PREDICTING LINEAR RELATIONAL DECODABILITY WITHOUT THE RELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear relational embeddings (LREs) approximate a language model's subject-to-object mapping with an affine function. Their faithfulness varies enormously across relations, from near zero for PERSON_FATHER to the ceiling for COUNTRY_CAPITAL_CITY. We investigate the factors that contribute to this variation and demonstrate that it can be significantly predicted from a readout that is not aware of the underlying relation. To do so, we propose a novel form of the Jacobian lens, the *entity lens*. Whereas the Jacobian lens is estimated across all tokens and concepts, the entity lens EL_t is trained to predict what the model will output for an entity of type t under the neutral prompt `"{} is known for"`; e.g., we train ELs for countries and persons. After estimating an entity lens, we measure how highly it ranks the model's answer for a relation without using any relation-specific information in estimation and prediction. Across 28 relations of the LRE dataset on Olmo-3-7B this relation-agnostic score correlates with LRE faithfulness at $\rho = 0.7355$ ($p < .01$). This suggests that linear decodability is largely a property of what a subject's representation already carries, rather than of the relation being decoded. Moving the unit of analysis from the relation to the individual fact, we conclude that new relation datasets and analysis methods are needed to address several problems in existing datasets, most prominently semantically weak first tokens (e.g., most predicted universities' first token is "University") and reference competition (e.g., the LLM confuses obscure actor names with competing non-actor entities that have the same last name). Our full pipeline is available here.

1 INTRODUCTION

Hernandez et al. (2024) showed that a language model's mapping from a subject to the object of a relation is often well approximated by an affine function of the subject's hidden state, a *linear relational embedding* (LRE). They also showed that this works far better for some relations than others. On OLMo-3-7B (Olmo et al., 2025), we reproduce the same spread: an LRE fitted on COUNTRY_CAPITAL_CITY agrees with the model's own prediction 87.5% of the time while one fitted on PERSON_FATHER agrees 7.5% of the time.

Why? The natural reading is that some relations are "more linear" than others: linearity is a property of the relation. We report evidence that it is better understood as a property of the *subject's representation*. Concretely, we show that one can predict how faithful an LRE will be *before fitting it*, and without using the relation at all.

Building on recent work by Gurnee et al. (2026), we introduce the *entity lens* EL_t, which is trained on subjects of type t placed in the neutral prompt `"{} is known for"` and evaluates the Jacobian at entities of type t. This readout provides factual knowledge about entities of a given type. The question it answers, *what factual knowledge is this model ready to produce about this specific entity?*, depends on the entity type, but not on the relation. We estimate this lens at one layer fixed in advance, and ask how highly it ranks the model's own answer. The lens sees no relation template, no objects and no (subject, object) pairs, yet its score correlates with LRE faithfulness across 28 relations at $\rho = 0.7355$.

We make four contributions. (i) We introduce the entity lens, a supervision-free readout of what a model is disposed to say about an individual entity of an entity type. (ii) We turn it into a relation-agnostic diagnostic of linear decodability and show that it predicts LRE faithfulness. (iii) We show

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the prediction holds up under leave-one-type-out and under controls for object cardinality and answer-token frequency. (iv) Moving the unit of analysis from the relation to the individual fact, we provide an analysis of several problems in existing datasets that hinder further progress on understanding LREs in language models, most prominently semantically weak first tokens and reference competition.

2 BACKGROUND

An LRE for relation r approximates the model's computation as

$$o \approx \beta W s + b_\beta, \tag{1}$$

where s is the hidden state at the subject's final token and o the output state. Write f for the model's own map between the two: f sends the subject's hidden state at an intermediate layer ℓ to the final-layer state at the last position. We average local first-order approximations of f over a **reference set** $\{s_i\}$ of reference subjects and scale their subject-dependent terms by β:

$$W = \mathrm{mean}_i\, J_i, \qquad b_\beta = \mathrm{mean}_i[\, f(s_i) - \beta J_i s_i \,], \tag{2}$$

where J_i is the Jacobian of f at s_i. At $\beta = 1$, this is the averaged first-order Taylor approximation (Hernandez et al., 2024).

This prediction decomposes into a subject-dependent term and a constant: expanding b_β gives

$$\beta W s + b_\beta \;=\; \mathrm{mean}_i[\, f(s_i) + \beta J_i (s - s_i) \,], \tag{3}$$

in which only the second term depends on which subject s is being decoded. Left unscaled, the shared anchor $\mathrm{mean}_i\, f(s_i)$ dominates and the operator returns much the same object for every subject. Setting $\beta > 1$ amplifies the subject-specific contribution relative to that anchor, and so does change which object is ranked first. Hernandez et al. (2024) fit β per relation alongside the layer; we follow their protocol and select it from $\{0.5, 1, 1.5, 2, 3, 5\}$.

Faithfulness is the rate at which decoding $\beta W s + b_\beta$ over the full vocabulary reproduces the *model's own* top-1 prediction (not the gold object). It is therefore a measure of how well an affine map captures what the model does. Faithfulness defined this way is the quantity we predict.

3 RELATED WORK

Linear relational embeddings. Paccanaro & Hinton (2000) proposed Linear Relational Embeddings (LREs) to model relations as linear maps between concept representations. Hernandez et al. (2024) applied LREs to transformers (Vaswani et al., 2017). Subsequent work has asked why LREs are faithful for some relations and not others. Merullo et al. (2025) trace part of this variation to pretraining data: LREs are more faithful when subjects and objects frequently co-occur in the corpus. Co-occurrence frequency also predicts whether facts are acquired during pretraining, though it does not fully explain learning (Gottesman et al., 2026). We report such corpus-frequency measures as baselines in Appendix A.1. A further question is whether such structure reflects what a model has learned or only the particular parametrization that realizes it, since many parametrizations induce the same next-token distribution. Existing results point to the former, since under suitable conditions, relational linearity holds either in all distribution-equivalent models or in none (Marconato et al., 2025). Using a KL-divergence probe, Valer et al. (2026) have shown that relational linearity varies across models, layers and phrasings of the relation. These works evaluate and explain linearity at the level of the relation. We ask instead how much of it can be predicted from how the model represents the subject, using a readout estimated without the relation or any of its objects.

Subject representations in factual recall. Mechanistic studies converge on the view that much factual knowledge is attached to the subject before the relation is processed. Feed-forward layers operate as key-value memories whose keys respond to patterns in the input and whose values induce distributions over the output vocabulary (Geva et al., 2021). Meng et al. (2022) show that it is in the middle layers at the last subject token that factual associations are localized. Building on these findings, factual recall has been decomposed into stages: MLP layers enrich the subject representation with many of its attributes, a lookup that depends largely on the subject's name rather than on the preceding context; the relation is propagated to the final position; and attention heads extract the

requested attribute (Geva et al., 2023; Nanda et al., 2023). Our approach can be read as a probe of this enrichment, since it reports what the model is prepared to say about a subject before any relation is given. This allows us to connect two lines of work: We hypothesize that the enrichment the lens diagnoses contributes to the relational linearity found by Hernandez et al. (2024). The lens predicts LRE faithfulness across relations in all seven models we test, and the association persists when controlling for how many subjects share each object and for answer-token frequency (§6).

How facts are stored. Theoretical work has analyzed two mechanisms by which transformers store facts. Nichani et al. (2025) model attention and MLP weights as associative memories and prove that a one-layer transformer can store a number of facts that scales linearly with its parameter count, with storage traded off between self-attention and MLP parameters. Ravfogel et al. (2026) study a geometric form of storage: in a controlled one-layer setting, subject embeddings encode linear superpositions of their associated attributes, and a small MLP selects among them conditioned on the relation. In pretrained models the two need not compete: lookup-like computation can enrich the subject representation over the early and middle layers, and the geometric picture then describes the enriched state from which a relation selects. Our hypothesis concerns this enriched state: a relation should be linearly decodable to the extent that its attribute is prominently encoded in the subject representation after enrichment.

Lenses. Lenses read out intermediate hidden states in terms of the output vocabulary and differ mainly in the map they apply: the model's own unembedding (nostalgebraist, 2020) or an affine translator learned per layer (Belrose et al., 2023). Patchscopes place these readouts in a common framework, in which a hidden state is decoded by patching it into a target prompt, and have been used to inspect how models resolve entity names (Ghandeharioun et al., 2024). More recently, Gurnee et al. (2026) apply the model's input-output Jacobian, averaged over a large corpus, to raw activations, with the aim of identifying representations the model is poised to verbalize. Our entity lens instead evaluates the Jacobian at reference entities of the subject's type under a single neutral factual prompt, so that the readout concerns what the model is prepared to say about entities of that type rather than about any concept. Its contrastive variant further applies the Jacobian to differences between entity representations, subtracting what the references share.

4 THE ENTITY LENS

Gurnee et al. (2026)'s Jacobian lens is broadly similar to LRE. But instead of a relation-specific prompt and a focused target position (the object), it computes the average $W = \mathrm{mean}_i\, J_i$ over many diverse prompts and many output positions. The Jacobian lens also does not use the "anchor", the zero-order term of the Taylor expansion; so it is simply $W = \mathrm{mean}_i\, J_i$.

We derive our *contrastive entity lens* from Eq. 3, restated as:

$$f(s) \approx \beta W s + b_\beta \;=\; \mathrm{mean}_i[\, f(s_i)] + \beta\, \mathrm{mean}_i[\, J_i(s - s_i)\,] \tag{4}$$

Following Gurnee et al. (2026), we also drop the anchor term $\mathrm{mean}_i[\, f(s_i)]$; see discussion in §7.3. This also means that we do not need β since the scaling does not change the output distribution over tokens. The final form of the contrastive entity lens is then:

$$\mathrm{CEL}(e) = \mathrm{mean}_i[\, J_i(s_e - s_i)\,] \tag{5}$$

CEL takes into account the difference between s_e and the reference point s_i, as it should if we want to approximate $f(s_e)$ using the first derivative. Our original motivation was that we are not interested in what the model associates with an entity like "Russia" in general (e.g., "Europe"). Instead we are interested in what the model associates with "Russia", but not with other reference entities like "Spain" and "Italy". Examples of such contrastive associations for "Russia" are "vodka" and "Kremlin". Our hypothesis was that such top-ranked contrastive associations give rise to linearity.

In initial experiments, we found that the contrastive entity lens is quite sensitive to the reference set. For example, if s_e is a French person and a Spanish person is the reference entity s_i, then the vector difference in Eq. 5 will enhance "French". But if s_i is another French person, then "French" will be demoted. This is bad for answering questions about s_e's nationality and native language. We therefore also consider the following noncontrastive entity lens, which drops the reference set differences:

$$\mathrm{NEL}(e) = \mathrm{mean}_i[\, J_i s_e\,] \tag{6}$$

In what follows, EL refers to both NEL and CEL.

To instantiate EL for factual relations, we need (i) a prompt that evokes factual associations and (ii) a reference set of entities for computing the Jacobians J_i (both NEL and CEL) and for serving as the s_i in the vector differences $s_e - s_i$ (for CEL).

As the prompt we adopt `"{} is known for"`. This is a neutral prompt, i.e., it is not customized for any specific factual relation. We use it for all our experiments unless otherwise stated.

As the reference set, we use entities of the same type as the subject, e.g., other countries if the subject is a country, other landmarks if the subject is a landmark etc. NEL applied to an input s outputs anything the model is likely to say in a context in which an entity of the same type as s occurs, including features it shares with other same-type entities. CEL outputs only what differentiates s from other same-type entities.

Computation. We do not compute J_i explicitly, but obtain each Jacobian–vector product $J_i u$ by the double-backward construction (Pearlmutter, 1994), at the cost of a constant number of forward and backward passes (Appendix B).

5 Experimental setup

Models. Our primary analysis is conducted on OLMo-3-7B because it combines performance with full replication resources for analysis. Results for six additional models appear in Appendix A. Inference details are in Appendix B.

Relations. The LRE dataset (Hernandez et al., 2024) contains 47 relations. Our prompt presupposes an entity subject, so we fixed a subject-type filter in advance using the dataset's declared `domain_name` field: the 28 relations with named-entity subjects (*person*, *country*, *name*, *company*, *landmark*, *superhero*, *city*, *pokemon*, *product*, *star*). Excluded are relations over words as words (*verb*, *adjective*, *word*) and over abstracts (*occupation*, *task*, *characteristic*, *degree*), for which `"{} is known for"` is a category error. No other relation was excluded.

Measurement. For LRE faithfulness, we follow Hernandez et al. (2024). A relation is presented to the model *in context*: eight (subject, object) examples of the relation are concatenated, one per line, and the query subject is appended without its object, so that the model's next token is its answer. This is what defines the m_e, the model's **top-1 answer**. The subjects of a relation are shuffled once and partitioned into three disjoint sets: eight in-context examples and eight reference subjects from which W and b_β are estimated.

Faithfulness is then recomputed for all 28 relations under this protocol. Layer and β are chosen on a 6×6 grid consisting of six candidate layers for the subject state $(5, 11, 14, 17, 20, 23)$ crossed with six values of β $(0.5, 1, 1.5, 2, 3, 5)$. Selection is based on the mean over the test subjects. Hernandez et al. (2024) likewise fit both per relation; the six layers span the network at roughly even spacing and the β range brackets the values they report.

To measure the accuracy of the predictions of the entity lens, we decode $\mathrm{EL}(e)$ over the full vocabulary and record the rank of m_e, the model's top-1 answer. Writing $\mathrm{rk}_{\mathrm{EL}(e)}(m_e)$ for that rank, the diagnostic for a relation is

$$A_r \;=\; -\,\mathrm{median}_e\ \log_{10} \mathrm{rk}_{\mathrm{EL}(e)}(m_e), \tag{7}$$

so higher A_r means the lens agrees with the model more.

The lens and the LRE score every row of a relation (19–994 rows), including the in-context examples and the reference set. For the results in §6, the lens and the LRE sample their splits independently, so the particular subjects entering each set differ between the two measurements. Independent splits introduce sampling variation into this comparison. Because the LRE hyperparameters are selected on the scored subjects, its reported faithfulness is an in-sample estimate; these results do not establish performance on a separate held-out evaluation set.

Note that the in-context prompt serves only to define the target m_e; the lens itself is always computed under the neutral prompt of §4.

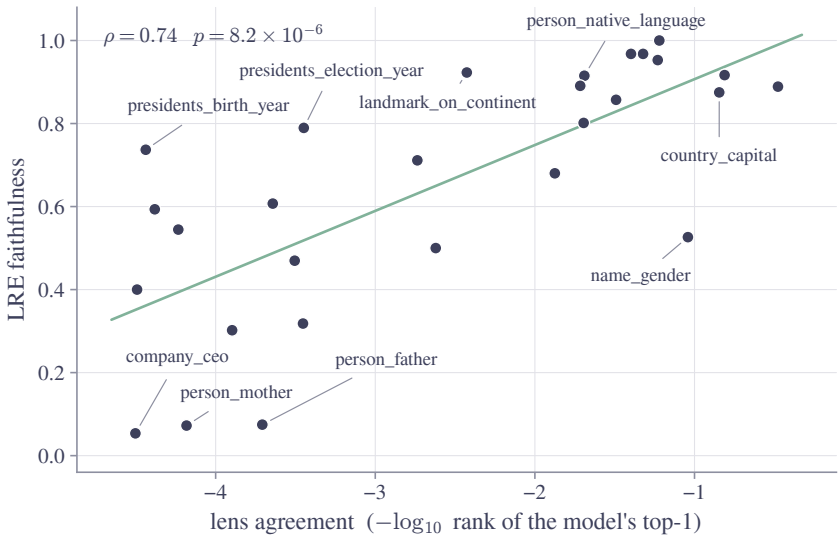

Figure 1: LRE faithfulness against the relation-agnostic CEL lens score, 28 relations. The lens sees no relation template, no objects, and no (subject, object) pairs, and uses one fixed layer throughout.

6 RESULTS

Figure 1 plots LRE faithfulness against the median entity-lens score across subjects for each of 28 relations, with a least-squares fit. The scores correlate at Spearman $\rho = 0.7355$ ($p = 8.24 \times 10^{-6}$). We next test three other predictors of faithfulness.

Object cardinality. Relations differ greatly in how many distinct objects they have; k, the mean number of subjects per object, ranges from 1.0 to 189. High k should make faithfulness easier, and k is indeed associated with faithfulness on its own, though only weakly ($\rho = 0.355$, $p = 0.064$). But k is weakly correlated with the lens score ($\rho = 0.086$). Removing k *raises* the correlation to $\rho = 0.757$ ($p = 4.90 \times 10^{-6}$): k and the lens are complementary predictors of faithfulness, so removing the variance attributable to k sharpens rather than weakens the lens signal. Since object cardinality is a complementary predictor, we will analyze it in more detail in §6.1 below.

Answer frequency. A relation whose answers are frequent tokens might be easy for both measures. We estimate the model's prior over each answer token by averaging its log-probability over ten short generic prefixes. This predicts neither faithfulness ($\rho = -0.111$) nor the lens score ($\rho = 0.098$), and removing it leaves the correlation unchanged at 0.755. The correlation is 0.784 with both cardinality and answer frequency removed.

Entity type. Eleven of the 28 relations have *person* subjects, raising the worry that one domain (i.e., one subject type) drives the result. Holding out each subject domain in turn leaves $\rho \in [0.696, 0.799]$, median 0.728.

6.1 OBJECT CARDINALITY

It turns out that object cardinality partially explains the tails of the distribution in Figure 1. We now investigate the two tails, i.e., those relations with the largest positive and negative deviations in Figure 1.

We rank relations by their residuals from the least-squares fit in Figure 1, excluding the two rescored year relations (Appendix D). For NAME_GENDER, we use the model's literal answer token, matching faithfulness, rather than all gender-matched alternatives. A residual of $+0.30$ means faithfulness is thirty percentage points above the lens prediction.

The largest positive residuals, faithful despite a low lens score, with their cardinality are NAME_GENDER ($+0.289$, $k = 9.5$), PERSON_UNIVERSITY ($+0.258$) and LAND-

Table 1: Object prominence, measured by the model's own free generation and therefore independently of the lens, against how often the lens surfaces the gold object in its top five. Spearman $\rho = 0.570$ ($p = 0.033$, $n = 14$).

relation	prom.	gold∈lens	n
city_in_country	0.96	0.71	24
person_plays_pro_sport	0.69	0.36	318
landmark_in_country	0.56	0.30	836
star_constellation	0.48	0.06	361
person_university	0.32	0.00	91
country_capital_city	0.29	0.42	24
person_occupation	0.25	0.12	821
company_hq	0.23	0.02	674
product_by_company	0.22	0.19	522
country_largest_city	0.21	0.42	24
country_currency	0.20	0.03	30
person_father	0.11	0.01	942
person_mother	0.10	0.01	986
superhero_archnemesis	0.06	0.01	94

MARK_ON_CONTINENT ($+0.239$, $k = 189$). In each, the LRE chooses among a handful of possible answers (for PERSON_UNIVERSITY, the model answers "University" for 54 of 91 subjects; §8), and can be faithful without the answer being salient for any particular subject: no name is chiefly "known for" its gender, but there are only two answers to pick between.

The largest negative residuals, unfaithful despite the lens, are PERSON_FATHER (-0.381), PERSON_MOTHER (-0.299) and COMPANY_CEO (-0.260), all with $k \approx 1$ and answer spaces in the hundreds. For these relations, faithfulness is near its floor due to the difficulty of the task and the details of the representation apparently do not change that.

The relationship between our NEL-based score and LRE faithfulness thus holds in the middle of the cardinality range and degrades at both ends, in opposite directions; across relations, residuals increase with k (Spearman $\rho = 0.60$, $p = 0.001$). Our analysis in this subsection helps us interpret the result from the beginning of this section: k is a true source of deviation, not merely a confound.

7 Analysis of the entity lens

7.1 Salience

Our approach is based on an assumption that is also shared by the Jacobian lens: Readouts of this kind are normally taken to surface what a model is internally disposed to say (Gurnee et al., 2026). We read EL(s) the same way: as what the subject's representation makes available, before any relation has been named. Our interpretation of Figure 1 rests on that reading. If it is right, a relation should be linearly decodable to the extent that its answer is already part of what the subject evokes; that is, the object should be *salient* for the subject. This section performs a sanity check (a check that is external to the lens) on that assumption by measuring salience by the model's own free generation.

For each test subject we greedily generate 90 tokens from each of `"{} is best known for"` and `"Here are some facts about {}. First,"`, and record whether any distinctive word of the gold object appears in either continuation.

We refer to this rate as the object's *prominence*. Table 1 gives both prominence and gold∈lens (how often the lens surfaces the gold object in its top five) for 14 relations. Prominence ranges from 0.96 (CITY_IN_COUNTRY) and 0.29 (COUNTRY_CAPITAL_CITY) down to 0.11 (PERSON_FATHER), 0.10 (PERSON_MOTHER) and 0.06 (SUPERHERO_ARCHNEMESIS). It predicts how often the lens surfaces the gold object in its top five: Spearman $\rho = 0.570$ ($p = 0.033$, $n = 14$). Thus, what the lens surfaces is predicted by a measurement of salience that never consults the lens.

Table 2: The reference set's effect on the two lenses, on PERSON_NATIVE_LANGUAGE. All four reference sets are scored on the same 61 French subjects. Left: median $\log_{10}$ rank of the model's own answer m_e. Right: median rank of the token "French" itself. CEL degrades as the reference entities come to share the subject's answer; NEL, which takes no difference, does not.

	median $\log_{10}$ rank of m_e		median rank of "French"	
reference set	CEL	NEL	CEL	NEL
8 non-French	1.62	1.81	47	64
8 from other relations	1.88	1.72	102	70
4 French, 4 non-French	1.93	1.79	115	82
8 French	2.72	1.71	665	55

7.2 REFERENCE SETS

§4 motivated NEL by the observation that CEL is sensitive to the reference set: if the reference entities s_i share the answer with s_e, the difference $s_e - s_i$ removes it. We now measure that effect.

We test PERSON_NATIVE_LANGUAGE on a fixed subset of 100 subjects (Table 2): 69 are French speakers, and the model answers "French" for 65 of them, allowing us to observe cancellation of a single-token answer. We preselect eight French and eight non-French reference entities and exclude them from testing. All four reference sets are evaluated on the same 61 French and 23 non-French subjects.

Table 2 shows the result. Under CEL, moving from non-French to French reference entities costs 1.10 in median $\log_{10}$ rank, a factor of 13. In contrast, NEL is robust to reference set composition: results are in a small range, $[1.71, 1.81]$. Tracking the answer token itself rather than each subject's own target makes the mechanism plain: the rank of "French" rises monotonically with the number of French reference entities, $47 \to 115 \to 665$ for zero, four and eight, and the effect is monotone for 52 of the 61 subjects individually. The control is the last column. *The same eight French reference entities* leave "French" at rank 55 under NEL, its best value for any reference set. The collapse is therefore a property of the subtraction, not of the reference entities.

Reversal of the effect. On the 23 non-French subjects the ordering under CEL is the other way round: the eight French reference entities are the *best* set (median $\log_{10}$ rank 0.78) and the eight non-French ones the worst (1.97). The same reference set helps or hurts depending on what is being asked, so reference quality is not a property of the set alone.

What is subtracted. CEL is $\mathrm{mean}_i[J_i s_e] - \mathrm{mean}_i[J_i s_i]$, and it is tempting to read the second term as the answer. It is not: decoded over the vocabulary it gives "Cin", "La", "Une", "fils". These are French vocabulary items, not the token "French". This is expected, since $\mathrm{mean}_i[J_i s_i]$ is what the model has loaded about those people under `"{} is known for"`, which is French cultural material rather than an answer to a question about language. The answer is suppressed because the two terms share the French direction, not because the subtracted term is the answer.

Summary. A reference set of the same type as the subject (§4) does not guarantee that the set is neutral with respect to the relation being read out, and no choice of reference entities can be neutral for every relation at once. This is why NEL is preferred to CEL.

7.3 AN INTERVENTIONAL ANALYSIS

The LRE of Eq. 3 is the sum of an *anchor* b_β and a *subject term* $\beta W s$. The subject term has the form of NEL (Eq. 6). We recombine the terms to determine what each captures. If the anchor captures the relation, pairing it with a *foreign* subject term from another relation should still produce this relation's object; and if NEL captures the subject, it should be able to stand in for the subject term.

We use two relation pairs that share their subjects: COUNTRY_CAPITAL_CITY and COUNTRY_LARGEST_CITY over 22 countries whose capital is not their largest city and for which the model answers both relations correctly, and PERSON_PLAYS_POSITION_IN_SPORT and PERSON_PLAYS_PRO_SPORT over 24 athletes. Each subject is read with the others as references, at $\beta = 3$. Appendix E describes the experiments in full.

		subject term			controls	
relation	n	own	foreign	NEL	anchor alone	wrong donor
COUNTRY_CAPITAL_CITY	22	12	12	7	0	0
COUNTRY_LARGEST_CITY	22	12	8	4	0	0
PERSON_PLAYS_POSITION	24	19	21	18	7	6
PERSON_PLAYS_PRO_SPORT	24	22	21	21	6	3

Table 3: Recombining an LRE's anchor with different subject terms, $\beta = 3$. Counts are how often the decoded top-1 is the relation's gold object. *Own* is the unmodified operator; *foreign* uses the paired relation's subject term; *NEL* is the noncontrastive entity lens, which never sees a relation; *anchor alone* drops the subject term; *wrong donor* uses another subject type's NEL. The model answers all 22 countries correctly, but the operator itself reproduces only 12 (Appendix E).

The anchor captures the relation. A foreign subject term leaves the answer largely intact (Table 3): with the largest-city subject term, the capital anchor still names 12 capitals, as many as with its own; each person anchor keeps 21 of 24 subjects with the other relation's subject term. The anchor alone names no country's capital or largest city, so these hits are composed rather than read off the anchor. Nor do they depend on the particular references: over 500 anchors estimated from references disjoint from the test subjects, the foreign subject term stays on average within one to two subjects of the operator's own on all four relations.

The lens stands in for the subject term, but only moderately. NEL, computed under the neutral prompt without relation-specific information, recovers the operator for sport (21 against 22) and largely for position (18 against 19), but for only 7 of the 12 capitals and 4 of the 12 largest cities that the operator itself gets right. What it supplies is subject-specific: another subject's NEL falls to or below the anchor alone. It is also less stable than the foreign subject term. Over the same 500 anchors it ranges from 7 to 20 of 24 for position, and for the country pair neither the anchor's references nor NEL's own references dominate its variance.

8 Fact-level analysis

Linearity emerges as a phenomenon at the relation level, not the fact level. It would not make sense to investigate linearity for the facts of a relation that is not represented linearly. But once there is evidence for at least partial linearity of a relation, we determine for individual facts how well the entity lens predicts the degree to which they are integrated into the linear structure.

To analyze relations at the fact level, we compute ρ between lens $\log_{10}$ rank and LRE $\log_{10}$ rank for each relation. The values of ρ are in the range [-0.401,0.664] with a median of 0.235. We found several shortcomings in the experimental setup currently used to investigate relations in LMs that result in these low values. We now analyze these shortcomings and give guidance for future relation datasets that may better support research on how LMs represent facts.

8.1 Reference competition

The three relations PERSON_UNIVERSITY (ρ=-0.401), PERSON_OCCUPATION (ρ=-0.388) and PERSON_PLAYS_INSTRUMENT (ρ=-0.035) have the lowest correlations. They exemplify the problem of *reference competition*. An example is Rolf Thommessen who is listed as an actor in LRE. The highest ranked token given by the lens is "whistlebl". This indicates that the most prominent entity associated with the last name Thommessen is Advokatfirmaet Thommessen, a law firm known for digital whistleblowing. Apparently, the first name "Rolf" does not disambiguate which subject is intended. In general, the last few tokens seem to dominate which entity is represented and used by the LM. For the actor Barry Sobel, the lens ranks "Jewish", "Israel" and "Holocaust" among its top five, associations for the Brazilian Rabbi Henry Sobel. The relation STAR_CONSTELLATION (which in Table 1 was an outlier: high prominence, low gold∈lens) also suffers from reference competition due to the complex names of many stars. For example, the most prominent associations of the star "rho01 Cnc" are "machining", "CNC", "manufacturing" (Cnc = Computerized Numerical Control).

8.2 SEMANTICALLY WEAK FIRST TOKENS

PERSON_UNIVERSITY exemplifies the problem of *semantically weak first tokens*. Many universities have names that start with "University". So the LM learns to predict this common prefix as the target token for many persons. The standard setup is to count such a prediction as correct. But this is not an effective evaluation: The correct generation of "University" is not a reliable indicator of the factual competence of the LM.

The collapse rate is the fraction of a relation's distinct objects that share a first token with another. **6 of 47 relations exceed 25%**, led by PERSON_MOTHER (61%), PERSON_FATHER (58%) and PERSON_UNIVERSITY (45%: 29 of 69 objects begin with UNIVERSITY).

Both LRE and the lens work on the token level, the residual stream vector at the last layer, which is focused on producing the correct next token and may not represent information about the tokens to follow. This single-token regime is problematic for the standard setup because we are effectively investigating a relation in which all objects have been shortened to their first token. This is clearly different from the intended relations whose objects are full entity names.

8.3 TOWARDS BETTER FACT DATASETS

Our analysis motivates two requirements for future evaluations. (i) Answer scoring should distinguish complete objects rather than credit a semantically weak first token. Evaluation methods that score complete answers are therefore needed for relations with unavoidable prefix overlap. (ii) A subject name should refer to the intended entity for the model under study. Dataset construction alone may not ensure this, because different models can resolve the same name differently. Evaluations should examine the entity that a model associates with each subject, document ambiguity, and distinguish errors of subject identification from errors in relational decoding. These are methodological directions motivated by the observed failures, rather than solutions evaluated in this study.

9 LIMITATIONS

The study only covers 28 relations, some of which have small sample sizes. Figure 1 is correlational: we show that a relation-agnostic readout predicts LRE faithfulness, not that the representation it reads is what the LRE exploits.

Our NEL-based score is only a good predictor of LRE faithfulness in the middle of the object cardinality range (§6.1). It is weakest where the answer space is very small, since an LRE can then be faithful by elimination rather than by reading anything from the subject, and where the answer space is very large and each answer near-unique, since faithfulness is then close to its floor regardless. We show that these are the two tails of a single variable, but we do not have a principled description of where the boundaries lie.

We also probe salience through a limited set of neutral prompts. That phrasing asks what an entity is *noted* for, which need not coincide with everything its representation carries; attributes such as gender may be well encoded without being something a name is "known for". A different neutral prompt, or an average over several, might track salience more broadly.

The smallest relations have only 19–24 examples, so their faithfulness estimates remain coarse.

Finally, the diagnostic is defined only for relations whose subjects are entities (and not words).

10 CONCLUSION

Whether a relation is linearly decodable from an LM is substantially predictable without the relation. A Jacobian lens estimated from subjects under a neutral prompt, at a fixed layer, ranks the model's own answer highly for exactly those relations whose LREs turn out to be faithful.

We favor salience as the mechanism: LRE faithfulness depends on how readily the subject's representation evokes the answer. On this view, variation across relations largely reflects the entities they concern.

AI USE STATEMENT

We used generative AI tools for brainstorming, assisting with software code, and initial drafting and rewriting of parts of the manuscript. All AI-assisted code and text were reviewed and verified by the authors, and generated code was tested for correctness. The authors take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We release the complete pipeline, and one command reproduces every number, table and figure in the paper from scratch: it downloads the LRE dataset and all models, runs every measurement, and builds the tables and figures. To confirm this, we ran the full pipeline in a fresh Docker container and compared every reported value against the paper. The pipeline pins the datasets, checkpoints, prompt templates, random seeds and hyperparameters in configuration files. Reproducing everything takes roughly 35 A100 (80 GB) GPU-hours for the seven models, plus about 1,200 CPU core-hours and 3.6 TB of disk for the OLMo training-corpus counts. All models and datasets are publicly available.

REFERENCES

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5484–5495, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Daniela Gottesman, Alon Gilaie-Dotan, Ido Cohen, Yoav Gur-Arieh, Marius Mosbach, Ori Yoran, and Mor Geva. Lment: A suite for analyzing knowledge in language models from pretraining data to representations. *Transactions of the Association for Computational Linguistics*, 14:1654–1691, 07 2026. ISSN 2307-387X. doi: 10.1162/TACL.a.746. URL https://doi.org/10.1162/TACL.a.746.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

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 *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Emanuele Marconato, Sébastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or None: Identifiable Linear Properties of Next-Token Predictors in Language Modeling. In *Proceedings of*

The 28th International Conference on Artificial Intelligence and Statistics, volume 258, pp. 4123–4131, 2025. URL https://mlanthology.org/aistats/2025/marconato2025aistats-all/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2504.12459.

Neel Nanda, Senthooran Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level. Alignment forum, December 2023. URL https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, volume 2025, pp. 4166–4207, 2025.

Rob nostalgebraist. Interpreting gpt: The logit lens. *Blog Post*, 2020.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Alberto Paccanaro and Geoffrey E. Hinton. Learning distributed representations by mapping concepts and relations into a linear space. In Pat Langley (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 711–718. Morgan Kaufmann, 2000.

Barak A. Pearlmutter. Fast exact multiplication by the Hessian. *Neural Computation*, 6(1):147–160, 1994.

Shauli Ravfogel, Gilad Yehudai, Joan Bruna, and Alberto Bietti. Geometric factual recall in transformers. *CoRR*, abs/2605.12426, 2026. doi: 10.48550/ARXIV.2605.12426. URL https://doi.org/10.48550/arXiv.2605.12426.

Giovanni Valer, Luigi Gresele, Marco Bronzini, and Emanuele Marconato. Relational linear properties in language models: An empirical investigation. *arXiv preprint arXiv:2605.22532*, 2026. URL https://arxiv.org/abs/2605.22532.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A BASELINE COMPARISONS

The first two tables report Spearman correlations with LRE faithfulness across the LRE Dataset.

Table 4: **Entity lenses and readout baselines across seven models.** Spearman correlation ρ with LRE faithfulness.

Model	Contrastive entity lens	Noncontrastive entity lens	Logit lens	Differential logit lens	Attribute rate
OLMo 3 7B	0.7355	**0.7650**	0.2056	0.4390	−0.4016
GPT-J 6B*	0.4346	0.4196	**0.4669**	0.2020	−0.1035
Mistral 7B	0.6561	**0.6711**	0.6581	0.5186	−0.2202
Gemma 2 9B	**0.6889**	0.6868	0.4728	0.3436	−0.5515
Qwen3 8B Base*	0.6647	0.6232	**0.7908**	0.5952	−0.0557
Qwen3.8 27B*	0.5539	**0.5717**	0.3404	0.4492	−0.0099
Pythia 6.9B*,†	0.5027	**0.5063**	−0.2047	0.1845	0.0041

Bold marks the largest point estimate within each model, not a significant difference. All columns use the contrastive entity lens's same evaluation rows, including demonstrations and references, and the same model-specific year exclusions. *GPT-J, Pythia, Qwen3 8B Base, and Qwen3.8 27B use one leading newline in the neutral prompt (`\n{s} is known for`) to reduce excessive activation norms. The other models use the prompt without this prefix. †Pythia has 2,663 empty filtered prediction sets; Gemma has one. Their attribute rates are assigned zero, retaining every evaluation row.

Readout definitions. For subject state s_e, reference states s_i, and reference Jacobians J_i, the contrastive and noncontrastive entity lenses decode $\mathrm{mean}_i\, J_i(s_e - s_i)$ and $\mathrm{mean}_i\, J_i s_e$, respectively. Logit lens decodes s_e; differential logit lens decodes $s_e - \mathrm{mean}_i\, s_i$. Each relation's score is the negative median $\log_{10}(\mathrm{rank} + 1)$ of the model's answer token.

Attribute rate. Following Geva et al. (2023), we measure overlap between filtered top-50 logit-lens predictions and subject attributes retrieved from Wikipedia. The index uses the union of current test subjects; the source stopword, length, and retrieval filters are retained. Relation scores average the percentage overlap. Undefined rates are handled as noted in the table; matched-row sensitivity results remain in the accompanying data.

NEL has a larger point estimate than CEL in four models. All seven paired 95% relation-bootstrap intervals for their difference include zero; these results do not establish a general advantage for either readout. LRE hyperparameter selection remains in sample.

A.1 OLMo CORPUS-FREQUENCY BASELINES

Table 5: **Frequency baselines for OLMo 3 7B.** Correlations with LRE faithfulness using document counts from OLMo's three published training mixes combined.

Frequency measure	Spearman ρ	p-value
Marginal frequencies		
Subject frequency, c_s	0.1380	0.484
Object frequency, c_o	0.5712	0.001
Co-occurrence		
Joint frequency, c_{so}	0.4568	0.015
Mean of $\log(1 + c_{so})$	0.4191	0.026
$\log(1 + \text{mean}(c_{so}))$	0.4568	0.015
Conditional frequencies		
Forward conditional, $P(o \mid s)$	0.4486	0.017
Reverse conditional, $P(s \mid o)$	-0.4149	0.028
Directional difference, $P(o \mid s) - P(s \mid o)$	0.4826	0.009
Micro-averaged $P(o \mid s)$	0.5504	0.002
Micro-averaged $P(s \mid o)$	-0.1453	0.461

Counts denote binary document presence. Means use the current test rows with both marginals observed; micro averages are ratios of summed counts. The p-values are two-sided and unadjusted.

Counting and aggregation. A document contributes at most one mention per subject/object string. Matching uses normalized literal names with Unicode word boundaries, without alias resolution or a proximity constraint. With subject, object, and joint counts c_s, c_o, and c_{so}, the conditionals are $P(o \mid s) = c_{so}/c_s$ and $P(s \mid o) = c_{so}/c_o$. Macro features average row-level quantities where both marginals are observed. Micro conditionals divide summed joint counts by summed marginals. The log of the mean joint count has the same Spearman correlation as the mean joint count because the transformation is monotonic.

Interpretation and reproducibility. These are exploratory associations. Relations can share subjects and objects, and the reported p-values are unadjusted. Conditional direction also reflects marginal frequency: when $c_{so} > 0$, $P(o \mid s)/P(s \mid o) = c_o/c_s$.

A.2 NEUTRAL-PROMPT VARIATIONS

Table 6 reports prompt variations for the Qwen3 8B Base contrastive entity lens. We tested these variations because initial correlations were low with the neutral prompt used for OLMo, `{s} is known for`. All prompts use the same checkpoint, layer 19, demonstrations, reference sets, 9,747 evaluation rows, gender scoring, relation-prompt answer tokens, and LRE outcomes as Table 4. Correlations use the same 26 relations after the model's two year-relation exclusions. Only the displayed neutral prompt varies; `\n` denotes one leading newline.

After observing unusually large subject-state norms at position zero, we added a leading newline (`\n`) to the original prompt. This moves every subject past position zero while preserving Qwen's original token sequence. The table marks this control with †.

Table 6: **Neutral-prompt variations.** Spearman ρ between contrastive entity lens scores and LRE faithfulness, using the evaluation setup of Table 4. †: leading newline added to address high subject-state norms at position zero.

Qwen3 8B Base: prompt wording and prefix	
Neutral prompt	ρ
`{s} is known for`	0.3430
`{s}`	0.2479
`{s} is`	0.2362
`{s} can be described as`	0.4297
`Information about {s}:`	0.6998
`Facts about {s}:`	0.6725
`Entity: {s} is known for`	0.7331
`\n{s} is known for`†	0.6647

B JACOBIAN–VECTOR PRODUCTS

Both the operator and the lens access J_i only through Jacobian–vector products $J_i u$. This is explicit for the lens in Eq. 5; for the LRE, the expansion in Eq. 3 expresses $W = \text{mean}_i J_i$ as a sum of these products. We compute each product using the standard double-backward construction (Pearlmutter, 1994): differentiate $\langle u', f(h)\rangle$ with respect to h, treating u' as free, and then differentiate with respect to u' in direction u. This takes a constant number of forward and backward passes, while materializing the Jacobian takes d_{model} passes. We use `float32` and eager attention because fused attention kernels do not support second derivatives.

C LAYER SELECTION

Layer 17 is roughly half-way through the network, the region in which such lenses are usable at all (Gurnee et al., 2026), and is the modal best layer for the LRE itself in our sweep (15 of 28 relations; the rest spread over 5, 11, 14, 20 and 23). A single layer is used for *every* relation, without per-relation tuning. However, this does not establish robustness to the layer choice: changing the layer can affect relations differently. A systematic sensitivity analysis is left to future work.

D NUMERIC-OBJECT TOKENIZATION CORRECTION

Both relations with numeric objects (PRESIDENTS_BIRTH_YEAR, PRESIDENTS_ELECTION_YEAR) initially scored faithfulness 1.000. This was spurious. In this tokenizer `" 1743"` segments as `[' ', '174', '3']`: digits never merge with a preceding space. First-token scoring therefore compared the model's top-1 against a bare space token on both sides, and faithfulness 1.000 meant only that the LRE and the model both emit a space.

The fix is to place the space in the prompt, so that the query ends with it and the next token is content. Rescored, the two relations give faithfulness 0.74 and 0.79 and lens scores of -4.44 and -3.45; the model does predict the year correctly once the space is consumed. Including the uncorrected values would have depressed the reported correlation from 0.7355 to 0.416.

E RECOMBINING ANCHORS AND SUBJECT TERMS

This appendix describes the experiments of §7.3 in full.

The decomposition. We write the LRE in the form of Eq. 1, $\beta W s + b_\beta$, and call b_β the *anchor* and $\beta W s = \beta\,\text{mean}_i J_i s$ the *subject term*. The anchor is fixed by the reference subjects alone; the subject term is NEL (Eq. 6) with the Jacobians taken under the relation's prompt. Their sum is the operator exactly, so the *own* arm is the unmodified LRE. The *foreign* arm replaces $\beta W_r s$ by $\beta W_{r'} s$, the paired relation's subject term, and keeps b^r_β. Because the constant $\text{mean}_i J_i s_i$ stays inside the anchor b_β, a foreign subject term imports the other relation's W and nothing else. The *NEL* arm replaces the subject term by $\beta\,\text{NEL}(e)$ under the neutral prompt; *anchor alone* is b_β; the *wrong donor* is $b_\beta + \beta\,\text{NEL}(e')$ for a different subject e'.

Choice of relation pairs. A term can be swapped with nothing else changing only if the two relations share their subjects, and a substitute can only be measured against an operator that produces the answer in the first place. In the benchmark two pairs meet both conditions: COUNTRY_CAPITAL_CITY with COUNTRY_LARGEST_CITY, and PERSON_PLAYS_POSITION_IN_SPORT with PERSON_PLAYS_PRO_SPORT, over the 24 athletes the two share. Two further pairs fail the second condition. COUNTRY_CURRENCY shares its countries with the other country relations, but its own operator recovers the currency for only 1 of 10 countries at $\beta = 3$, 0 at $\beta = 5$ and none at $\beta = 10$; where an operator is this weak there is nothing for a substitute to reproduce. PERSON_FATHER and PERSON_MOTHER would be the strongest test: they share 748 subjects and have answers of the same kind, so the two relations cannot be told apart by the kind of answer. But their faithfulness is 0.07 and 0.07, and the operators essentially never produce the gold object.

Country subjects. A country tests the swap only if its capital differs from its largest city; otherwise one answer satisfies both relations and a hit cannot say which relation won. The benchmark has 9 such countries among the 18 that the two relations share. We extended the set from the list of sovereign states whose capital is not their largest city.[1] Of its 37 states we excluded seven whose capital or largest city the list marks as contested: Israel and Palestine (limited recognition), Taiwan (sovereignty), Nauru (de facto capital only), Bolivia (La Paz is the seat of government), South Africa (three capitals) and Equatorial Guinea (capital in transition). Of the 31 remaining we kept those the model answers correctly for *both* relations under the prompts used here, scoring the model's greedy continuation rather than its first token, with diacritics folded. All 9 benchmark countries and 13 of the 22 added ones pass, giving 22: Australia, Belize, Benin, Brazil, Cameroon, Canada, China, India, Ivory Coast, Kazakhstan, Morocco, Myanmar, New Zealand, Nigeria, Pakistan, Palau, Switzerland, Tanzania, Turkey, the United Arab Emirates, the United States and Vietnam. Every rejected country is one for which the model gives the *same* city for both questions: the capital for the largest city (Ecuador, Gambia, Liechtenstein, Malta, Micronesia, the Philippines, San Marino, Trinidad and Tobago), or the largest city for the capital (Burundi). The filter therefore removes exactly the countries on which the model does not distinguish the two relations.

Protocol. All states are read at layer 17. Each relation's prompt is its first template with eight in-context demonstrations drawn from its own samples outside the candidate set, so no test subject appears in any prompt. Each test subject is read with the remaining test subjects as references, for the operators and for NEL alike. We fix $\beta = 3$. It was chosen by inspecting an earlier version of this analysis on the same 24 athletes and on 12 countries, 7 of which are among the 22, so it is not independent of the evaluation; it is not re-tuned here. At this value the subject term is smaller than the anchor, $\beta\|Ws\|/\|b_\beta\| \approx 0.8$ for the country relations and 0.4–0.6 for the person relations. A hit is a decoded top-1 equal to the first token of the gold object; since the model answers every test country correctly, gold and model answer coincide for the country pair.

Full results. Table 7 adds one arm to Table 3: a subject term estimated under the prompt `"The official currency of {} is the"`, which shares nothing with either country relation. Conditional on the operator being right, the foreign subject term preserves the answer for 11 of 12 capitals and 8 of 12 largest cities, NEL for 7 of 12 and 4 of 12, and the currency subject term for 6 of 12 and 5 of 12.

relation	n	own	foreign	NEL	currency	anchor alone	wrong donor
COUNTRY_CAPITAL_CITY	22	12	12	7	6	0	0
COUNTRY_LARGEST_CITY	22	12	8	4	5	0	0
PERSON_PLAYS_POSITION	24	19	21	18	–	7	6
PERSON_PLAYS_PRO_SPORT	24	22	21	21	–	6	3

Table 7: All arms, $\beta = 3$, each subject read with the others as references. *Currency*: a subject term estimated under the official-currency prompt, which shares nothing with either country relation.

Where the operator fails. Although the model answers all 22 countries correctly, the operator reproduces only 12 per relation. It does much better on the benchmark's nine (7 capitals, 7 largest cities) than on the 13 added countries (5 and 5), which are rarer in text. On those it fails in recognizable ways: it echoes the country's own name (Kazakhstan, Tanzania, Vietnam), names a famous associated city (Paris for Ivory Coast and Morocco, Geneva for Switzerland), or produces a degenerate token. This is consistent with linear structure weakening for entities that are rarer in pretraining (Merullo et al., 2025).

Dependence on the reference sets. In the leave-one-out design above the references are the other test subjects, so they barely vary. To measure how much the results depend on them we held every subject term fixed and drew the *anchor's* eight references at random, 500 times, from subjects outside the test set and outside every demonstration: 16 countries, and a fixed random 150 of the 1,206 eligible athletes in the two sport relations. Table 8 gives the result. The operator and the foreign

[1] `https://en.wikipedia.org/wiki/List_of_countries_whose_capital_is_not_their_largest_city`, retrieved 2026-09-24.

subject term barely move; NEL is the arm that depends on the anchor, most for position, whose worst anchor brings it down to the anchor-alone level.

arm	capital (22)	largest (22)	position (24)	sport (24)
own	10.7 (10–14)	11.3 (10–12)	18.3 (17–20)	22.0 (21–23)
foreign	9.7 (9–11)	9.3 (8–10)	17.3 (15–21)	20.4 (16–23)
NEL	5.2 (1–8)	4.6 (4–5)	14.7 (7–20)	21.5 (16–24)
anchor alone	0.0 (0–1)	0.0 (0–0)	7.0 (7–7)	4.5 (1–7)

Table 8: Mean (range) hits over 500 anchors, each estimated from eight references drawn from outside the test set; subject terms fixed.

Part of the spread for position is composition. The athlete pool has fewer baseball players and more hockey players than the test set, and an anchor with more soccer players gives fewer position hits (Spearman $\rho = -0.32$ across anchors), plausibly because the anchor's typical position leans toward soccer's.

For the country pair we also varied the two reference sets jointly: 40 anchors crossed with 40 sets of references for NEL, drawn from the 30 countries outside the test set. Of the variance in NEL's hits, the anchor's references account for 36% (capital) and 25% (largest city), NEL's references for 38% and 42%, and their interaction for the rest. Neither source dominates.

Scoring. Hits compare the first token of the decoded output with the first token of the gold object. Three cases in the country pair are affected: the model writes São Paulo as "Sao", a different first token, and the gold "New Delhi" begins with "New", so the answer "Delhi" scores as a miss; conversely the first token of Ngerulmud, Palau's capital, is the single letter "N", so a bare "N" scores as a hit. In the own column these roughly cancel. Scoring the full continuation would remove all three.

F ALL RELATIONS

Table 9 gives faithfulness and the lens score for every relation in the study. Table 1 gives object prominence and the lens's top-five hit rate for the 14 relations on which prominence can be measured (Section 6.1).

G THE ANCHORED CONTRASTIVE ENTITY LENS

Here we report on an experiment with an "anchored" contrastive entity lens rather than the non-anchored entity lens defined in Eq. 5. Keeping the anchor requires choosing how to weight the anchor and subject term, which is what β does in Eq. 3.

In sample no such β outperforms dropping the anchor: the correlation rises towards the anchor-free value as β grows, and selection does not help either: choosing β on 15 relations and evaluating on the remaining 13, across 2,000 random splits, beats the anchor-free readout in only 1% of them. We therefore drop the anchor, which leaves a readout with no hyperparameter at all. The decoder normalizes, so once the anchor is gone, scaling cannot change a ranking and β has nothing left to do. This gives the *contrastive entity lens* that we defined in Eq. 5.

In summary, a single β shared by every relation would be a legitimate option, adding nothing relation-specific, but it turns out to have no advantage compared to the simpler form we adopted.

Table 9: All 28 relations, sorted by LRE faithfulness. k is the mean number of subjects per distinct object; A_r is the relation-agnostic lens score, higher meaning closer agreement with the model. †rescored for the numeric-object artifact (Appendix D); *fewer than ten test items, so faithfulness is coarsely quantized.

relation	k	faith.	A_r
country_language*	1.7	1.00	-1.22
country_largest_city*	1.0	0.92	-0.81
landmark_on_continent	189.4	0.92	-2.43
person_native_language	30.6	0.92	-1.69
name_birthplace	3.9	0.97	-1.32
name_religion	6.2	0.97	-1.40
person_plays_pro_sport	63.6	0.95	-1.23
person_plays_instrument	85.5	0.89	-1.72
landmark_in_country	9.2	0.80	-1.69
city_in_country	1.3	0.89	-0.48
person_occupation	26.5	0.71	-2.74
country_capital_city*	1.0	0.88	-0.85
name_gender*	9.5	0.53	-1.04
presidents_birth_year†	1.3	0.74	-4.44
presidents_election_year†	1.1	0.79	-3.45
product_by_company	17.4	0.68	-1.88
person_plays_position_in_sport	68.0	0.61	-3.64
person_band_lead_singer*	1.0	0.86	-1.49
person_university	1.3	0.59	-4.38
company_hq	4.1	0.54	-4.23
superhero_person	1.0	0.50	-2.62
star_constellation	11.7	0.47	-3.50
pokemon_evolutions	1.0	0.32	-3.45
country_currency	1.3	0.40	-4.49
superhero_archnemesis	1.1	0.30	-3.90
person_father	1.0	0.07	-3.71
company_ceo	1.0	0.05	-4.50
person_mother	1.0	0.07	-4.18