f},\mathbf{g}}(y_1 \mid \mathbf{s} \frown \mathbf{q}_1)}{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s} \frown \mathbf{q}_1)} = \beta \cdot \log \frac{p_{\mathbf{f},\mathbf{g}}(\hat{y}_1 \mid \mathbf{s} \frown \mathbf{q}_2)}{p_{\mathbf{f},\mathbf{g}}(\hat{y}_0 \mid \mathbf{s} \frown \mathbf{q}_2)} \tag{236}$$

$$\log \exp\left((\mathbf{g}(y_1) - \mathbf{g}(y_0))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1)\right) = \beta \log \exp\left((\mathbf{g}(\hat{y}_1) - \mathbf{g}(\hat{y}_0))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2)\right)$$

(Substituting Equation (1) for the conditional probabilities)

$$(\mathbf{g}(y_1) - \mathbf{g}(y_0))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) = \beta (\mathbf{g}(\hat{y}_1) - \mathbf{g}(\hat{y}_0))^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) \,, \tag{237}$$

where in the last line we removed the logarithm of the exponential on both sides. Define $\mathbf{g}_0(y) := \mathbf{g}(y) - \mathbf{g}(y_0)$ for a pivot $y_0 \in \mathcal{Y}_1$ and define $\hat{\mathbf{g}}_0(\hat{y}) := \mathbf{g}(\hat{y}) - \mathbf{g}(\hat{y}_0)$ for the corresponding pivot $\hat{y}_0 \in \mathcal{Y}_2$. Then consider q elements $y_i \in \mathcal{Y}_1$ and their correspondents $\hat{y}_i \in \mathcal{Y}_2$ such that the matrices:

$$\mathbf{G} := \left(\mathbf{g}_0(y_1), \ldots, \mathbf{g}_0(y_q)\right) \,, \quad \hat{\mathbf{G}} := \left(\hat{\mathbf{g}}_0(\hat{y}_1), \ldots, \hat{\mathbf{g}}_0(\hat{y}_q)\right) \tag{238}$$

have rank equal to $\dim(\Gamma_1)$ and $\dim(\Gamma_2)$, respectively. Then, we make use of these matrices with their transpose in Equation (237), obtaining:

$$\mathbf{G}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) = \beta \hat{\mathbf{G}}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) \tag{239}$$

$$(\mathbf{G}^\top)^+ \mathbf{G}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) = \beta (\mathbf{G}^\top)^+ \hat{\mathbf{G}}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) \qquad \text{(Multiplying on the left for } (\mathbf{G}^\top)^+\text{)}$$

$$\mathbf{P}_{\Gamma_1} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) = \beta \mathbf{O} \mathbf{P}_{\Gamma_2} \mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) \tag{240}$$

where in the last line we used $(\mathbf{G}^\top)^+\mathbf{G}^\top = \mathbf{P}_{\mathrm{Im}(\mathbf{G})} = \mathbf{P}_{\Gamma_1}$, since $\mathbf{G}$ in Equation (238) spans Γ_1, and we defined $\mathbf{O} := (\mathbf{G}^\top)^+\hat{\mathbf{G}}^\top$. Proceeding similarly but from $\hat{\mathbf{G}}^\top\mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) = \mathbf{G}^\top\mathbf{f}(\mathbf{s} \frown \mathbf{q}_1)$ we obtain a similar equation to Equation (240) :

$$\mathbf{P}_{\Gamma_2}\mathbf{f}(\mathbf{s} \frown \mathbf{q}_2) = \beta^{-1}\hat{\mathbf{O}}\mathbf{P}_{\Gamma_1}\mathbf{f}(\mathbf{s} \frown \mathbf{q}_1) \tag{241}$$

where we defined $\hat{\mathbf{O}} := (\hat{\mathbf{G}}^\top)^+\mathbf{G}^\top$.

Let $k_1 := \dim(\Gamma_1)$, $k_2 := \dim(\Gamma_2)$, and $o := \mathrm{rank}(\mathbf{O})$. Notice that, by definition, $\mathbf{O} = \mathbf{P}_{\Gamma_1}\mathbf{O}$, which means that:

$$\begin{aligned}
\mathrm{rank}(\mathbf{O}) &= \mathrm{rank}(\mathbf{P}_{\Gamma_1}\mathbf{O}) && (242)\\
o &\leq \min(k_1, o) && (\text{Using } \mathrm{rank}(\mathbf{AB}) \leq \min\big(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B})\big)) \\
\implies o &\leq k_1 && (243)
\end{aligned}$$

Now, consider ℓ points, or sequences, $\mathbf{x}_i \in \mathsf{Seq}(\mathcal{A})$ such that:

$$\mathbf{F}_1 := \big(\mathbf{f}(\mathbf{x}_1 \frown \mathbf{q}_1), \ldots, \mathbf{f}(\mathbf{x}_\ell \frown \mathbf{q}_1)\big)\,, \quad \mathbf{F}_2 := \big(\mathbf{f}(\mathbf{x}_1 \frown \mathbf{q}_2), \ldots, \mathbf{f}(\mathbf{x}_\ell \frown \mathbf{q}_2)\big)\,, \tag{244}$$

have rank equal to $\dim(\mathcal{F}_1)$ and to $\dim(\mathcal{F}_2)$, respectively. Therefore,by substituting $\mathbf{F}_1$ and $\mathbf{F}_2$ in Equation (240) and in Equation (241), we obtain:

$$\begin{aligned}
\mathbf{P}_{\Gamma_1}\mathbf{F}_1 &= \beta\mathbf{O}\mathbf{P}_{\Gamma_2}\mathbf{F}_2 && (245)\\
\mathbf{P}_{\Gamma_2}\mathbf{F}_2 &= \beta^{-1}\hat{\mathbf{O}}\mathbf{P}_{\Gamma_1}\mathbf{F}_1\,. && (246)
\end{aligned}$$

Then, by Lemma 17 it holds that:

$$\mathrm{rank}(\mathbf{P}_{\Gamma_1}\mathbf{F}_1) = \dim(\Gamma_1) = k_1, \quad \mathrm{rank}(\mathbf{P}_{\Gamma_2}\mathbf{F}_2) = \dim(\Gamma_2) = k_2\,. \tag{247}$$

We use this in Equation (245) to obtain the following:

$$\begin{aligned}
\mathrm{rank}(\mathbf{P}_{\Gamma_1}\mathbf{F}_1) &= \mathrm{rank}(\beta\mathbf{O}\mathbf{P}_{\Gamma_2}\mathbf{F}_2) && (248)\\
k_1 &\leq \min\big(\mathrm{rank}(\mathbf{O}), \mathrm{rank}(\mathbf{P}_{\Gamma_2}\mathbf{F}_2)\big) && (\text{Using } \mathrm{rank}(\mathbf{AB}) \leq \min\big(\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\mathbf{B})\big)) \\
k_1 &\leq \min(o, k_2) && (249)\\
\implies o \geq k_1, &\quad k_2 \geq k_1\,. && (250)
\end{aligned}$$

This shows that $k_1 \leq o \leq k_1$, meaning that $\mathrm{rank}(\mathbf{O}) = \dim(\Gamma_1)$. Similarly, taking the rank of Equation (246) we obtain a similar implication for k_1 and k_2 (as obtained for k_2 and k_1 for Equation (245)), that is:

$$\begin{aligned}
\mathrm{rank}(\mathbf{P}_{\Gamma_2}\mathbf{F}_2) &= \mathrm{rank}(\beta^{-1}\hat{\mathbf{O}}\mathbf{P}_{\Gamma_1}\mathbf{F}_1) && (251)\\
&\implies k_2 \leq k_1\,, && (252)
\end{aligned}$$

which shows that $k_1 \leq k_2 \leq k_1$, meaning that $\dim(\Gamma_2) = \dim(\Gamma_1)$. Moreover, the matrix $\mathbf{O}$, being of rank $o = k_1 = k_2$ defines an invertible transformation from Γ_2 to Γ_1. This concludes the proof. □

E.2 Tautologies

Next, we investigate how tautological aspects can be encoded by a model. A tautology in our context can be considered as a context-independent sentence, whose reply to it is not influenced by the previous context. For example, we can consider as $\mathbf{q}$ = *"No matter what was written before. Whatever follows reply with 42!"* as a tautology. These strings constitute peculiar cases in natural language where the previous input context does not influence the replies to the query $\mathbf{q}$. Formally:

Definition 23 (Tautology)**.** *We say that* $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ *is a tautology for the model* $(\mathbf{f}, \mathbf{g}) \in \Theta$ *if, for every* $\mathbf{s} \in \mathsf{Seq}(\mathcal{A})$ *and all* $y \in \mathcal{A}$*, it holds:*

$$\log p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{s} \frown \mathbf{q}) = \log p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{q})\,. \tag{253}$$

Next, we show that, for such tautologies, a "trivial" form of linear relational embedding holds.

Proposition 24 (Tautologies)**.** *Let* $\mathbf{q} \in \mathsf{Seq}(\mathcal{A})$ *be a tautology for the model* $(\mathbf{f}, \mathbf{g})$*, then* $\mathbf{f}$ *linearly represents* $\mathbf{q}$ *on* $\mathcal{G} := \mathrm{SIm}(\mathbf{g}_0)$ *with:*

$$\mathbf{P}_{\mathcal{G}} \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_{\mathcal{G}} \mathbf{a}_{\mathbf{q}} \,. \tag{254}$$

Proof. From Definition 23, consider a pivot token $y_0 \in \mathcal{A}$ and define $\mathbf{g}_0 := \mathbf{g}(y) - \mathbf{g}(y_0)$ for every $y \in \mathcal{A}$. We have that:

$$\begin{aligned}
\log \frac{p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{s} \frown \mathbf{q})}{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{s} \frown \mathbf{q})} &= \log \frac{p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{q})}{p_{\mathbf{f},\mathbf{g}}(y_0 \mid \mathbf{q})} && \text{(Take log of the ratio between } p_{\mathbf{f},\mathbf{g}}(y \mid \cdot) \text{ and } p_{\mathbf{f},\mathbf{g}}(y_0 \mid \cdot)) \\
\log \frac{\exp \mathbf{g}(y)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})}{\exp \mathbf{g}(y_0)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q})} &= \log \frac{\exp \mathbf{g}(y)^\top \mathbf{f}(\mathbf{q})}{\exp \mathbf{g}(y_0)^\top \mathbf{f}(\mathbf{q})} && \text{(Write with the exponentional)} \\
\log \exp \mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \log \exp \mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{q}) && \text{(Use the definition of } \mathbf{g}_0) \\
\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{q}) && \text{(Remove log and exponential)} \\
\mathbf{g}_0(y)^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{g}_0(y)^\top \mathbf{a}_{\mathbf{q}} \,, && (255)
\end{aligned}$$

where in the last line with denoted with $\mathbf{a}_{\mathbf{q}} := \mathbf{f}(\mathbf{q})$. Consider ℓ tokens $y_i \in \mathcal{A}$ such that

$$\mathbf{G} := \big(\mathbf{g}_0(y_1), \dots, \mathbf{g}_0(y_\ell)\big) \tag{256}$$

spans $\mathcal{G}$. We use this and consider the following expression ofr the transpose:

$$\begin{aligned}
\mathbf{G}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{G}^\top \mathbf{a}_{\mathbf{q}} && (257) \\
(\mathbf{G}^\top)^+ \mathbf{G}^\top \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= (\mathbf{G}^\top)^+ \mathbf{G}^\top \mathbf{a}_{\mathbf{q}} && \text{(Multiply by pseudo-inverse of } \mathbf{G}^\top) \\
\mathbf{P}_{\mathcal{G}} \mathbf{f}(\mathbf{s} \frown \mathbf{q}) &= \mathbf{P}_{\mathcal{G}} \mathbf{a}_{\mathbf{q}} && (258)
\end{aligned}$$

where we used the fact that $(\mathbf{G}^\top)^+ \mathbf{G}^\top = \mathbf{P}_{\mathcal{G}}$. This shows the claim. □