Relational Linear Properties in Language Models: An Empirical Investigation

Giovanni Valer [1 2 3] **Luigi Gresele** [4] **Marco Bronzini** [*1 5] **Emanuele Marconato** [*5]

Abstract

Linear properties are ubiquitous in the representations of language models; however, testing them experimentally remains a challenging task. This work focuses on *relational linearity*: the hypothesis that, for a fixed relation (*e.g.*, *"plays"*), the unembedding of an object (*e.g.*, *"trumpet"*) can be predicted from the embedding of its subject (*e.g.*, *"Miles Davis"*) by a linear map. We present an experimental method to test the formulation of relational linearity by Marconato et al. (2025). Specifically, we introduce a probing method, based on Kullback-Leibler divergence, to evaluate this property and examine its variation across layers and paraphrased relational queries. It is also more efficient than previous work; for example, it avoids the crude Jacobian approximations used in Linear Relational Embeddings by Hernandez et al. (2024). Our findings across four datasets show that relational linearity varies across models, exhibits layer-wise patterns consistent with prior observations about linguistic information in model representations, and is differently affected by changes in how the relation is phrased.

1. Introduction

How do large language models encode knowledge about the world, specifically about its entities and their relations, in their latent representations? A growing line of work approaches this question by studying whether semantic properties of entities and relations are reflected in linear structure in the representation spaces of neural language models (Mikolov et al., 2013; Elhage et al., 2022; Marks & Tegmark, 2023; Park et al., 2023; Engels et al., 2024; Wu et al., 2025). Such linear structure has been argued to support both the interpretation of a model's internal computations (Geiger et al., 2021) and the control of model behavior through activation steering (Marks & Tegmark, 2023; Turner et al., 2023; Wu et al., 2024).[1]

Relational linearity in language models (Paccanaro & Hinton, 2002; Hernandez et al., 2024; Marconato et al., 2025) investigates whether relations between text sequences are encoded linearly in model representations. In subject-relation-object triples, or in context-query-answer triples, this means asking whether the object/answer associated with a relation/query can be recovered by a linear map from the representation of the subject/context alone. For example, for the subject *"Jimi Hendrix"* and the relation *"played at"*, the question is whether the unembedding of *"Woodstock"* (object) can be predicted linearly from the representation of *"Jimi Hendrix"* (subject), without appending the relation to the subject. While relational linearity was detected in small datasets with annotated triplets and by resorting to linear expansions of transformer computations (Hernandez et al., 2024; Chanin et al., 2024), an extensive experimental evaluation of this property is currently missing.

We contribute to fill this gap and further investigate relational linearity in large language models, analyzing the degree to which it arises in their latent representations using a novel probing method based on Kullback-Leibler (KL) divergence. *Our main contribution is to introduce an experimental procedure that permits evaluating relational linear properties*, extending Marconato et al. (2025). Specifically, our probing procedure only requires accessing the model's latent representations both when it is prompted with (a) the context and the query of interest, and with (b) the context only. With this, we investigate the presence of relational linearity in language models through a *novel KL-based linear probe* (`KL-RP`), and compare it with an adaptation of the Linear Relational Embedding (`LRE`) method proposed by Hernandez et al. (2024). Computationally, our approach avoids crude Jacobian approximations of previous

[*]Shared last author [1]Fondazione Bruno Kessler, Trento, Italy [2]University of Bologna, Bologna, Italy [3]University of Pisa, Pisa, Italy [4]University of Copenhagen, Copenhagen, Denmark [5]University of Trento, Trento, Italy. Correspondence to: Giovanni Valer <gvaler@fbk.eu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Appendix A expands the discussion of these linear properties and reviews prior work on linear and relational properties of language models, as well as layer-wise interpretability.

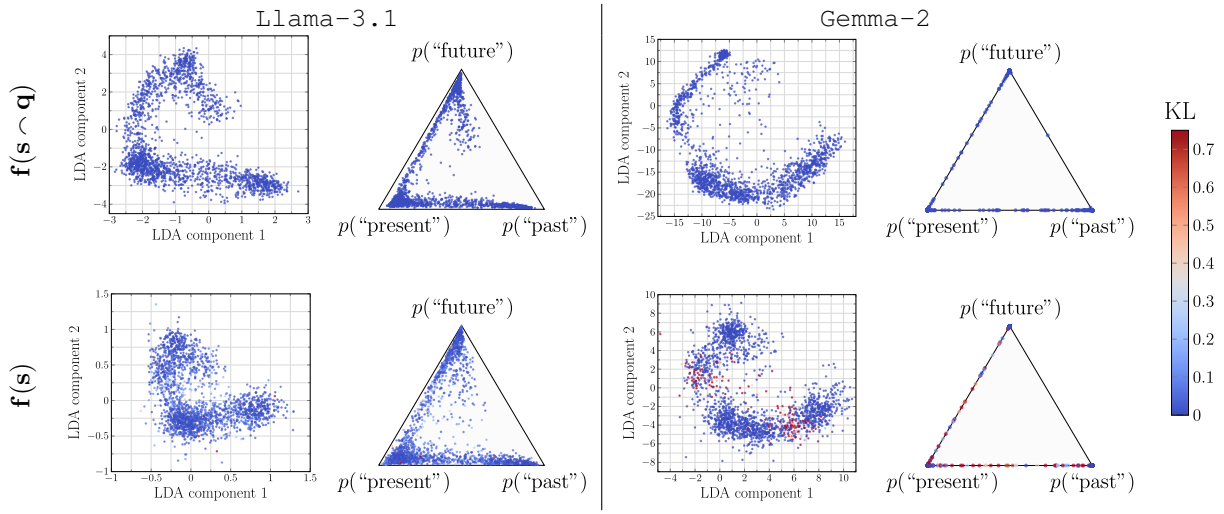


Figure 1. **Relational linearity of final-layer representations for the TENSE dataset**. For `Llama-3.1` (two left columns) and `Gemma-2` (two right columns), we compare the embeddings computed from contexts concatenated with the query $\mathbf{q}$ = *"What is the tense of the previous sentence?"* $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$, with embeddings obtained by KL-based Relational Probe (`KL-RP`) on the context-only embeddings $\mathbf{f}(\mathbf{s})$, for contexts $\mathbf{s}$ from the TENSE dataset (Ayman et al., 2024). The top row shows the reference embeddings $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$, while the bottom row shows the corresponding `KL-RP` embeddings computed from $\mathbf{f}(\mathbf{s})$. For each model, the left panels show two-dim LDA projections of the embeddings, and the right panels show the next-token probabilities for *"present"*, *"past"*, and *"future"* on the probability simplex. Points are coloured by their KL divergence to the reference next-token distribution induced by $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$, with reference KL set to zero and values above 0.75 clipped for visualization. For `Llama-3.1`, `KL-RP` largely preserves both the embedding geometry and the induced next-token distributions, yielding $d_{\mathrm{KL}} \approx 0.02$ on the test set. For `Gemma-2`, the relation to the reference embedding geometry is weaker, and the induced next-token distributions differ more from the reference, yielding $d_{\mathrm{KL}} \approx 0.23$. Overall, the contrast between the top and bottom rows indicates whether the target representations and distribution (top) can be recovered from the unprompted model (bottom), giving relational linearity: this is highly possible for `Llama-3.1`, but less so for `Gemma-2`.

works (Hernandez et al., 2024) and is thus more efficient. In experiments, our methods uncovers relational linear properties in language model representations that are missed by previous approaches, as we show in Section 4.3.

We perform experiments using two language models: a medium-sized model (Llama-3.1, Grattafiori et al., 2024) with 8B parameters, and a smaller 2B-parameter model (Gemma 2, Gemma Team et al., 2024); and we study their internal activations with different sizes: 4096 for Llama, and 2304 for Gemma. Our experimental analysis reveals the following:

a) language models encode relational properties in a largely linear manner, especially for tense (Fig. 1) and truthfulness, with model-dependent or weaker effects for language and subjectivity (Q1; Section 4.3);

b) relational linearity varies across layers, with surface features emerging in early layers and more abstract properties peaking in middle layers (Q2; Section 4.4);

c) changing the way the query is phrased produces varied effects on linear probing (Q3; Section 4.5).

2. Preliminaries

We denote by $\mathcal{Y}$ a finite vocabulary of tokens. Textual inputs are elements $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is a set of finite ordered sequences of tokens from $\mathcal{Y}$. We denote the operation of concatenation of token sequences by "$\frown$"; for example, if tokens y_1, y_2, y_3 are rendered as *"A"*, *"perfect"*, and *"circle"*, then $y_1 \frown y_2 \frown y_3$ is rendered as *"A perfect circle"*.

Given a textual input $\mathbf{x} \in \mathcal{X}$, autoregressive language models compute a sequence of hidden states through $L > 1$ transformer layers. We denote by $\mathbf{f}^{(\ell)}(\mathbf{x}) \in \mathbb{R}^d$, with $\ell \in \{1, \ldots, L\}$, the hidden state at layer ℓ used to predict the next token,[2] and write $\mathbf{f} := \mathbf{f}^{(L)}$ for the final-layer *embedding*. We use the term *representation* broadly to refer to such hidden states, covering both intermediate-layer hidden states and final-layer embeddings. The model also has an output *unembedding* map $\mathbf{g} : \mathcal{Y} \to \mathbb{R}^d$, which projects hidden states into the output vocabulary by associating each token $y \in \mathcal{Y}$ with an unembedding vector. Together, $\mathbf{f}$ and $\mathbf{g}$ define the next-token distribution:

$$p_{\mathbf{f},\mathbf{g}}(y \mid \mathbf{x}) = \frac{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y'))} \,. \tag{1}$$

[2]Each transformer layer produces hidden states at all input positions; $\mathbf{f}^{(\ell)}(\mathbf{x})$ denotes the final-position hidden state used for next-token prediction. We consider only these layer outputs, not internal components such as attention heads or MLP blocks.

2.1. Relational Linearity

Following the setting by Paccanaro & Hinton (2002), we consider textual strings that can be decomposed as triplets $\mathbf{s} \frown \mathbf{q} \frown y$, where $\mathbf{s}, \mathbf{q} \in \mathcal{X}$ and $y \in \mathcal{Y}$. These can be either *subject-relation-object* triplets, *e.g.*, $\mathbf{s} =$*"David Gilmour"*, $\mathbf{q} =$*"plays the"*, and $y =$*"guitar"*, or, more generally, as *context-query-answer* triplets, *e.g.*, $\mathbf{s} =$*"Today I'm happy."*, $\mathbf{q} =$*"Is it a positive statement?"*, and $y =$*"Yes"*.[3] We will refer to $\mathbf{s}$ as to the *context* (or the *subject*), to $\mathbf{q}$ as the *query* (or the *relation*), and to y as the *answer* (or the *object*).

Relational linearity (Hernandez et al., 2024; Marconato et al., 2025) describes the property under which it is possible to linearly decode the (last-layer) representation of the context concatenated with the query, $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$, from the representation of the context, $\mathbf{f}(\mathbf{s})$, alone. We mainly follow the setup described in Marconato et al. (2025) and analyze *linear relational probing*,[4] that is, when relational linearity holds for a subset $\mathcal{Y}_P \subseteq \mathcal{Y}$ of specific tokens related to query $\mathbf{q}$. As with Hernandez et al. (2024), we extend it to all layers of language models. Formally, let the conditional distribution restricted to tokens in $\mathcal{Y}_P$ defined by

$$p(y \mid \cdot; \mathcal{Y}_P) := p(y \mid \cdot) \Big/ \textstyle\sum_{y' \in \mathcal{Y}_P} p(y' \mid \cdot) \,. \tag{2}$$

Definition 2.1 (Linear Relational Probing)**.** For a query $\mathbf{q} \in \mathcal{X}$ and a subset of tokens $\mathcal{Y}_P \subseteq \mathcal{Y}$ of dimension $k = |\mathcal{Y}_P|$, a model $(\mathbf{f}, \mathbf{g})$ can be *linearly probed for* $\mathbf{q}$ *on the subset* $\mathcal{Y}_P$ *at layer* ℓ if there exist k weights $\mathbf{w}_{\mathbf{q}}^{y} \in \mathbb{R}^d$ and biases $b_{\mathbf{q}}^{y} \in \mathbb{R}$ such that, for all $y \in \mathcal{Y}_P$ and $\forall \mathbf{s} \in \mathcal{X}$,

$$p(y \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) = \frac{\exp(\mathbf{w}_{\mathbf{q}}^{y,\top} \mathbf{f}^{(\ell)}(\mathbf{s}) + b_{\mathbf{q}}^{y})}{\sum_{y' \in \mathcal{Y}_P} \exp(\mathbf{w}_{\mathbf{q}}^{y',\top} \mathbf{f}^{(\ell)}(\mathbf{s}) + b_{\mathbf{q}}^{y'})} \,. \tag{3}$$

The following example illustrates the setting:

> *Example* 2.2. Consider a query $\mathbf{q} =$*"What is the tense of the previous sentence?"* and consider three possible answers $\mathcal{Y}_P = \{$*"past"*, *"present"*, *"future"*$\}$. Take as context $\mathbf{s} =$*"Yesterday, we watched a movie"*, to which the language model assigns high-probability to *"past"* token (with probability $1 - 2\epsilon$) and vanishing probability to *"present"* and *"future"* (both with probability ϵ) when prompted with input $\mathbf{s} \frown \mathbf{q}$.
>
> The language model can be linearly probed for $\mathbf{q}$ on $\mathcal{Y}_P$ at layer ℓ if this probability distribution can be *exactly* reconstructed from the embeddings $\mathbf{f}(\mathbf{s})$ with a linear classifier and softmax activation (according to Equation 3 and for multiple different contexts $\mathbf{s} \in \mathcal{X}$).
>
> This happens, to some degree, in `Llama-3.1` and `Gemma-2`, cf. Fig. 1.

Intuitively, Definition 2.1 ensures that, for all prompted contexts, the probability distribution on the relevant next-tokens for a query can be obtained through a linear probe of the representations of the context alone.

2.2. Relaxation of Linear Relational Probing

Similarly to Hernandez et al. (2024), we also evaluate a weaker version of linear relational probing that only considers the likeliest tokens. In this setting, we do not need to consider the full equivalence between distributions as in Equation 3, but only the match among the likeliest tokens. This specializes Equation 3 to the following (see Definition D.2 for a complete description):

$$\operatorname*{argmax}_{y \in \mathcal{Y}_P} p(y \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) = \operatorname*{argmax}_{y \in \mathcal{Y}_P} \left(\mathbf{w}_{\mathbf{q}}^{y,\top} \mathbf{f}^{(\ell)}(\mathbf{s}) + b_{\mathbf{q}}^{y} \right) . \tag{4}$$

Notice that, for linear relational probing, Equation 3 is a stronger condition than Equation 4, as the former equation implies the latter, but the opposite direction does not hold. A model can satisfy weak linear relational probing and not linear relational probing when, for example, both the linear probe and the query-prompted reference distribution $p(y \mid \mathbf{s} \frown \mathbf{q})$ agree on the most-likely token, but with different probabilities. E.g., *"past"* in Example 2.2 may be the likeliest token for both the linear relational probe and the query-prompted reference, but with probability $(1 - \epsilon)$ for the former and $(1 - 2\epsilon)$ for the latter.

3. Estimating Relational Linearity

In this section, we present our main experimental setting and discuss different methods to estimate relational linearity. We first remark one key aspect of relational linearity that allows for straightforward evaluation (Section 3.1), then we introduce the main setup for testing (Section 3.2), and finally we present the different probes that we test (Section 3.3).

3.1. Practical Implication of Definition 2.1

We notice that the notion of relational linear probing and its variant we discussed so far only depend on how the language model defines the next-token distribution. Upon selecting a query $\mathbf{q}$ to be tested and the set of tokens $\mathcal{Y}_P$, one only needs to access the model representations $\mathbf{f}(\mathbf{s})$ and $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$, for multiple contexts $\mathbf{s} \in \mathcal{X}$, and the embossed probabilities on $\mathcal{Y}_P$ to validate if Definition 2.1 and/or a weaker version hold. This has two immediate consequences.

[3] The definition is not restricted to well-formed sentences: both the query $\mathbf{q}$ and sequences $\mathbf{s}$ could be arbitrary strings, such as *"@!as1smf kqw#ù"*

[4] Marconato et al. (2025) introduce a more general framework for relational linear properties of which our Definition 2.1 is a special case. We provide the general definition in Definition D.1.

On the downside, the likeliest tokens for the model might be nonsensical w.r.t. the prompted pair of context and query. E.g., for the context $\mathbf{s}$ =*"I watched a movie."* and query $\mathbf{q}$ =*"What is the tense of the sentence?"*, one would expect higher likelihood on the token y_1 =*"past"* rather than y_2 =*"present"*. However, the language model might not be able to attribute high-likelihood on y_1 and predict y_2 or another token at random; this might happen for poorly trained language models. Whether this happens or not, it is irrelevant for Definition 2.1 (and variants) to hold. We observe this in one scenario when `Llama-3.1` is prompted to classify subjective *vs* objective statements (see Fig. 3, third column). For example, while `Llama-3.1` struggles to correctly recognize subjective statements, middle layers show that the model's (wrong) answers can be approximately retrieved in a linear subspace.

Moreover, model probabilities might result in trivial scenarios, where probabilities are collapsed in a small region of the simplex. E.g., if probabilities $p(y \mid \mathbf{s} \frown \mathbf{q}) = \text{const}$ for all choices in contexts $\mathbf{s} \in \mathcal{X}$ and tokens in $\mathcal{Y}$,[5] Definition 2.1 is easily satisfied by setting the weights $\mathbf{w}_{\mathbf{q}}^{y}$ to zero and properly tuning the biases $b_{\mathbf{q}}^{y}$. We propose to detect these trivial scenarios introducing a **collapse on the simplex score** (CSS) that leverages the normalized Shannon entropy $\mathrm{H}_k(\cdot)$[6] on a dataset of model probabilities $\{p(\cdot \mid \mathbf{s}_i \frown \mathbf{q})\}_{i=1}^{N}$, that is

$$\mathrm{CSS} := 1 - \mathrm{H}_k\big(\overline{p(\cdot \mid \mathbf{s}_i \frown \mathbf{q})}\big) + \overline{\mathrm{H}_k(p(\cdot \mid \mathbf{s}_i \frown \mathbf{q}))}\,, \quad (5)$$

where $\bar{\cdot}$ denotes the average over the dataset. CSS ranges in $[0, 1]$, where high values indicate a collapse towards $p(y \mid \mathbf{s} \frown \mathbf{q}) = \text{const}$, whereas low values indicate that model probabilities span multiple regions of the simplex.

On the upside, estimating relational linearity does not require accessing to predefined and well-constructed triplets $(\mathbf{s}, \mathbf{q}, y)$ used in previous works (*e.g.*, Hernandez et al., 2024; Chanin et al., 2024). This benefits a more thorough investigation upon deciding: (i) the dataset of contexts, (ii) the specific queries, and (iii) their relevant tokens. We discuss next how to make use of this effectively.

3.2. Setup

We assume to have access to a set of contexts $\mathcal{S} = \{\mathbf{s}_i \in \mathcal{X}\}_{i=1}^{N}$. For each given query $\mathbf{q} \in \mathcal{X}$, we design a prompt $\tilde{\mathbf{q}}$ to elicit the model for being more predictive of the relevant tokens $\mathcal{Y}_P$ chosen for the query. In this setting, we test a slight variation of Definition 2.1 where the order of $\mathbf{s}$ and $\tilde{\mathbf{q}}$ is inverted and use an object-inducing string $\tilde{\mathbf{t}}$:[7]

[5] A "tautology" in Marconato et al. (2025, Definiton 23).

[6] It uses a base-k logarithm, where k is the token count in $\mathcal{Y}_P$.

[7] *e.g.*, a system prompt prepended to an LLM user's input.

> *Example* 3.1. For the query $\mathbf{q}$ =*"Does the previous sentence reflect a truthful statement?"* and tokens $\mathcal{Y}_P = \{$*"Yes"*, *"No"*$\}$, the system prompt is:
>
> $\tilde{\mathbf{q}}$ =*"You are a linguistic classifier. Respond with only one word. Task: Identify the truthfulness of a sentence. Options: Yes, No."*.
>
> Accordingly, the object-inducing string is given by $\tilde{\mathbf{t}}$ = *"Truthfulness:"*.

From this, we evaluate the probability distribution of the model $p(y_j \mid \tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}}; \mathcal{Y}_P)$ for all relevant tokens $y_j \in \mathcal{Y}_P$. This gives a dataset of input contexts embeddings and probabilities pairs $\mathcal{D}_{\mathbf{q}} = \{\big(\mathbf{f}(\mathbf{s}_i), p(\cdot \mid \tilde{\mathbf{q}} \frown \mathbf{s}_i \frown \tilde{\mathbf{t}}; \mathcal{Y}_P)\big)\}_{i=1}^{N}$.

Paraphrases. We note that some relations admit different query's surface forms (i.e., paraphrases), offering alternative ways to construct the prompt ($\tilde{\mathbf{q}}$) and determine its relevant token set ($\mathcal{Y}_P$). The following illustrates an example of query paraphrasing for a binary relation:

> *Example* 3.2. As in example 3.1, we consider the query $\mathbf{q}$ =*"Does the previous sentence reflect a truthful statement?"* but now with tokens $\mathcal{Y}'_P$ = $\{$*"True"*, *"False"*$\}$, The system prompt we use in this setting is a paraphrase of $\tilde{\mathbf{q}}$:
>
> $\tilde{\mathbf{q}}'$ =*"You are a linguistic classifier. Respond with only one word.* ***Task: Classify a sentence as true or false****. Options: True, False."*.
>
> The object-inducing string is $\tilde{\mathbf{t}}'$ = *"Truthfulness:"*.

A natural question is whether different textual alternatives of a query exhibit comparable relational linearity, which we empirically evaluate in Section 4 for two distinct relations.

3.3. Methods

To address the question of whether a language model encodes any relational linearity, we consider the following estimation methods. First, we repurpose the technique of Linear Relational Embedding (`LRE`) by Hernandez et al. (2024). In short, `LRE` estimates the weights and biases for linear relational probing by resorting to a first-order Taylor expansion of the context concatenated with the query. In practice, by choosing a small set of contexts (usually way smaller than N), the joint embedding $\mathbf{f}(\tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}})$ is approximated as:

$$\mathbf{f}(\tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}}) \approx \beta \hat{\mathbf{W}}_{\mathbf{q}} \mathbf{f}(\mathbf{s}) + \hat{\mathbf{b}}_{\mathbf{q}} \quad (6)$$

where $\beta \in \mathbb{R}$ is a fine-tuned term, $\hat{\mathbf{W}}_{\mathbf{q}}$ and $\hat{\mathbf{b}}_{\mathbf{q}}$ are obtained by averaging the terms from first-order Taylor expansion on the subset of contexts. This is similar in spirit to estimating

the broader notion of relational linearity (see Definition D.1), but requires estimating the Jacobian, which can make the procedure computationally intensive and introduce additional numerical or estimation-related variability. For more details, we refer to Hernandez et al.'s work (2024).

Second, we introduce the KL-based Relational Probe (`KL-RP`), specifically tailored to test relational linear probing. `KL-RP` builds on a linear model, with k weights $\mathbf{w}_y \in \mathbb{R}^d$ and biases $b_y \in \mathbb{R}$, and is trained to align to the probability of the language model by minimizing the Kullback-Leibler (KL) divergence. `KL-RP` defines a conditional distribution over the set of tokens in $\mathcal{Y}_P$ given by

$$p^{(\ell)}_{\mathbf{W},\mathbf{b}}(y \mid \mathbf{s}) = \frac{\exp(\mathbf{w}_y^\top \mathbf{f}^{(\ell)}(\mathbf{s}) + b_y)}{\sum_{y' \in \mathcal{Y}_P} \exp(\mathbf{w}_{y'}^\top \mathbf{f}^{(\ell)}(\mathbf{s}) + b_{y'})} . \quad (7)$$

The weights and biases are chosen by minimizing the KL divergence to the model conditional distribution, that is

$$d_{\mathrm{KL}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{y' \in \mathcal{Y}_P} p(y' \mid \tilde{\mathbf{q}} \frown \mathbf{s}_i \frown \tilde{\mathbf{t}}) \log \frac{p^{(\ell)}_{\mathbf{W},\mathbf{b}}(y' \mid \mathbf{s}_i)}{p(y' \mid \tilde{\mathbf{q}} \frown \mathbf{s}_i \frown \tilde{\mathbf{t}})} . \quad (8)$$

4. Empirical Verification

We aim to answer the following research questions:

Q1 Do language models encode any linear relational property, and to what degree?

Q2 How does relational linearity vary across layers of these models?

Q3 Are paraphrases of the queries similarly encoded in model representations?

Section 4.1 describes our experimental setup, and Section 4.2 evaluates the next-token performance of language models under the proposed query prompt. We then address the three research questions in Sections 4.3 to 4.5.

4.1. Setting

Language Models. We adopt two different language models: `Llama-3.1`-8B,[8] which has 32 hidden layers and an embedding size of 4096. `Gemma-2`-2B,[9] which has 26 hidden layers and embedding size of 2304. For both models, we adopt the instruction-tuned version and apply 8-bit quantization (Dettmers et al., 2022).

Methods. We investigate the linear relational properties encoded in the neural embeddings of both language models using different methods:

1. Linear Relational Embedding (`LRE`) is implemented according to Hernandez et al. (2024). See Appendix B.2.
2. `KL-RP` is a linear layer that minimizes[10] the KL divergence in Equation 8. Softmax is used to evaluate the final probability of the tokens.
3. `Random` baseline defines the null hypothesis for relational linearity. It uses the `KL-RP` method, but is trained with randomly permuted embeddings (see Equation 9).

A further probing method, based on Support Vector Machines (SVMs), is implemented to test the relaxed variant of linear relational probing (Equation 4); details and discussion of this are left to Appendix B.4.

Metrics. To assess relational linearity, we measure:

- `F1`(GT): the linear probe F1-score to ground-truth labels of the task;
- `F1`(LLM): F1-score for the model's most likely tokens;
- KL divergence between LLM and linear probe predictions (Equation 8).

As sanity check, we measure the collapse on the simplex score (Equation 5) on model probabilities when concatenating contexts and queries. Note that F1(GT) provides a further sanity check, since a low F1(GT) combined with good performance in predicting the LLM output (either a high F1(LLM) or a low d_{KL}) might suggest some bias of the LLM in next-token prediction. This is especially important when comparing `KL-RP` to LRE, as the two methods require distinct prompt templates, causing the LLM to behave differently.

Datasets. In our experiments, we consider four sentence-level datasets, each one annotated with a distinct linguistic or semantic attribute. We investigate relational linearity by examining language, tense, subjectivity, and truthfulness. For each of these, we adopt a different dataset, so that we have the ground truth label of the given relation:

- LANG (Papariello, 2021). Each sentence is annotated with the origin language, for a total of 20 distinct idioms (each one ranging from 3.9% to 5.2% of the total). For our experiments, we consider only a subset of the total data, including 9542 samples.
- TENSE (Ayman et al., 2024). Here, sentences are written in English and are labelled according to their main verbal tense (either *past*, *present*, or *future*). Labels are in proportion 36.9% *present*, 33.8% *future*, 29.3% *past*. The dataset includes 10k samples.

[8] huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[9] huggingface.co/google/gemma-2-2b-it
[10] The model is trained using Adam with a learning rate of 10^{-4}.

- SUBJ (Alam et al., 2025). Sentences are labeled based on whether they express an opinion or a fact. This multilingual dataset covers Arabic (37%), German (23%), English (14%), Italian (14%), and Bulgarian (12%). The dataset counts 8090 samples.
- TRUTH (Marks & Tegmark, 2023). Sentences are binary labelled on factual accuracy, indicating whether they reflect true or false statements. The two classes are perfectly balanced. All sentences are written in English. The dataset contains 9123 samples.

Input sentences exceeding 250 characters (about 1.6% of the total) are excluded to avoid exceedingly long contexts and to reduce computational costs. Moreover, for all datasets, we use 80% of the entries as a training split, and the remaining 20% for evaluation. All datasets are well-balanced across classes. Appendix B reports the descriptive statistics of the four datasets (Table 3).

Queries. For each dataset, we separately construct a textual query designed to elicit model outputs that match the dataset's labels. For example, for the LANG relation, we use a system prompt with the query "What is the written language?", allowing 20 possible token-level answers corresponding to the languages in the dataset (see also Table 4). Throughout the paper, we refer to each query by its dataset name; for example, the above is denoted as the LANG query.

4.2. Performance of language models on next-token prediction

Initially, we explore the capabilities of language models to correctly predict the ground truth label on the four datasets. Results are presented in Table 1. The majority class baseline provides a reference for comparing LLM capabilities. The collapse score (Equation 5) is also reported as a preliminary measure of how indicative subsequent probing results are of relational linearity; values near 1 suggest less meaningful probing outcomes. We highlight that `Gemma-2` on the SUBJ query shows high collapse (≈ 0.91), implying the relation is almost trivially encoded, as model output probabilities are nearly collapsed into a narrow region of the simplex.

We observe that `Llama-3.1` outperforms `Gemma-2` across all datasets, with TENSE with marginal difference (0.89 vs. 0.90). However, it is worth noting that for SUBJ, both models perform worse than the majority-class baseline. We can argue that predicting input's subjectivity is challenging, and even more so in a multilingual setting without additional context, as in this case. Whereas on the LANG relation, `Llama-3.1` outperforms `Gemma-2` by a large margin (0.97 vs 0.59). This is expected, as the `Gemma-2` is mostly trained on English sentences, whereas `Llama-3.1` encounters multiple languages during training, More broadly, this performance disparity is likely attributable to the difference in model size (8B vs 2B).

Table 1. Performance of language models in predicting the ground truth label when prompted with the query appended. It also reports the CSS, and a majority class `baseline`. Best across columns are marked in **bold**.

	LANG		TENSE		SUBJ		TRUTH	
	F1	CSS	F1	CSS	F1	CSS	F1	CSS
`Baseline`	0.05	–	0.64	–	**0.64**	–	0.50	–
`Llama-3.1`	**0.97**	0.46	0.89	0.57	0.63	**0.46**	**0.86**	0.35
`Gemma-2`	0.59	**0.22**	**0.90**	**0.05**	0.41	0.91	0.70	**0.22**

Interestingly, `Gemma-2` prompted with queries of all relations exhibits more peaked probability distributions over the tokens than `Llama-3.1` (see Fig. 5, reported in Appendix B). This can also be observed in the rightmost simplex plots in Fig. 1. Since our KL-based probe (`KL-RP`) is trained directly on LLM predictions instead of ground truth, these preliminary findings on model's output distribution have to be taken into account as they set different baselines for the two language models.

4.3. Q1: Do language models encode any linear relational property?

We compare `KL-RP` to `LRE` and `Random` baseline. Due to the poor scalability of `LRE` (requiring approximately 30 mins per relation to obtain the linear approximation for a single layer) compared to `KL-RP` (approximately 40 mins per relation to validate the probe across all layers), we focus on middle layers of `Llama-3.1` and `Gemma-2`.

Table 2 shows the performance of different linear probing approaches across the four datasets. We observe that `KL-RP` shows consistently higher F1(LLM) scores and lower KL divergence compared to the random baseline, across LANG, TENSE, and TRUTH. Such results indicate that $\mathbf{f}(\mathbf{s})$ linearly encodes at least some information about the queries. In particular, this is more evident for TENSE and TRUTH in both LLMs, with F1(LLM) scores above 0.90 and KL divergence lower than 0.50. For the LANG relation, `Gemma-2` performs considerably worse than `Llama-3.1` (0.62 vs. 0.98), though it still exceeds the random baseline. This suggests that its latent representations retain some degree of linearity for this relation, yet less than `Llama-3.1` 's. We observe different outcomes on SUBJ: `KL-RP` shows less distinctive linearity for `Llama-3.1` with F1(LLM) equal to 0.88 and no degree of linearity for `Gemma-2` (0.50).

Concerning the results of `LRE`, there are two cases where good performance in predicting the LLM output is combined with a very low F1(GT) score: `Llama-3.1` on SUBJ, and `Gemma-2` on LANG. Upon further inspection, we find that in both cases they exhibit a strong bias toward specific tokens: "False" in the first case, and "Arabic" and "Swahili" in the second. This can be caused by the different prompt

Table 2. **Comparison between KL-RP, LRE, and a Random baseline**. Results are obtained with representations at middle layers ($\ell = 16$ for Llama-3.1; $\ell = 13$ for Gemma-2). Best across columns are marked in **bold**.

		LANG			TENSE			SUBJ			TRUTH		
		F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}
Llama-3.1	Random	0.04	0.05	2.04	0.28	0.36	0.57	0.41	0.47	0.49	0.50	0.49	0.49
	LRE	0.06	0.36	0.51	0.15	0.20	1.46	0.28	0.50	**0.02**	0.84	0.90	0.25
	KL-RP	**0.98**	**0.98**	**0.06**	**0.89**	**0.95**	**0.02**	**0.63**	**0.88**	0.10	**0.86**	**0.94**	**0.04**
		F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}
Gemma-2	Random	0.03	0.13	6.16	0.29	0.30	1.70	**0.39**	**0.50**	**0.16**	0.37	0.47	0.82
	LRE	0.01	0.48	**0.45**	0.47	0.48	0.88	0.27	0.48	0.38	**0.80**	**0.92**	**0.10**
	KL-RP	**0.51**	**0.62**	2.19	**0.89**	**0.95**	**0.23**	**0.39**	**0.50**	0.26	0.72	0.90	0.42

template required by LRE (see Section B.1). We observe results close to KL-RP only in the TRUTH relation for both models, with slightly worse performance for Llama-3.1 and slightly better performance for Gemma-2. In all other relations, it was not possible to linearly approximate the model's behavior using a LRE. This could suggest that the model's vector space separates the binary property of input's truthfulness more clearly, making it easier to approximate with linear methods. LRE not only reduces probing capability by *capturing only strongly-encoded relational linearity due to its approximation*, but is also *highly sensitive to hyperparameter choices* (*e.g.*, β, Appendix B.2), adding an additional degree of freedom and leading to inconsistent probing outcomes. Overall, LRE misses some linear relational properties detected by KL-RP, thus we focus on our method for the next research questions.

4.4. Q2: How does relational linearity vary across layers?

Figure 3 presents the layer-wise probing results with KL-RP, including F1 scores and KL divergence. First, we analyze the F1 score. When focusing on Llama-3.1, we observe distinct layer-wise dynamics between surface-level syntactic relations (LANG and TENSE) and semantic and more abstract ones (SUBJ and TRUTH). Surface-level relations already achieve high F1 scores in early layers, whereas abstract relations show increasing F1(LLM), peaking in the middle layers. A similar pattern is also observed in Gemma-2, except for SUBJ, as expected due to its triviality score for this relation equal to 0.91 (Section 4.2; Table 1). These layer-wise findings align with experimental observations from prior work (Appendix A; *e.g.*, Tenney et al. (2019); Cheng et al. (2025)): *surface linguistic features emerge in early layers* (*i.e.*, LANG and TENSE), while *middle layers peak on abstract ones* (*i.e.*, SUBJ and TRUTH)

About d_{KL}, Llama-3.1 representations show a sensible reduction to Random already at first layers, with noticeable improvements in middle layers (with major drops in the TENSE and TRUTH datasets). Last-layers often increase in d_{KL} but not sensibly enough to downgrade F1 scores. Overall, middle layers for all datasets show high strong and weak relational linear probing for Llama-3.1. Instead, Gemma-2 representations in first and middle layers fare higher in d_{KL} but always below random chance. Here, KL scores are lower for SUBJ, however, for a trivial relation (with Random at 0.16). Last-layers always degrade in d_{KL}, showing that Gemma-2 encode weak relational linearity at first and middle layers for linguistic datasets.

4.5. Q3: Are paraphrases of the queries similarly encoded in model representations?

We evaluate how changes in the query's surface form affect linear relational probing by varying the queries for: (1) LANG: The query is translated in four different languages: *English*, *Italian*, *Spanish*, and *Chinese*. (2) TRUTH: Three paraphrased formulations are evaluated, each corresponding to a distinct set of target tokens $\mathcal{Y}_P$: *True/False*, *Valid/Invalid*, and *Yes/No*.

We focus on these two relations to compare how paraphrases impact on linguistic *vs* abstract queries. In particular, the binary TRUTH relation allows variation of the relevant target tokens, while the LANG relation enables the study of cross-lingual queries. System prompts are in Appendix C.

Figure 2 shows the layer-wise dynamics of KL-RP in F1(LLM) tested with different paraphrases, with compre-

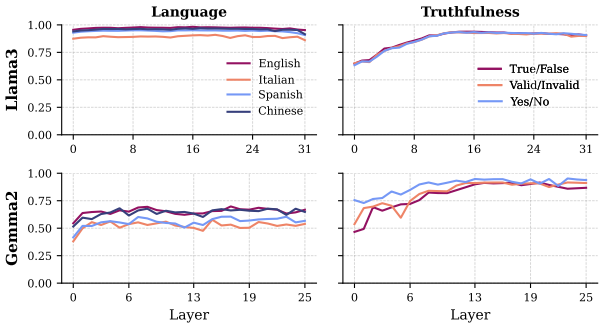


Figure 2. Relational linear probing (KL-RP) with different paraphrases. The graphs show F1(LLM) for both models.

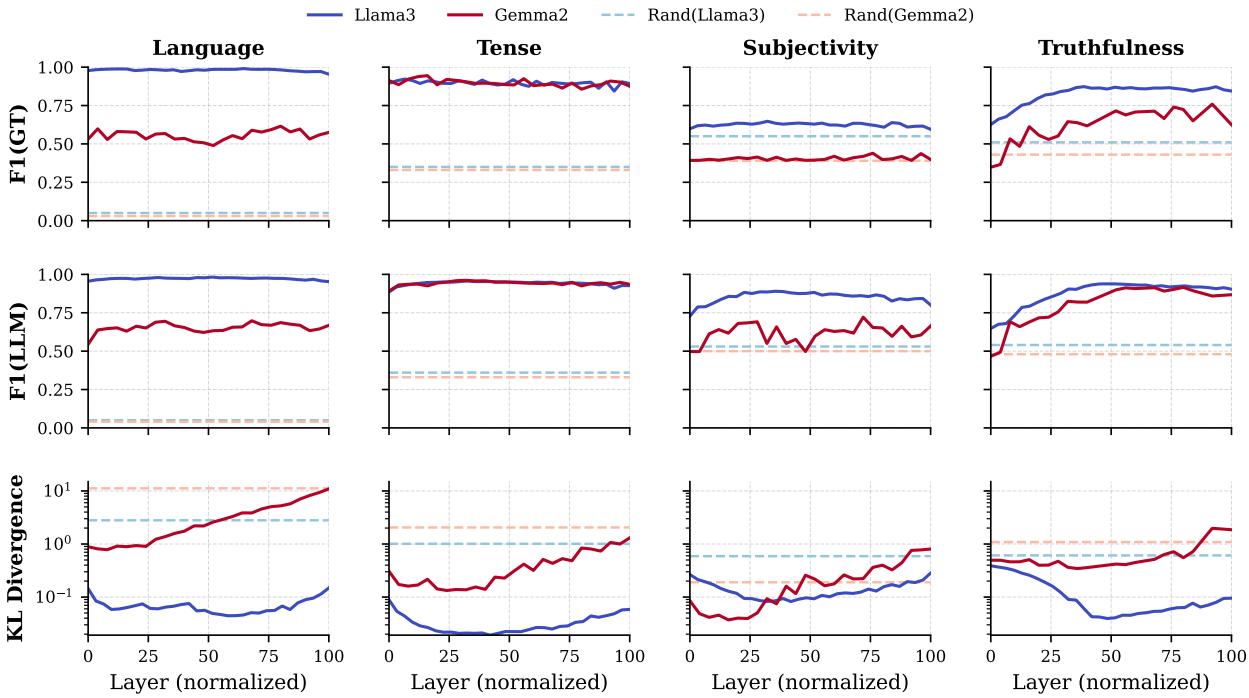


Figure 3. Relational linear probing across layers (normalized into the range 0–100), comparing `Llama-3.1` and `Gemma-2`.

hensive results available in Appendix C. Changing the language of the query slightly impact on the probing results on both language models: for `Gemma-2`, Spanish and Italian lead to lower scores (-12.5 on average) with respect to English and Chinese across all layers consistently; while for `Llama-3.1`, only Italian reduces probing performance of `KL-RP` by 6.7. For the paraphrases on TRUTH, `Llama-3.1` shows robustness to label variation, whereas `Gemma-2` exhibits substantial deviations in performance, with larger variations in earlier layers. This suggests that: (i) the *effect of surface-level query variations diminishes progressively* in `Gemma-2`, indicating that its middle layers may capture more abstract representations of this latent property, and (ii) `Llama-3.1` appears to encode relational linearity earlier, as changes in query surface form have no effect across all layers.

5. Limitations

Our work presents some limitations. While our analysis provides the first and comprehensive empirical validation of the relational linearity formulation proposed by Marconato et al. (2025), we experiment with a slight variation where each query is structured using the model's system prompt (Section 3.2). Despite this, the experimental results highlight that both language models we tested often encode strong relational linearity in low-dimensional subspaces across layers. We remark that this might not hold for the original formulation of relational linearity, *i.e.*, when considering embeddings $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$ instead of $\mathbf{f}(\tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}})$, as the former might not induce well-defined answers as the latter. We intend to evaluate this formulation in future work.

While the proposed probe does not require annotated data, unlike prior work on relational linearity (*e.g.*, Hernandez et al., 2024), the annotations of our datasets (*e.g.*, written language in LANG) enable us to benchmark probe's performance against ground truth via F1 score. However, this restricts the analysis of relational linearity to inputs that are likely to elicit well-defined answers to relational queries. This narrows the evaluation of Definition 2.1 to specific subsets of contexts, where relations are drawn from a bounded feature space. A direction of improvement is to consider more generic datasets of contexts, *e.g.*, DBpedia (Lehmann et al., 2015), enabling the assessment of relational linearity even when contexts do not yield well-defined answers to queries. We plan to investigate this in future experiments.

6. Conclusion

This work investigates relational linearity in language models, providing an empirical validation of the formulation proposed by Marconato et al. (2025). We introduce a novel probing approach based on Kullback-Leibler divergence to investigate it for two distinct language models (Q1; Section 4.3), evaluate its layer-wise dynamics (Q2; Section 4.4), and the impact of paraphrases on relational queries (Q3;

Section 4.5). The experimental results indicate that (i) relational linearity holds in several cases, (ii) layer-wise patterns emerge consistent with prior evidence on linguistic dynamics (i.e., LANG appear in early layers, while TRUTH peaks in intermediate layers), and (iii) varying the query's surface form has a slight effect on the results.

As future work, we plan to study relational linearity further, by choosing new queries, investigate emerging patterns in relational linearity caused by model distillation (*e.g.*, Llama 8B vs. 70B), and whether it can enable sensible steering.

Acknowledgments

E.M. and M.B. acknowledge 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. M.B. has been also partially supported by the European Union's Horizon Europe research and innovation program under grant agreement No. 101120237 (ELIAS). L.G. was supported by the Pioneer Centre for AI, DNRF grant number P1.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., and Søgaard, A. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th conference on computational natural language learning*, pp. 109–132, 2021.

Alam, F., Struß, J. M., Chakraborty, T., Dietze, S., Hafid, S., Korre, K., Muti, A., Nakov, P., Ruggeri, F., Schellhammer, S., Setty, V., Sundriyal, M., Todorov, K., and V., V. The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval. In Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F. M., Pinelli, F., Silvestri, F., and Tonellotto, N. (eds.), *Advances in Information Retrieval*, pp. 467–478, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-88720-8.

Ayman, U., Rahman, M. H., and Islam, M. S. Englishtense: A large scale english texts dataset categorized into three categories: Past. *Present, Future tenses*, 2024.

Chanin, D., Hunter, A., and Camburu, O.-M. Identifying linear relational concepts in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1524–1535, 2024.

Cheng, E., Doimo, D., Kervadec, C., Macocco, I., Yu, L., Laio, A., and Baroni, M. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0fD3iIBhlV.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. *arXiv e-prints*, pp. arXiv–2405, 2024.

Ferrando, J., Sarti, G., Bisazza, A., and Costa-Jussà, M. R. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Goyal, N., Daumé III, H., Drouin, A., and Sridhar, D. Causal differentiating concepts: Interpreting lm behavior via causal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hernandez, E., Li, B. Z., and Andreas, J. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. Linearity of relation decoding in transformer language models. 2024. URL https://arxiv.org/abs/2308.09124.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

Lv, A., Chen, Y., Zhang, K., Wang, Y., Liu, L., Wen, J.-R., Xie, J., and Yan, R. Interpreting key mechanisms of factual recall in transformer-based language models. *arXiv preprint arXiv:2403.19521*, 2024.

Marconato, E., Lachapelle, S., Weichwald, S., and Gresele, L. All or none: Identifiable linear properties of next-token predictors in language modeling. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824, 2023. URL https://api.semanticscholar.org/CorpusID:263831277.

Merullo, J., Eickhoff, C., and Pavlick, E. Language models implement simple word2vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5030–5047, 2024.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Paccanaro, A. and Hinton, G. E. Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):232–244, 2002.

Papariello, L. Language-identification. https://huggingface.co/datasets/papluca/language-identification, November 2021. Accessed: 2025-27-06.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, pp. 39643–39666. PMLR, 2024.

Popovič, N. and Färber, M. Tracing relational knowledge recall in large language models. *arXiv preprint arXiv:2604.19934*, 2026.

Rajendran, G., Buchholz, S., Aragam, B., Schölkopf, B., and Ravikumar, P. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37:101250–101296, 2024.

Roeder, G., Metz, L., and Kingma, D. On linear identifiability of learned representations. In *International Conference on Machine Learning (ICML)*, pp. 9030–9039. PMLR, 2021.

Sakata, M., Heinzerling, B., Ito, T., Yokoi, S., and Inui, K. Linear representations of hierarchical concepts in language models. *arXiv preprint arXiv:2604.07886*, 2026.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shu, D., Wu, X., Zhao, H., Rai, D., Yao, Z., Liu, N., and Du, M. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

Tenney, I., Das, D., and Pavlick, E. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Wang, Z., Whyte, B., and Xu, C. Locating and extracting relational concepts in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4818–4832, 2024.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., and Potts, C. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, pp. 67035–67080. PMLR, 2025.

A. Additional Related Work

Linear Properties of Language Models. The latent representations of transformer-based models, also known as residual stream, is a high-dimensional vector space that aggregates the outputs of all hidden layers (Elhage et al., 2021; Roeder et al., 2021; Park et al., 2024). Notably, these representations are hard to interpret due to their high dimensionality and the superposition phenomenon (Sharkey et al., 2025). More recently, a lot of interest lies in the possibility of identifying linearly encoded, human-interpretable features in these vectors spaces (Elhage et al., 2022; Hernandez et al., 2023; Marks & Tegmark, 2023; Engels et al., 2024; Park et al., 2024; Rajendran et al., 2024; Ferrando et al., 2024; Goyal et al., 2025). Linear features, or concepts, can be recovered using supervised learning by training probes on input's weak labels or resorting to Sparse AutoEncoders (SAEs) (Shu et al., 2025). Our work focuses on relational linearity, which can differ from linearly encoded concepts (Chanin et al., 2024; Wang et al., 2024; Sakata et al., 2026), and does not directly require any annotation step for probing or post-hoc alignment (as in SAEs).

Relational Properties. Paccanaro & Hinton (2002) introduce the notion of linear relational embeddings to model relations between entities as linear transformations in vector spaces, enabling the encoding of relations with distributed representations. The works by (Abdou et al., 2021; Hernandez et al., 2024) provides a first empirical evidence that certain relational mappings in language models can be approximated by linear transformations. Hernandez et al. (2024) frame `LRE` to inspect relational linear properties, testing it on curated datasets of annotated (subject, relation, object) triplets across 47 relations. Subsequent work by Chanin et al. (2024) invert `LRE`s to derive concept direction vectors. This and subsequent works testing relational linear properties (Wang et al., 2024; Merullo et al., 2024; Lv et al., 2024; Popovič & Färber, 2026; Sakata et al., 2026) likewise rely on annotated subject–object pairs or relation-type labels to supervise or validate the linear structure.

This work empirically investigates the formulation of relational linearity proposed by Marconato et al. (2025), which is grounded in the mechanisms by which language models compute next-token probability distributions. Our approach enables testing relational linearity by training probes to match the model's output probabilities, eliminating the need for object annotations entirely: only a dataset of contexts and a choice of query are required.

Layer-wise interpretability. Tenney et al. (2019) have detected hierarchical distributed features, where lower layers capture syntax (e.g., POS tags), while higher layers encode semantics (e.g., relations). Similarly, Cheng et al. (2025) have shown that language is represented on a low–intrinsic-dimension manifold, with a high-dimensional phase in early–middle model's hidden layers. Our layer-wise findings on relational linearity (Section 4.4) reflect this observed progression from surface to abstract linguistic properties.

B. Additional Experimental Details and Results

Table 3. Overview of the datasets after pre-processing.

Dataset	# Samples	# Classes
LANG	9,542	20
TENSE	10,000	3
SUBJ	8,090	2
TRUTH	9,123	2

Table 4. Examples from the datasets with the target class, alongside the textual query used in our work. Note that we hereby present only the *Task* within the whole $\tilde{\mathbf{q}}$, which is followed by the relation-specific tokens $\mathcal{Y}_P$ (refer to Definition 3.1 for a complete prompt example).

Relation	Sentence	Query task	Target
LANG	*Een man zingt en speelt gitaar.*	Identify the language of a sentence.	Dutch
TENSE	*I will be going to the concert.*	Identify the grammatical tense of the sentence below.	Future
SUBJ	*They're probably right.*	Identify if a sentence is a fact or an opinion.	Opinion
TRUTH	*The capital city of Laos is Vienna.*	Classify a sentence as true or false.	False

B.1. Prompt design

As introduced in Section 3.2, we adopt a structured prompt that consists not only in $(\mathbf{s} \frown \mathbf{q})$, but rather in $(\tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}})$. Since our experiments are conducted with two instruction-tuned models, we use the system prompt applying the chat template as follows:

Example B.1. Prompt for `Llama-3.1` (`KL-RP`).

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a linguistic classifier.  Respond with only one word.
Task:  Classify a sentence as true or false.
Options:  True, False.

<|eot_id|><|start_header_id|>user<|end_header_id|>

Sentence:  The capital city of Laos is Vienna.
Truthfulness:  <|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Example B.2. Prompt for `Gemma-2` (`KL-RP`).

```
<bos><start_of_turn>user

Sentence:  The capital city of Laos is Vienna.  Truthfulness:  <end_of_turn>

<start_of_turn>system

You are a linguistic classifier.  Respond with only one word.
Task:  Classify a sentence as true or false.
Options:  True, False.  <end_of_turn>

<start_of_turn>model
```

Due to the different setup required by `LRE`, the prompt template used for such method differs:

Example B.3. Prompt for `Llama-3.1` (`LRE`).

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>

You are a linguistic classifier.  Respond with only one word.
Task:  Classify a sentence as true or false.
Options:  True, False.

Sentence:  Oranges are not Naturally Orange.
Truthfulness:  False

Sentence:  Volkswagen Group operates in the industry of consumer durables.
Truthfulness:  True

Sentence:  Seventy-one is larger than ninety-seven.
Truthfulness:  False

Sentence:  The capital city of Laos is Vienna.
Truthfulness:  <|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Example B.4. Prompt for `Gemma-2` (`LRE`).

```
<bos><start_of_turn>user

You are a linguistic classifier.  Respond with only one word.
Task:  Classify a sentence as true or false.
Options:  True, False.

Sentence:  Oranges are not Naturally Orange.
Truthfulness:  False

Sentence:  Volkswagen Group operates in the industry of consumer durables.
Truthfulness:  True

Sentence:  Seventy-one is larger than ninety-seven.
Truthfulness:  False

Sentence:  The capital city of Laos is Vienna.
Truthfulness:  <start_of_turn>model
```

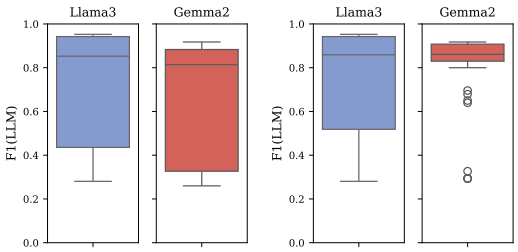
B.2. LRE setup & hyperparameters search

LREs have four hyper-parameters: ℓ, the layer after which $\mathbf{s}$ is to be extracted; n, the number of examples used to estimate the approximation; β, the rescaling factor applied to the linear map W_r; and ρ, the rank of the inverse $W^\dagger$. To ensure the scalability of the method, we set the hyper-parameters ℓ, β, ρ per LLM and relation via grid search on a subset of data (10%), while n is set to 5 as Hernandez et al. (2024) show that for most relations scores start plateauing after $n = 5$. Similarly, the values for the grid search are chosen according to their results, and are: $\ell \in \{6, 11, 16, 23, 32\}$ for `Llama-3.1`, while $\ell \in \{5, 9, 13, 20, 26\}$ for `Gemma-2`; $\beta \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$; and $\rho \in \{8, 16, 32, 64, 100\}$. Table Table 5 shows the best value found, in terms of F1(LLM), for each hyperparameter.

Table 5. Best configuration for each model and relation.

		LANG	TENSE	SUBJ	TRUTH
`Llama-3.1`	ℓ	16	16	6	16
	β	5.0	5.0	5.0	2.5
	ρ	100	100	100	64
`Gemma-2`	ℓ	13	13	13	13
	β	5.0	2.5	0.5	3.5
	ρ	64	32	32	64

Focusing on the relation leading to higher scores for `LRE`, *i.e.*, TRUTH, Fig. 4 shows the results of the different hyper-parameter combinations.



(a) All layers *(b)* Middle layer

Figure 4. **Sensibility of `LRE` when changing hyperparameters**. F1(LLM) scores on TRUTH using different configurations of `LRE` hyper-parameters, **(a)** varying ℓ, β, and ρ, or **(b)** with a fixed layer ℓ.

B.3. Randomly permuted embeddings for baseline

The random baseline is defined by rearranging the embeddings with π, a random permutation of $\{1, \ldots, N\}$, so that the randomized training data is:

$$\mathcal{D}_{\mathbf{q}}^{\text{rand}} = \left\{ \left(\mathbf{f}(\mathbf{s}_{\pi(i)}),\ p(\cdot \mid \tilde{\mathbf{q}} \frown \mathbf{s}_i \frown \tilde{\mathbf{t}}; \mathcal{Y}_P) \right) \right\}_{i=1}^{N} \tag{9}$$

B.4. Probing with SVMs

The Support Vector Machine (`SVM`) with a linear kernel is trained to match the likeliest next-token of the language model for each pair $\tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}}$, according to Equation 4. This rensembles the linear probing technique introduced by (Popovič & Färber, 2026). Preliminary experiments involving this method have been conducted using `Llama-3.1`. Table 6 presents the results, showing how the two approaches reach close performance in terms of F1(LLM), but when considering d_{KL}, `KL-RP` outperforms `SVM`, as it is specifically trained to minimize the Kullback-Leibler (KL) divergence to the LLM predictions, rather than the most probable token only.

Table 6. Comparison of KL-RP and SVM.

	LANG			TENSE			SUBJ			TRUTH		
	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}
Random	0.04	0.05	2.04	0.28	0.36	0.57	0.41	0.47	0.49	0.50	0.49	0.49
SVM	**0.99**	**0.98**	0.79	0.88	0.94	0.76	**0.65**	0.85	0.17	0.84	0.93	0.43
KL-RP	0.98	**0.98**	**0.06**	**0.89**	**0.95**	**0.02**	0.63	**0.88**	**0.10**	**0.86**	**0.94**	**0.04**

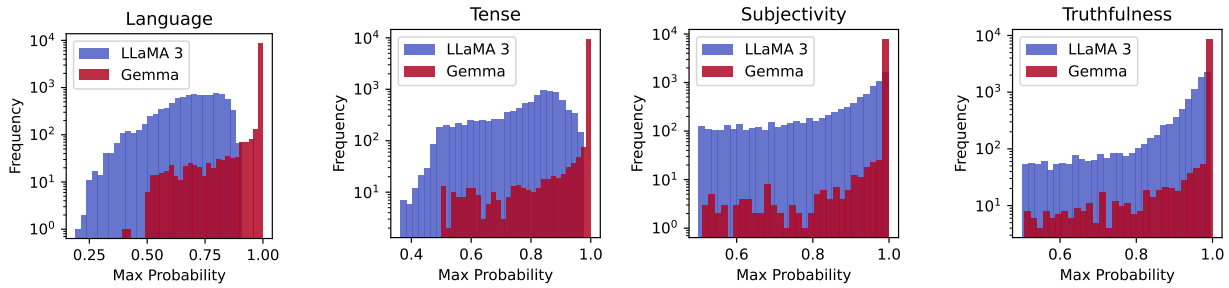


Figure 5. **Histograms of probabilities for `Llama-3.1` and `Gemma-2`.** The x-axis reports $\max p(y \mid \tilde{\mathbf{q}} \frown \mathbf{s} \frown \tilde{\mathbf{t}})$ when prompted with the query. The y-axis is the frequency in log scale. We observe that Gemma-2 tends to be more overconfident than Llama-3.1, collapsing output probabilities to high confidence regions two times more in order of magnitude than low confidence intervals.

C. Paraphrases

Table 7 shows the result of linear probing with KL-RP across different paraphrases. We varied the language of the prompt for the LANG relation, experimenting with English (i), Italian (ii), Spanish (iii), and Chinese (iv). While for TRUTH we investigated different set of labels: *True/False* (i), *Valid/Invalid* (ii), and *Yes/No* (ii). The comprehensive list of system prompts is provided in Table 8.

Table 7. Result of different paraphrases. All results refer to middle layer, *i.e.*, $\ell = 16$ for Llama-3.1 and $\ell = 13$ for Gemma-2. $-$ marks invalid measurements.

	Llama-3.1						Gemma-2					
	LANG			TRUTH			LANG			TRUTH		
	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}	F1(GT)	F1(LLM)	d_{KL}
i	0.98	0.98	0.06	0.86	0.94	0.04	0.51	0.62	2.19	0.72	0.90	0.42
ii	–	0.90	0.08	–	0.93	0.04	–	0.51	2.15	–	0.91	0.39
iii	–	0.95	0.10	–	0.93	0.04	–	0.51	2.28	–	0.92	0.27
iv	–	0.97	0.07	–	–	–	–	0.65	2.17	–	–	–

Table 8. System prompt of each query paraphrase explored.

	LANG	TRUTH
i	You are a linguistic classifier. Respond with only one word. Task: Identify the language of a sentence. Options: Arabic, Bulgarian, Greek, Hindi, Japanese, Russian, Thai, Urdu, Chinese, Dutch, Spanish, Italian, Turkish, Polish, Vietnamese, French, Portuguese, English, German, Swahili	You are a linguistic classifier. Respond with only one word. Task: Classify a sentence as true or false. Options: True, False
ii	Sei un classificatore linguistico. Rispondi con una sola parola. Compito: Identifica la lingua di una frase. Opzioni: Arabo, Bulgaro, Greco, Hindi, Giapponese, Russo, Tailandese, Urdu, Cinese, Olandese, Spagnolo, Italiano, Turco, Polacco, Vietnamita, Francese, Portoghese, Inglese, Tedesco, Swahili	You are a linguistic classifier. Respond with only one word. Task: Classify the truthfulness of a sentence. Options: Valid, Invalid
iii	Eres un clasificador lingüístico. Responde con una sola palabra. Tarea: Identifica el idioma de una oración. Opciones: Árabe, Búlgaro, Griego, Hindi, Japonés, Ruso, Tailandés, Urdu, Chino, Neerlandés, Español, Italiano, Turco, Polaco, Vietnamita, Francés, Portugués, Inglés, Alemán, Suajili	You are a linguistic classifier. Respond with only one word. Task: Classify the truthfulness of a sentence. Options: Yes, No

D. Other Definitions of Relational Linearity

We collect here additional definitions of relational linearity by Marconato et al. (2025); Hernandez et al. (2024). We first review *relational linearity* of a query $\mathbf{q}$, that is the capability of decoding the joint embedding $\mathbf{f}(\mathbf{s}\ cat\mathbf{q})$ from $\mathbf{f}(\mathbf{s})$. This may hold only on a linear subspace of the representation space, indicated as $\Gamma \subseteq \mathbb{R}^d$. The formal definition follows:

Definition D.1 (Relational Linearity of $\mathbf{q}$ in Γ by Marconato et al., 2025)**.** For a query $\mathbf{q} \in \mathcal{X}$ and a subspace $\Gamma \subset \mathbb{R}^d$, a model $(\mathbf{f}, \mathbf{g})$ *linearly represents* $\mathbf{q}$ *in* Γ if there exist a matrix $\mathbf{L}_\mathbf{q} \in \mathbb{R}^{d \times d}$ and a vector $\mathbf{d}_\mathbf{q} \in \mathbb{R}^d$ such that

$$\forall \mathbf{s} \in \mathcal{X},\ \mathbf{P}_\Gamma \mathbf{f}(\mathbf{s} \frown \mathbf{q}) = \mathbf{P}_\Gamma \left(\mathbf{L}_\mathbf{q} \mathbf{f}(\mathbf{s}) + \mathbf{d}_\mathbf{q}\right) . \tag{10}$$

In other words, all the information of the representation $\mathbf{f}(\mathbf{s} \frown \mathbf{q})$ in the linear subspace Γ is contained in (a subspace of) $\mathbf{f}(\mathbf{s})$, modulo an affine transformation. Interestingly, this property can be declined in three settings that mirror empirical notions of linearly encoded concepts (Marconato et al., 2025), including *linear relational subspaces*, *linear relational probing* (Definition 2.1), and *linear relational steering*. For a broader discussion, we refer the reader to (Marconato et al., 2025).

Next, we provide the formal definition of *weak linear relational probing*, accounting for argmax's of probability distribution of the model (measured with F1 scores). This is more aligned to the notion tested by Hernandez et al. (2024) in their experiments:

Definition D.2 (Weak Linear Relational Probing)**.** For a query $\mathbf{q} \in \mathcal{X}$ and a subset of tokens $\mathcal{Y}_P \subseteq \mathcal{Y}$ of dimension $k = |\mathcal{Y}_P|$, a model $(\mathbf{f}, \mathbf{g})$ can be *weakly linearly probed for* $\mathbf{q}$ *on the subset* $\mathcal{Y}_P$ if there exist k weights $\mathbf{w}_\mathbf{q}^y \in \mathbb{R}^d$ and biases $b_\mathbf{q}^y \in \mathbb{R}$ such that, for all $y \in \mathcal{Y}_P$ and $\forall \mathbf{s} \in \mathcal{X}$,

$$\operatorname*{argmax}_{y \in \mathcal{Y}_P} p(y \mid \mathbf{s} \frown \mathbf{q}; \mathcal{Y}_P) = \operatorname*{argmax}_{y \in \mathcal{Y}_P} \left(\mathbf{w}_\mathbf{q}^{y,\top} \mathbf{f}(\mathbf{s}) + b_\mathbf{q}^y\right) . \tag{11}$$