PLAUSIBILITY CONFOUNDS LINEAR TRUTH PROBES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has shown that large language models have linear directions in their internal representations that can distinguish true from false factual statements. Conventional studies use labeled true and false statements to probe the truth direction. Surprisingly, we find that a direction probed solely by our plausible false statements and random negatives, *without using any true statement* to fit the direction, can still achieve strong truth-discrimination performance: its AUC on standard truth-probing benchmarks reaches $> 95\%$ of that achieved by a conventional truth direction trained with true statements. We further identify a generalization risk for standard truth probes on more challenging falsehoods: replacing random false statements in the test set with plausible falsehoods that better resemble realistic errors will reduce test AUC. In the more realistic setting of single-statement truth classification, plausible falsehoods yield false-positive rates 2.6–7.6 times those of random negatives at a fixed threshold on main datasets. We investigate this plausibility confound across six datasets spanning statements with entity answer slots, free-form claims, and long-form generated statements. Based on these observations and analyses, we propose a threshold-free local contrastive readout that compares each statement with a plausible alternative from the same factual context. On datasets with identifiable answer slots and candidate entities of given types, this method substantially reduces single-statement false-positive rates at matched true-positive rates.

1 INTRODUCTION

Large language models (LLMs) often generate fluent but factually incorrect statements, motivating methods that can assess factuality without relying solely on the model's observable output. A particularly promising line of work probes the model's internal activations for representations associated with truth. Marks & Tegmark (2024) showed that the residual-stream representations of simple factual statements exhibit striking linear structure: true and false statements can often be separated by a single direction, probes trained on one factual domain can transfer to others, and intervening along such directions can causally alter the model's treatment of statements as true or false. They further showed that simple difference-in-means, or mass-mean, probes can recover directions that generalize competitively while being more causally implicated than standard discriminative probes. These findings suggest an appealing possibility: rather than asking an LLM whether a claim is true—or trusting the answer it generates—we may be able to inspect its internal representation of the claim directly.

We pose a testable challenge to the interpretation of these directions: *can a direction fitted without a single true statement nevertheless distinguish true from false statements?* Surprisingly, a direction fitted only to our constructed plausible falsehoods and random falsehoods achieves more than 95% of the conventional truth probe's AUC on standard truth-discrimination benchmarks. This provides a concrete test of whether strong truth-discrimination performance alone identifies a representation specific to truth.

Marks & Tegmark (2024) explicitly considered one version of this concern by asking whether their truth direction was merely a text-likelihood direction. They constructed a likely control dataset consisting of nonfactual, unconditioned generations, and showed that probes trained to distinguish likely from unlikely text do not reproduce the generalization behavior of truth probes. Their results

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.

inspire our further exploration and we figure out a key reason as these nonfactual statements are out of domain relative to the factual statements used to probe truth. Our test therefore addresses a different prediction: whether plausibility varies systematically along the truth direction within factual statements whose truth labels are held fixed.

Another line of subsequent work has found that truth directions do not always generalize reliably across models or more challenging task settings (Bürger et al., 2024; Bao et al., 2025; Orgad et al., 2025), and are sensitive to context (Schouten et al., 2025; Long et al., 2025). These generalization failures motivate us to investigate potential confounding factors in truth directions. We focus on plausibility because the next-token prediction objective used in LLM pretraining maximizes the likelihood of observed text, giving us reason to investigate plausibility as a potential confound within the factual semantic space of interest.

To study this question, we augment the factual benchmarks of Marks & Tegmark (2024) with near-miss falsehoods, i.e., for each factual slot, we use an LLM's conditional distribution to identify a high-probability incorrect candidate of the same relation or semantic type. This yields three kinds of statements: a true factual statement, the benchmark's original random or random-like false statement, and a model-plausible but factually incorrect alternative. Crucially, this construction allows us to estimate a plausibility direction without using any true statements at all since we contrast plausible-false with random-false examples to construct the plausibility direction, which is shown strongly aligned with the conventional truth direction. Conversely, the standard truth direction also substantially separates plausible falsehoods from random falsehoods. These results indicate that relative to random negatives, true facts and plausible falsehoods move representations in remarkably similar directions.

This alignment matters because the false statements used in standard truth-probing benchmarks are often much easier than the errors that matter in actual deployment. For example, a hallucinated statement challenging LLMs is typically not caused by a random object, but instead often an alternative that is logically coherent and much more probable than most candidates. When we replace benchmark random negatives with near-miss falsehoods mined once by a reference LLM, conventional truth-probe AUC decreases remarkably, plausible falsehoods are accepted as true at 2.6–7.6 times the false-positive rate of standard random negatives in single statement truth classification. Our further analysis show that straightfoward projection-based removal of the plausibility direction does not yield a clean purification of the truth direction. We view this as evidence for a partially entangled, potentially superposition-like representation in which factual correctness and factual plausibility share representational structure rather than occupying cleanly independent axes.

Based on the observations and analysess, We design a threshold-free method named local contrastive readout to mitigate the plausibility confound in truth discrimination. It compares each statement with a model-selected plausible alternative from the same factual context, keeping the truth direction unchanged. On datasets with identifiable answer slots and candidate pools, this readout can largely reduce false positives on both standard and plausible negatives at matched true-positive rates.

In summary, our work makes the following main contributions:

- We identify and characterize a factual plausibility direction using only plausible-false and random-false statements, and show that it is strongly aligned with conventional truth directions despite never being trained on a true example.
- We show that this alignment exposes a substantial blind spot in standard truth-probe evaluation: near-miss falsehoods produce remarkable AUC degradation and a 2.6–7.6$\times$ increase in fixed-threshold false-positive rate, motivating the use of plausible-negative stress tests and single-statement operating metrics alongside standard AUC.
- We show that plausible falsehoods have substantial cross-model transferability and that their interaction with truth is not cleanly removed by straightforward post-hoc processing, suggesting that plausibility is not merely a dataset artifact but a structural confound that should be explicitly considered when interpreting and deploying truth probes.
- We propose a threshold-free local contrastive readout for statements with identifiable answer slots, reducing false-positive rates on both standard and plausible negatives at matched recall without retraining the truth probe.

2 Related Work

Azaria & Mitchell (2023) showed that classifiers trained on internal activations can distinguish true from false statements, while Marks & Tegmark (2024) demonstrated that simple difference-of-means directions generalize across several factual datasets and can causally influence model outputs. Bürger et al. (2024) decomposed the truth subspace into a general truth component and a polarity-sensitive component, yielding probes that are robust to negation. Bao et al. (2025) further studied the consistency and transfer of truth directions across models, logical transformations, and question-answering settings, finding substantial variation in their generalizability.

Unsupervised approaches such as Contrast-Consistent Search (Burns et al., 2023) can recover salient features that need not correspond uniquely to truth, a limitation analyzed further by Farquhar et al. (2023). More recently, Bertolazzi et al. (2026) showed that logical validity and semantic plausibility can be strongly entangled in representation space for LLM reasoning, demonstrating that internal features related to reasoning logics may encode correlated semantic factors. Our factual-probing analogue provides tighter control by holding the subject and relation fixed and varying only the asserted value, contrasting model-mined plausible falsehoods with random falsehoods. Levinstein & Herrmann (2024) similarly argue that standard probing datasets may be insufficient to distinguish truth, belief, and other correlated properties. We identify factual plausibility as a contaminating property and quantify its overlap with the truth direction and its effect on probe errors. For evaluation, prior truth-direction work validates through transfer and causal interventions (Marks & Tegmark, 2024) but reports no matched null, whereas we provide one using random balanced partitions of the same false statements (Section 3.3). Appendix A.1 provides the full related-work discussion.

3 Experimental Setup

3.1 Datasets

In our work, we use six datasets covering various domains. Among them, three are used as the core datasets for most of the controlled replacement experiments, and the other three are used for broader evaluation and extended analyses.

Datasets for controlled replacement experiments. We use the factual datasets distributed with the probing framework of Marks & Tegmark (2024). **Cities** contains city–country statements, such as "The city of Krasnodar is in Russia" (true) and "The city of Krasnodar is in South Africa" (false). There are 748 cities before near-miss filtering. **CounterFact**, originally introduced by Meng et al. (2022), contains subject–relation–object facts covering several relations. We use its declarative true/false version; for example, "The mother tongue of Danielle Darrieux is French" and its false counterpart with "English." We select candidate replacements from objects that occur with the same relation. **sp_en_trans** contains Spanish–English translation statements, such as "The Spanish word 'uno' means 'one'." Its candidate answers are English glosses from the dataset. These three datasets support comparisons among a true statement, a standard false statement, and a plausible false statement with the same subject and relation.

Datasets for broader evaluation. **Companies** includes labeled factual statements about companies, headquarters and business activities. For example,"Meta Platforms has headquarters in United States" (true) and "Thermo Fisher Scientific engages in the manufacture and sale of semiconductor products" (false). **CommonClaim** contains free-form claims originating from Casper et al. (2023), such as "Spiders can use surface tension to walk on water" (true) and "Oxygen is actually purple" (false). We use the versions distributed by Marks & Tegmark (2024). For these two datasets, we sample up to 1,000 statements per dataset, with equal numbers of true and false statements, and use a seeded 50/50 train/test split. **hallucination-probes** provides span-level factuality annotations for long-form generations (Obeso et al., 2025).We use the Llama-3.1-8B-Instruct responses to LongFact (Wei et al., 2024) prompts and keep spans labeled *Supported* or *Not Supported*, while preserving their original context.

3.2 GENERATING PLAUSIBLE FALSE STATEMENTS

Controlled construction. For each factual item i, let p_i denote the prefix containing its subject and relation, a_i^{T} its labeled correct answer, and $\mathcal{C}_i$ the candidate answer pool. We construct three statements: the true statement x_i^{T}, the standard random or random-like false statement x_i^{R}, and a model-mined plausible false statement x_i^{P}. Candidates are countries for **Cities**, relation-matched objects for **CounterFact**, and English glosses for **sp_en_trans**. We retain the benchmark's supplied false answer where available. Because the published translation dataset does not pair every true word with a false counterpart, we assign a seeded random incorrect gloss to each true word.

Let $c_i(a) = (c_{i,1}, \ldots, c_{i,m_i(a)})$ be the tokenized continuation that completes prefix p_i with answer a (including the closing punctuation), and $m_i(a)$ denote the total sequential length. For a mining model M and a given prefix p_i, we define the conditional plausibility for an answer a as

$$\ell_M(a \mid p_i) = \sum_{t=1}^{m_i(a)} \log P_M(c_{i,t} \mid p_i, c_{i,<t}) \,, \tag{1}$$

and construct the plausible-but-false answer [1] by

$$a_i^{\mathrm{P}} = \underset{a \in \mathcal{C}_i \setminus \{a_i^{\mathrm{T}}\}}{\arg\max}\, \ell_M(a \mid p_i). \tag{2}$$

We score the full continuation, including multi-token answers, rather than comparing only the first answer token. Here $\ell_M(a \mid p_i)$ measures the probability of the complete continuation, which is not length-normalized because we aim to mine the falsehood the model is most likely to produce. Candidate pools $\mathcal{C}_i$ are constructed directly from answer values already present in each benchmark, while preserving the answer type or the factual relation, according to the property of each benchmark. We describe more details in Appendix.

Controls and data quality. We hold the prefix, sentence template, and punctuation fixed and change only the answer. We additionally flag an item as *ranking-ambiguous* when $\ell_M(a_i^{\mathrm{P}} \mid p_i) > \ell_M(a_i^{\mathrm{T}} \mid p_i)$. Ranking-ambiguity rates are 0% for **Cities**, 0% for **sp_en_trans**, 23% for **CounterFact**, and 41% for **Companies**. The primary held-out direction analysis excludes these items from its test split, while retaining them in the released data. This filter identifies disagreement between the mining model's ranking and the factual label, but does not change the label or imply that the fact itself is ambiguous. We also remove multi-valued relations, which cover 9.8% of **CounterFact** items. Across all the evaluated models, removing these relations changes each model's true-versus-standard-false AUC, true-versus-plausible-false AUC, and plausible-negative FPR by less than 0.01. We also did alias screen to avoid correct answers with different spellings being selected as plausible falsehoods. We found no alias foils in **Cities** or **sp_en_trans**, and in **Companies** 65/578 case variants were found and fixed. Appendix A.2 reports the tie checks, alias-screening and normalization procedure, and per-dataset multi-answer audit, including aggregate results before and after removing multi-valued relations.

Truth and plausibility probes. For direction comparisons, we remove items for which the mined and standard false answers coincide. We split the answer-slot data 50/50 by factual item with a fixed seed, keeping all statement variants of an item in the same split. Let $h(x)$ be the residual-stream representation at the final token of statement x at the selected model layer, and let $\mu_{\mathrm{T}}, \mu_{\mathrm{R}}, \mu_{\mathrm{P}}$ be the corresponding training-set means. Following the mass-mean construction of Marks & Tegmark (2024), we define

$$d_{\mathrm{truth}} = \frac{\mu_{\mathrm{T}} - \mu_{\mathrm{R}}}{\|\mu_{\mathrm{T}} - \mu_{\mathrm{R}}\|_2}, \qquad d_{\mathrm{plaus}} = \frac{\mu_{\mathrm{P}} - \mu_{\mathrm{R}}}{\|\mu_{\mathrm{P}} - \mu_{\mathrm{R}}\|_2}. \tag{3}$$

Both endpoints used to fit d_{plaus} are false: no true-statement representations enter its fit, although factual labels are used when selecting incorrect candidates. We measure geometric alignment by $d_{\mathrm{truth}}^{\top} d_{\mathrm{plaus}}$ and evaluate scores $s_d(x) = d^{\top} h(x)$ on three held-out tasks: true versus standard false, plausible false versus standard false, and true versus plausible false. Cross-domain tests apply a direction learned in one domain to another domain within the same model. We use mass-mean probes to keep alignment, transfer, and ranking evaluations tied to the same explicitly defined directions.

[1] We mostly use the single term *plausible* for simplicity for the remaining of the paper.

For fixed-threshold truth classification, we also fit logistic probes on the standard labeled training distribution and retain the probability threshold of 0.5 when evaluating plausible negatives. A logistic probe outputs $q(x) = \sigma(w^\top h(x) + b)$, where σ is the sigmoid function and w, b are learned by logistic regression. This provides a standard supervised classifier for evaluating probability calibration and fixed-threshold reliability under plausible-negative replacement. Thus, Table 2 reports mass-mean AUC and separately fitted logistic classification results. In Section 4.3, we instead keep the mass-mean direction fixed to isolate the effect of changing the global or local decision rule.

Likelihood-based plausibility measurement. For more general free-form statements, we use length-normalized statement log-probability as a proxy for naturalness. More generally, for a scored token sequence $z = (z_1, \ldots, z_m)$ and preceding context u, we compute

$$\bar{\ell}_M(z \mid u) = \frac{1}{m} \sum_{t=1}^{m} \log P_M(z_t \mid u, z_{<t}). \tag{4}$$

For standalone claims (e.g., **CommonClaim**), z comprises the scored statement tokens and u is the initial model context; for annotated generations (e.g., **hallucination-probes**), z is the span and u is its preceding prompt and response context. We partition held-out false statements, or *Not Supported* spans, into likelihood terciles and compare their probe scores and false-positive rates at the same fixed threshold.

3.3 Permutation Null Test

We compare the plausibility direction against a matched null fitted from random balanced partitions of the same training false statements using the same mass-mean procedure. The train/test split, representations, and held-out evaluation labels remain fixed. Across 300 permutations, we evaluate orientation-invariant AUC and report the null mean, standard deviation, and 97.5th percentile to assess whether the plausibility grouping carries more discriminative structure than arbitrary groupings. Appendix A.3 gives the full procedure, including matched projection controls and the test's interpretive limits.

4 Results and Analyses

The core comparisons include Llama-3-8B(Grattafiori et al., 2025), Llama-2-13B(Touvron et al., 2023), Mistral-7B-v0.3(Jiang et al., 2023), OLMo-2-7B(OLMo et al., 2025), Qwen2.5-14B(Qwen et al., 2025), and Gemma-2-9B(Team et al., 2024).

4.1 Alignment between Plausibility and the Truth Direction

A false-only direction recovers much of the truth signal. Our core finding is that, the plausibility direction d_{plaus} defined in Equation 3, despite being fitted entirely from false statements, is strongly aligned with the conventional truth direction d_{truth} fitted from true and false statements in most experimental groups. For example, on **CounterFact**, their cosine similarity is 0.803–0.879, and the false-only direction achieves 0.902–0.954 AUC on true versus standard false statements (T/R), close to the truth probe's 0.930–0.969, as shown in the first three columns of Table 1. Across all three domains, the plausibility direction retains approximately $> 95\%$ of the truth probe's AUC. On P/R, both directions have lower AUC than on T/R, but d_{plaus} still exceeds the permutation null's 97.5th percentile in 16 of 18 settings, with exceptions for Qwen2.5-14B and Gemma-2-9B on **sp_en_trans**.

Transferability of the plausibility direction. We further find that the alignment between the plausibility and truth directions is not necessarily dependent on the model's own mined falsehoods. We test cross-model transfer of the plausibility probe by using a shared set of plausible negatives mined once by Llama-3-8B. On **CounterFact**, the false-only direction achieves 0.882–0.944 AUC on true versus standard false statements (T/R), still close to the truth probe's 0.913–0.952 and retains $> 88\%$ of the truth probe's AUC. In addition, d_{plaus} shows some cross-dataset transferability. We find that the transfer effect is mainly dominated by the target domain, e.g., T/R performance remains high if the target is **Cities** or **sp_en_trans**, regardless of the source domain, whereas transfer

Table 1: **Truth and plausibility directions on held-out statements.** T/R denotes true versus standard false; P/R denotes plausible versus standard false. Columns report cosine similarity and orientation-invariant AUC for d_{truth} (d_T) and d_{plaus} (d_P). For example, d_P: T/R means applying the probed d_P direction on the test data of true versus standard false. Null is the 97.5th percentile of 300 random-partition AUCs on P/R. Each evaluated model mines its own plausible negatives.

Dataset	Model	Cosine	d_T: T/R	d_P: T/R	d_T: P/R	d_P: P/R	Null
CounterFact	Llama-3-8B	0.805	0.952	0.943	0.671	0.737	0.559
	Llama-2-13B	0.824	0.930	0.912	0.716	0.761	0.565
	Mistral-7B	0.811	0.969	0.954	0.698	0.760	0.581
	OLMo-2-7B	0.879	0.939	0.941	0.709	0.753	0.580
	Qwen2.5-14B	0.858	0.932	0.902	0.677	0.707	0.555
	Gemma-2-9B	0.803	0.933	0.941	0.649	0.748	0.552
Cities	Llama-3-8B	0.479	1.000	0.963	0.619	0.850	0.759
	Llama-2-13B	0.732	1.000	0.999	0.717	0.809	0.742
	Mistral-7B	0.748	1.000	0.997	0.740	0.816	0.796
	OLMo-2-7B	0.809	1.000	0.998	0.809	0.853	0.837
	Qwen2.5-14B	0.685	1.000	0.994	0.726	0.789	0.775
	Gemma-2-9B	0.484	0.998	0.999	0.578	0.887	0.722
sp_en_trans	Llama-3-8B	0.819	1.000	0.999	0.852	0.871	0.847
	Llama-2-13B	0.847	1.000	0.999	0.834	0.852	0.805
	Mistral-7B	0.825	1.000	1.000	0.880	0.905	0.830
	OLMo-2-7B	0.886	1.000	1.000	0.901	0.916	0.875
	Qwen2.5-14B	0.823	1.000	1.000	0.718	0.780	0.784
	Gemma-2-9B	0.796	1.000	1.000	0.720	0.764	0.812

into the harder **CounterFact** domain is consistently weaker. As in the in-domain evaluation, P/R AUC is always lower than T/R AUC. We put the complete cross-model and cross-dataset results in Appendix A.4.

Separating truth and plausibility incurs a cost. We next ask whether the plausibility component can be removed from the truth representation while preserving truth-discrimination performance. We test two linear separation strategies on CounterFact. These tests use the shared Llama-3-8B negatives. First, we directly remove the component of the mass-mean truth direction along the plausibility direction, as $d_{\text{truth}\perp\text{plaus}} = d_{\text{truth}} - \text{Proj}_{d_{\text{plaus}}}(d_{\text{truth}})$. Second, rather than projecting the learned probe itself, we remove the plausibility direction from every representation, $h_{\perp\text{plaus}}(x) = h(x) - \text{Proj}_{d_{\text{plaus}}}(h(x))$, and refit a logistic truth reader in this orthogonal feature subspace. Both procedures reduce truth-task performance. The model-level AUCs and losses are reported in Appendix Tables 6 and 7. Across models, direct projection produces a 0.038–0.091 loss in T/R AUC, while refitting in the orthogonal feature space preserves more performance but still incurs a smaller, consistent 0.015–0.038 loss. This pattern implies that truth and plausibility share information through a deeper coupling in the representation, making them difficult to disentangle losslessly with simple geometric operations.

Factual plausibility beyond generic token likelihood. The **likely** control of Marks & Tegmark (2024) contrasts nonfactual continuations whose final token is the most likely or the 100th-most-likely choice. Their negative result proves that a truth direction is not simply a generic token-likelihood direction. Our main findings do not conflict with theirs. We compare alternative factual bindings while holding the subject, relation, and truth label fixed, exposing plausibility-related variation within factual statements. In particular, plausible and standard falsehoods share the same negative truth label but differ systematically along the learned plausibility direction, providing an in-domain test that is complementary to the generic likelihood control. As illustrated in Fig. 1 in Appendix A.6, the PCA visualization compares generic nonfactual representations with the factual statements used to fit the truth and plausibility directions. Within the factual domain, we further find that length-normalized statement log-probability correlates positively with a logistic truth probe's predicted probability of truth. This association holds across the evaluated Llama models on **Companies** and **CommonClaim** with $r = 0.26 - 0.42$. We put detailed descriptions and numbers in Appendix A.6.

Table 2: **Plausible negatives expose fixed-threshold errors on CounterFact.** AUC columns use the mass-mean truth direction; TPR and FPR columns use logistic probes at the unchanged probability threshold of 0.5. All columns use shared Llama-3-8B negatives. R denotes standard false statements and P plausible false statements.

	Mass-mean AUC		Logistic classification		
Model	True/R	True/P	TPR	FPR R	FPR P
Llama-3-8B	0.952	0.876	0.958	0.040	0.305
Llama-2-13B	0.927	0.828	0.948	0.140	0.370
Mistral-7B	0.952	0.856	0.950	0.083	0.393
OLMo-2-7B	0.913	0.798	0.877	0.122	0.390
Qwen2.5-14B	0.932	0.821	0.948	0.107	0.372
Gemma-2-9B	0.923	0.831	0.938	0.075	0.320

4.2 Standard Truth Evaluation Hides Plausibility Risk

Plausibility risk in truth discrimination. The findings in the last section suggest risks related to plausible falsehoods in truth evaluation that are not captured by random negatives used in the current benchmarks. We therefore quantify the extent to which plausible falsehoods are accepted by truth directions that are probed with and perform well on standard true/false statements. Specifically, based on the same held-out items as in Section 4.1, we create an additional true-vs-plausible test set by replacing the standard negatives with plausible false statements. While existing papers mostly focus on the standard truth benchmarks, we suggest paying equal attention to the plausibility evaluation since a real-world challenge for truth probe is to detect those hallucinated statements which are not obviously false.

To be more practical, we use the fixed plausible false statements generated by Llama-3-8B rather than mining each model's own plausibility falsehoods. Besides the standard AUC metric which measures the ranking performance, we also consider a more realistic evaluation scenario where a single statement needs to be classified as true or false based on a fixed threshold. We particularly care about the false-positive rate (FPR), which measures the probability of incorrectly classifying a false statement as true. For a logistic truth probe with output $q(x)$, we define

$$\mathrm{FPR}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[q(x_i^k) > 0.5\right], \qquad k \in \{\mathrm{R}, \mathrm{P}\}, \tag{5}$$

where R and P denote standard and plausible false statements. The fitted classifier and threshold remain unchanged when the negative set is replaced.

The analysis is mainly based on the more challenging **CounterFact** dataset. As shown in Table 2, by replacing the standard negatives with plausible alternatives for Llama-3-8B, the mass-mean AUC decrease from 0.952 to 0.876. More importantly, the logistic probe's FPR rises from 0.040 to 0.305, which is a $7.6\times$ amplification of false positives. Furthermore, we find that the standard-data expected calibration error (ECE) is only 0.02–0.07, which looks good, yet plausible falsehoods are frequently accepted as true, implying thattruth probes can be *calibrated but wrong*. This exposes a failure of the evaluation protocol: standard-distribution AUC and calibration do not establish fixed-threshold reliability under plausible falsehoods in truth classification. For **Cities** and **sp_en_trans**, the test AUC is almost saturated and does not decrease as much, but the FPR on plausible negatives is still significantly higher than on standard negatives. Detailed results for the above evaluations are provided in Appendix A.7.

Broader evaluation of plausibility risk. We observe the similar risk of plausible falsehoods by using the more general statement likelihood proxy defined by Equation 4 without relying on entity replacement (This can also be applied to datasets with entity slots though). Specifically, we calculate FPR for the most-likely third and least-likely third of false statements, using the same logistic truth probes learned from the standard training set and fixed threshold of 0.5 for prediction. We further assess the plausibility risk on a more realistic dataset **hallucination-probes**, which contains 6,025 labeled test spans from 380 responses generated by Llama-3.1-8B-Instruct. By using the same analytical method, we find that the most-likely third of unsupported spans is accepted as supported at 0.49–0.57, versus 0.35–0.42 for the least-likely third, consistently across all probe con-

figurations. The results show that FPR amplification persists across increasingly realistic settings including structured replacements, free-form claims, and real generated spans. Detailed experimental setting and results are provided in Appendix Section A.8.

Implications from the results. The experiments above show that directly applying conventional truth-probe classifiers to single-statement truth discrimination can lead to high FPR on plausible falsehoods. Here we do not claim that this represents the performance ceiling of conventional truth probes, since the threshold can always be adjusted to trade off TPR against FPR. Our main point is that directly applying a classifier trained on T/R to T/P can pose a substantial generalization risk. A more fundamental issue is that there is also a marked decrease in threshold-independent AUC, indicating that more false-but-plausible examples receive higher truth scores than genuinely true statements. We think this is a fundamental limitation arising from the standard truth-score readout, which uses a global truth direction as the reference and compare the magnitudes of individual statements' projections onto it. If each statement is viewed as a small data domain, some domains may naturally be more aligned with this global truth direction, such that the projection contributed by their plausibility confound alone may exceed the contribution of another domain's clean truth component. We will explore a preliminary solution to address this in the next subsection.

4.3 Mitigating the Plausibility Confound in Truth Detection

From a global threshold to a local comparison. In Section 4.1 we show that removing a single plausibility direction can harm useful truth information, and Section 4.2 shows that a fixed global threshold is sensitive to the negative distribution. These observations motivate changing the inference-time readout with a local comparison, i.e., contrasting each statement with its own plausible alternative. Figure 2 in Appendix A.9 illustrates how local comparison preserves more true positives than raising a global threshold at the same false-positive rate.

Local contrastive readout. We keep the standard mass-mean truth direction unchanged. For a statement x with representation $h(x)$, its score and baseline decision (i.e., the global readout) are

$$s(x) = d_{\text{truth}}^{\top} h(x), \qquad \widehat{y}_{\text{global}}(x) = \mathbf{1}[s(x) > \tau]\,. \tag{6}$$

Let $x = \text{render}(p, a)$ contain an identifiable answer a and factual prefix p, with candidate set $\mathcal{C}(p)$. Using the conditional sequence score from Equation 2, we select the highest-probability *different* candidate,

$$a^{\star}(x) = \underset{b \in \mathcal{C}(p) \setminus \{a\}}{\arg\max}\, \ell_M(b \mid p), \qquad x' = \text{render}(p, a^{\star}(x)). \tag{7}$$

We then classify the input by its score relative to this alternative:

$$\widehat{y}_{\text{local}}(x) = \mathbf{1}[s(x) - s(x') > 0] = \mathbf{1}\left[d_{\text{truth}}^{\top}\big(h(x) - h(x')\big) > 0\right]. \tag{8}$$

Unlike plausible-negative construction for evaluation, selecting the plausible counterpart x' for local contrastive readout excludes the input answer instead of a known correct answer. In other words, it does not require ground-truth label. Therefore, a false input may be compared with the correct answer if the model ranks it highest among alternatives. The rule also requires no fitted global threshold or probe retraining.

Comparison between global and local readouts. We consider two global-threshold versions of the same mass-mean probe: a threshold calibrated on training data, and a threshold chosen to match the TPR of our local contrastive readout, so that their FPR are directly comparible. The Table 3 reports means over the six core models, with FPR measured separately on standard and plausible negatives. F1 is evaluated on true versus standard false (T/R), true versus plausible false (T/P), and a pooled task containing one true and two negative variants per item.

Compared with Global, matched TPR, local contrastive readout reduces both FPR R and FPR P at the same TPR across all three datasets. For example, on **CounterFact** with TPR of 0.873, our local contrastive readout reduces FPR R from 0.179 to 0.035 (about 5.1 times) and FPR P from 0.432 to 0.134 (about 3.1 times). These consistent improvements support the effectiveness of local contrastive readout in reducing instance-level plausibility interference in truth discrimination, rather than trading errors on plausible falsehoods against errors on random falsehoods.

Table 3: **Local contrastive readout improves single-statement classification.** Values are means over six core models. Each comparison uses the same mass-mean truth direction. The matched-TPR global baseline has a test-set oracle threshold. R denotes standard negatives and P plausible negatives; pooled F1 uses both. Bold marks the best FPR and F1 values within each dataset, including ties (higher F1 and lower FPR are better).

Dataset	Decision rule	TPR	FPR R	FPR P	F1 T/R	F1 T/P	F1 pooled
CounterFact	Global, train-calibrated	0.916	0.235	0.512	0.852	0.755	0.689
	Global, matched TPR	0.873	0.179	0.432	0.851	0.758	0.704
	Local contrastive	0.873	**0.035**	**0.134**	**0.915**	**0.870**	**0.855**
Cities	Global, train-calibrated	1.000	0.048	0.125	0.977	0.941	0.921
	Global, matched TPR	0.997	0.012	0.037	0.993	0.981	0.975
	Local contrastive	0.997	**0.000**	**0.003**	**0.999**	**0.997**	**0.997**
sp_en_trans	Global, train-calibrated	0.998	0.008	0.211	0.995	0.904	0.901
	Global, matched TPR	1.000	0.004	0.120	0.998	0.944	0.943
	Local contrastive	1.000	**0.000**	**0.000**	**1.000**	**1.000**	**1.000**

5 Limitations and Future Work

The coupling between truth and plausibility remains unresolved. We find that the plausibility direction is strongly aligned with the conventional truth direction, and that plausibility-related variation appears at both the domain and instance levels. However, our experiments do not identify the precise mechanism that couples these signals in the representation, nor do they provide a clean procedure for disentangling plausibility from truth while preserving useful factual information. Projection and constrained readers characterize this difficulty but do not solve it. Future work should investigate the underlying features and circuits with more targeted controls and causal interventions, and develop disentanglement methods that are validated on held-out domains and realistic negative distributions.

The local contrastive readout has a restricted application scope. Our local contrastive readout provides an effective test of the hypothesis that truth directions contain domain-specific and instance-level plausibility confounds: comparing a statement with a model-selected alternative improves fixed-operating-point performance on the structured benchmarks. However, from an application perspective, the method requires an identifiable entity slot and a candidate pool from which an alternative can be constructed. Therefore, it does not apply directly to arbitrary free-form claims or long-form hallucination spans. Extending the method to those settings will require reliable context-preserving alternative generation, along with safeguards for paraphrases, aliases, multiple valid answers, and additional inference cost.

6 Conclusion

Truth probes are sensitive to more than factual correctness. By constructing plausible false statements that preserve the subject and relation while changing the asserted answer, we show that a direction learned only from false statements can align strongly with the conventional truth direction. This plausibility signal creates a substantial fixed-threshold false-positive risk and also cause standard AUC decrease. These results suggest that truth-probe evaluation should include model-plausible negatives and report operating-point metrics alongside ranking metrics. When an entity slot and candidate pool are available, a local contrastive readout can reduce this risk without retraining or rotating the truth direction. Together, the findings recast plausibility as a structural part of truth-probe geometry and provide a concrete basis for evaluating and improving factuality probes in more realistic settings.

AI USE STATEMENT

We used generative AI tools, including ChatGPT and Claude, to assist with manuscript drafting and revision, literature summarization, and code generation and debugging. All AI-assisted code, analyses, citations, and scientific claims were reviewed and verified by the authors. Generative AI was not used to design ideas,generate experimental data, determine labels, or replace the authors in quantitative experiments. The authors take full responsibility for the final content of this work.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yuntai Bao, Xuhong Zhang, Tianyu Du, Xinkui Zhao, Zhengwen Feng, Hao Peng, and Jianwei Yin. Probing the geometry of truth: Consistency and generalization of truth directions in LLMs across logical transformations and question answering tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 682–700, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.38. URL https://aclanthology.org/2025.findings-acl.38/.

Leonardo Bertolazzi, Sandro Pezzelle, and Raffaella Bernardi. How language models conflate logical validity with plausibility: A representational analysis of content effects. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9316–9347, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.454. URL https://aclanthology.org/2026.findings-acl.454/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.12831.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*, 2023.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023. URL https://arxiv.org/abs/2312.10029.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. 2025.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02207.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 2024. doi: 10.1007/s11098-023-02094-3. URL https://doi.org/10.1007/s11098-023-02094-3. Published online 17 February 2024.

Jane Li, Abhinav Patil, and Kyle Rawlins. Propositional compositionality in neural language models. In Rob Voigt, Alex Warstadt, Naomi Feldman, and Tal Linzen (eds.), *Proceedings of the Society for Computation in Linguistics 2026*, pp. 365–368, San Diego, CA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-412-5. doi: 10.18653/v1/2026.scil-main.34. URL https://aclanthology.org/2026.scil-main.34/.

Xianxuan Long, Yao Fu, Runchao Li, Mu Sheng, Haotian Yu, Xiaotian Han, and Pan Li. When truthful representations flip under deceptive instructions? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16315–16335, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.826. URL https://aclanthology.org/2025.emnlp-main.826/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*, 2025. URL https://arxiv.org/abs/2509.03531.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.02707.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Stefan F. Schouten, Peter Bloem, Ilia Markov, and Piek Vossen. Truth-value judgment in language models: 'truth directions' are context sensitive. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2404.18865.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

A Appendix

A.1 Extended Related Work

A.1.1 Mechanistic Interpretability of LLMs

Mechanistic interpretability seeks to explain model behavior through the internal representations and computations that produce it. Building on feature visualization and circuit analysis (Olah et al., 2020; Elhage et al., 2021), studies of transformers have identified functional roles for attention heads and traced how components interact to implement specific behaviors. Causal interventions, including activation and path patching, test these roles by measuring how changes to intermediate activations affect model outputs (Wang et al., 2022). Automated circuit discovery extends this approach by searching for subnetworks that preserve a target behavior (Conmy et al., 2023). Complementary work uses sparse autoencoders and dictionary learning to extract interpretable features from high-dimensional activations (Cunningham et al., 2023; Bricken et al., 2023). Together, these approaches connect the organization of learned features to the computational mechanisms underlying model predictions.

Within this broader program, a central question is how LLMs store and access knowledge about the world. Geva et al. (2021) characterize feed-forward layers as key-value memories, while Meng et al. (2022) use causal tracing and targeted model editing to localize and modify factual associations. Beyond individual facts, Gurnee & Tegmark (2024) identify linear representations of spatial and temporal coordinates across multiple scales and entity types, providing evidence for structured internal representations of world knowledge. These findings motivate studying knowledge through representation geometry, while leaving open which properties a recovered direction encodes.

A.1.2 Internal Truth Directions in LLMs

A growing body of work suggests that factual truth is linearly represented in the hidden states of large language models. Azaria & Mitchell (2023) showed that classifiers trained on internal activations can distinguish true from false statements, while Marks & Tegmark (2024) demonstrated that simple difference-of-means directions generalize across several factual datasets and can causally influence model outputs. Subsequent work has examined the structure and generality of these representations. Bürger et al. (2024) decomposed the truth subspace into a general truth component and a polarity-sensitive component, yielding probes that are robust to negation. Bao et al. (2025) further studied the consistency and transfer of truth directions across models, logical transformations, and question-answering settings, finding substantial variation in their generalizability. Extending this line of work to propositional composition, Li et al. (2026) showed that individual conjuncts' truth values can be decoded independently of the truth value of the full conjunction, and that interventions on conjunct representations can causally affect the model's truth-value judgment of the whole. Related studies have shown that truth representations are sensitive to additional aspects of model behavior and context: Schouten et al. (2025) found that truth-value judgments change with supporting or contradicting context, while Long et al. (2025) showed that deceptive instructions can systematically reshape truth-related representations. Orgad et al. (2025) extended internal-state probing to naturally generated hallucinations and found that truthfulness information is often localized near answer tokens but transfers imperfectly across tasks.

These findings have also raised concerns about what a probed "truth direction" actually represents. Unsupervised approaches such as Contrast-Consistent Search (Burns et al., 2023) can recover salient features that need not correspond uniquely to truth, a limitation analyzed further by Farquhar et al. (2023). More recently, Bertolazzi et al. (2026) showed that logical validity and semantic plausibility can be strongly entangled in representation space for LLM reasoning, demonstrating that internal features related to reasoning logics may encode correlated semantic factors. Our factual-probing analogue provides tighter control by holding the subject and relation fixed and varying only the asserted value, contrasting model-mined plausible falsehoods with random falsehoods. Levinstein

& Herrmann (2024) similarly argue that standard probing datasets may be insufficient to distinguish truth, belief, and other correlated properties. We identify factual plausibility as a contaminating property and quantify its overlap with the truth direction and its effect on probe errors.

For evaluation, prior truth-direction work validates through transfer and causal interventions (Marks & Tegmark, 2024) but reports no matched null, whereas we provide one using random balanced partitions of the same false statements (Section 3.3).

A.2 Additional Details of Plausible-Answer Construction

Candidate answer pools. For each structured factual item, we choose a plausible false answer from values already in the benchmark. Let p_i be the prefix containing the subject and relation, a_i^T the labeled correct answer, and $\mathcal{C}_i$ the candidate answer pool. For **Cities**, the pool contains all 108 country labels in the dataset. For **CounterFact**, candidates are objects listed as true answers for the same relation elsewhere in the dataset. We consider at most 40 candidates per item. For **sp_en_trans**, we consider at most 60 English glosses from the dataset for each Spanish word. The published true and false rows are not paired by Spanish word, so we use a fixed random seed to assign an incorrect gloss to each true word as its standard false answer. For **Companies**, candidates come from values used with the same template. The three templates are `X has headquarters in __`, `X operates in the industry of __`, and `X engages in __`. We evaluate at most 40 candidates per item. These pools preserve the answer type and, where applicable, the factual relation.

For each candidate $a \in \mathcal{C}_i$, we sum the log-probabilities of its continuation tokens under the mining model,

$$\ell_M(a \mid p_i) = \sum_{t=1}^{m_i(a)} \log P_M(c_{i,t} \mid p_i, c_{i,<t}),$$

where the continuation includes the full answer and closing punctuation. Each token is scored given the prefix and all preceding continuation tokens. We do not normalize by length, so the score measures the probability of the complete continuation. We select the highest-scoring candidate other than the labeled answer,

$$a_i^P = \arg \max_{a \in \mathcal{C}_i \setminus \{a_i^T\}} \ell_M(a \mid p_i).$$

Here, "most plausible" means most probable among the evaluated candidates.

Filtering and direction comparisons. We use a fixed 50/50 train/test split by factual item and keep all variants of an item in the same split. For direction comparisons, we first remove items whose mined plausible false answer matches the standard false answer, because they provide no plausible-versus-standard contrast. We also flag items when the mining model ranks an incorrect candidate at least as highly as the labeled answer, using the dataset-specific comparisons below. We call these items *ranking-ambiguous*. The flag records disagreement with the benchmark label and does not mean that the fact itself is ambiguous. For **Cities**, we flag an item when

$$\max_{a \in \mathcal{C}_i \setminus \{a_i^T\}} \ell_M(a \mid p_i) \geq \ell_M(a_i^T \mid p_i),$$

so ties are included. For **Companies**, we use the strict condition

$$\max_{a \in \mathcal{C}_i \setminus \{a_i^T\}} \ell_M(a \mid p_i) > \ell_M(a_i^T \mid p_i).$$

The auxiliary **smaller_than** construction also uses $>$. Ranking-ambiguous items remain in the released data with their flags, but we exclude them from the primary held-out direction metrics. This keeps items that the mining model ranks incorrectly out of that evaluation.

We found no exact ties in 1,496 **Cities** rows, 348 **sp_en_trans** rows, 578 **Companies** rows, or 861 **smaller_than** rows. The smallest nonzero margin $|\ell_M(a_i^{\mathrm{P}} \mid p_i) - \ell_M(a_i^{\mathrm{T}} \mid p_i)|$ is 3×10^{-3}. The only tie is a **CounterFact** item about the official language of Equatorial Guinea. The benchmark lists the same value as both the true and false target for this item. The preceding filter removes it, so exact ties do not affect the reported held-out direction metrics.

Semantic equivalence and label ambiguity. A candidate differs from the benchmark's labeled answer when $a \neq a_i^T$. This alone does not make it factually incorrect. Synonyms, aliases, abbreviations, spelling variants, or alternative translations can express the same answer. Some relations also have more than one valid answer. The ranking-ambiguity flag only checks the model's scores, so we separately audited each mined set for these issues.

Surface variants. Cities uses 108 country labels with one spelling per country. No mined false answer is an alias of the labeled answer. The screen flagged some similar names, including Niger/Nigeria and the Democratic Republic of the Congo/the Republic of the Congo, but these pairs refer to distinct countries. In **sp_en_trans**, no mined false answer has string overlap with its labeled gloss. In **Companies**, the source file uses two capitalizations for 28 industry values, such as "banking" and "Banking". As a result, 65 of 578 mined false answers (11.2%) were the labeled answer with different capitalization. We now normalize capitalization and punctuation when comparing candidates with the labeled answer and exclude matches. No such item remains after this change.

Multiple valid answers. Each **Cities** city has one labeled country, and each **sp_en_trans** word has one attested gloss. No mined translation answer matches another attested gloss for that word. Each **Companies** company–slot pair also has one labeled value. In **CounterFact**, 12 of the 150 relations allow multiple valid answers and cover 9.8% of items. For example, 38.8% of subjects in the relation "{} is a twin city of" have several attested true targets. We exclude these relations. Across seventeen models, their removal changes mean held-out AUC from 0.963 to 0.960 on true versus standard false statements and from 0.863 to 0.859 on true versus plausible false statements. Mean fixed-threshold FPR on plausible negatives changes from 0.374 to 0.373. Each of these metrics changes by less than 0.01 for every individual model.

A.3 Permutation Null Test

A direction obtained by contrasting two groups of false statements can appear predictive because of structure shared by those statements, even if the grouping does not identify plausibility. We therefore compare the learned plausibility direction against directions fitted from random partitions of the *same* training false statements. This gives a control matched to the data and fitting procedure, rather than an arbitrary random vector in representation space.

Let $\mathcal{F}_{\text{train}}$ contain the n plausible-false and n standard-false training representations. For each permutation b, we randomly partition these $2n$ representations into equal-sized groups A_b and B_b and recompute the unit mass-mean direction:

$$d^{(b)} = \frac{\mu_{A_b} - \mu_{B_b}}{\|\mu_{A_b} - \mu_{B_b}\|_2}. \tag{9}$$

The factual train/test split, representations, and held-out plausible-versus-standard-false labels remain fixed. Each null direction is evaluated on precisely the same test statements as d_{plaus}. Following the direction analyses, we use orientation-invariant AUC,

$$T(d) = \max\{\text{AUC}(d),\, 1 - \text{AUC}(d)\}\,, \tag{10}$$

where plausible falsehoods are the positive class and standard falsehoods the negative class. Applying the same statistic to observed and permuted directions accounts for the arbitrary sign of a random partition.

We repeat the procedure 300 times and report the null mean, standard deviation, and 97.5th percentile, comparing the observed held-out AUC with this empirical reference. For analyses of the component orthogonal to the truth direction, we apply the same projection and renormalization to every null direction while keeping the training-fitted d_{truth} fixed. This prevents the projection itself from giving the observed direction an advantage over its controls. The test asks whether the mined plausible-versus-standard grouping carries more held-out discriminative structure than random balanced groupings of the same false examples; it does not by itself establish a causal mechanism or a domain-independent plausibility representation.

A.4 Transferability across Models and Datasets

Cross-model transfer of plausible falsehoods. Table 4 tests transfer of plausible falsehoods mined by Llama-3-8B to other models, each using its own representations and fitted directions.

Table 4: **Truth and plausibility directions on held-out statements.** T/R denotes true versus standard false; P/R denotes plausible versus standard false. Columns report cosine similarity and orientation-invariant AUC for d_{truth} (d_T) and d_{plaus} (d_P). For example, d_P: T/R means applying the probed d_P direction on the test data of true versus standard false. Null is the 97.5th percentile of 300 random-partition AUCs on P/R. Negatives are mined by Llama-3-8B and shared across evaluated models.

Model	Cosine	d_T: T/R	d_P: T/R	d_T: P/R	d_P: P/R	Null
CounterFact						
Llama-3-8B	0.805	0.952	0.943	0.671	0.737	0.559
Llama-2-13B	0.793	0.927	0.913	0.690	0.754	0.551
Mistral-7B	0.800	0.952	0.944	0.717	0.781	0.574
OLMo-2-7B	0.828	0.913	0.901	0.684	0.726	0.555
Qwen2.5-14B	0.837	0.932	0.909	0.689	0.728	0.560
Gemma-2-9B	0.779	0.923	0.882	0.660	0.726	0.548
Cities						
Llama-3-8B	0.479	1.000	0.963	0.619	0.850	0.759
Llama-2-13B	0.779	1.000	0.999	0.710	0.777	0.755
Mistral-7B	0.748	1.000	1.000	0.762	0.842	0.808
OLMo-2-7B	0.803	1.000	0.999	0.782	0.829	0.820
Qwen2.5-14B	0.681	1.000	0.991	0.718	0.787	0.769
Gemma-2-9B	0.230	0.998	0.881	0.534	0.819	0.689
sp_en_trans						
Llama-3-8B	0.819	1.000	0.999	0.852	0.871	0.847
Llama-2-13B	0.861	1.000	0.994	0.871	0.875	0.839
Mistral-7B	0.822	1.000	1.000	0.893	0.908	0.846
OLMo-2-7B	0.864	1.000	0.999	0.872	0.894	0.836
Qwen2.5-14B	0.788	1.000	1.000	0.810	0.838	0.822
Gemma-2-9B	0.752	1.000	0.999	0.762	0.824	0.809

Table 5: **Cross-dataset transfer of the plausibility direction.** Entries are mean AUC across six core models, with the across-model range in parentheses. The direction is trained on the source domain and tested on the target domain.

Source → target	d_{plaus}: T/R	d_{plaus}: P/R
Cities → CounterFact	0.773 (0.624–0.883)	0.638 (0.570–0.695)
Cities → sp_en_trans	0.983 (0.923–1.000)	0.813 (0.757–0.866)
CounterFact → Cities	1.000 (0.998–1.000)	0.749 (0.584–0.831)
CounterFact → sp_en_trans	1.000 (1.000–1.000)	0.852 (0.830–0.874)
sp_en_trans → Cities	0.999 (0.997–1.000)	0.704 (0.601–0.774)
sp_en_trans → CounterFact	0.776 (0.751–0.813)	0.645 (0.629–0.669)

Cross-dataset transfer of the plausibility direction. Table 5 reports cross-dataset transfer for the plausibility direction. For each source–target pair, d_{plaus} is fitted on the source-domain training split and evaluated on the target-domain test split. Each entry is the mean AUC across the six core evaluated models, with the range across models in parentheses. The plausible negatives are the shared Llama-3-8B negatives used throughout the corresponding audit; T/R denotes true versus standard false and P/R denotes plausible versus standard false.

Table 6: **Direct projection of the mass-mean truth direction on CounterFact.** AUC is measured on held-out true versus standard-false statements with shared Llama-3-8B negatives. Loss is the unprojected AUC minus the projected AUC.

Model	Original	Projected	Loss
Llama-3-8B	0.900	0.815	0.085
Llama-2-13B	0.817	0.726	0.091
Mistral-7B	0.850	0.784	0.066
OLMo-2-7B	0.846	0.783	0.063
Qwen2.5-14B	0.753	0.684	0.069
Gemma-2-9B	0.857	0.819	0.038

Table 7: **Refitted logistic truth readers in the orthogonal feature subspace.** Results use shared Llama-3-8B negatives on CounterFact. The reader is fit either in the full representation or after projecting every representation orthogonally to d_{plaus}.

Model	Full	Orthogonal	Loss
Llama-3-8B	0.989	0.974	0.015
Llama-2-13B	0.970	0.952	0.018
Mistral-7B	0.981	0.965	0.016
OLMo-2-7B	0.948	0.916	0.032
Qwen2.5-14B	0.973	0.935	0.038
Gemma-2-9B	0.981	0.964	0.017

A.5 Separating Truth and Plausibility

Tables 6 and 7 give the CounterFact results for the two linear separation strategies described in Section 4.1. We report the standard held-out T/R AUC before and after removing the plausibility component, together with the AUC loss. All results use the shared Llama-3-8B negatives; "orthogonal" denotes refitting after projecting every representation away from d_{plaus}.

A.6 Within-False Correlations between Likelihood and Truth Scores

Table 8 reports Pearson correlations between length-normalized statement log-probability and the logistic truth reader's predicted probability of truth. The reader is fitted on training statements with both truth labels, whereas each correlation is computed only over false statements in the held-out split. Thus, the association reflects variation within a fixed truth label. All eight correlations are positive. The Llama-3.1-70B run uses 4-bit quantized weights.

A.7 Ranking Performance and Fixed-Threshold Reliability

Table 9 gives the model-level results on **Cities** and **sp_en_trans** summarized in Section 4.2. The mass-mean truth direction is evaluated on held-out true versus standard false (T/R) and true versus plausible false (T/P) statements. The separate logistic reader uses the unchanged probability threshold of 0.5 for TPR and both FPRs. Plausible negatives are mined by Llama-3-8B and shared across evaluated models. AUC and classification columns describe different readers and should not be interpreted as measurements of the same classifier.

Table 10 supplements Table 2 with standard-distribution calibration and accuracy for the logistic reader on **CounterFact**. ECE and accuracy are evaluated on true and standard false statements; the FPR ratio compares plausible and standard false statements at the same fixed threshold. Thus, calibration on the standard distribution does not establish reliability under plausible-negative replacement.

A.8 Likelihood-Stratified Errors on Claims and Generated Spans

Free-form claims. Table 11 details the free-form evaluation in Section 4.2. A logistic truth reader is trained on the original true/false labels and evaluated at probability threshold 0.5. Among held-out false statements, the lowest and highest likelihood thirds are defined using the evaluated model's

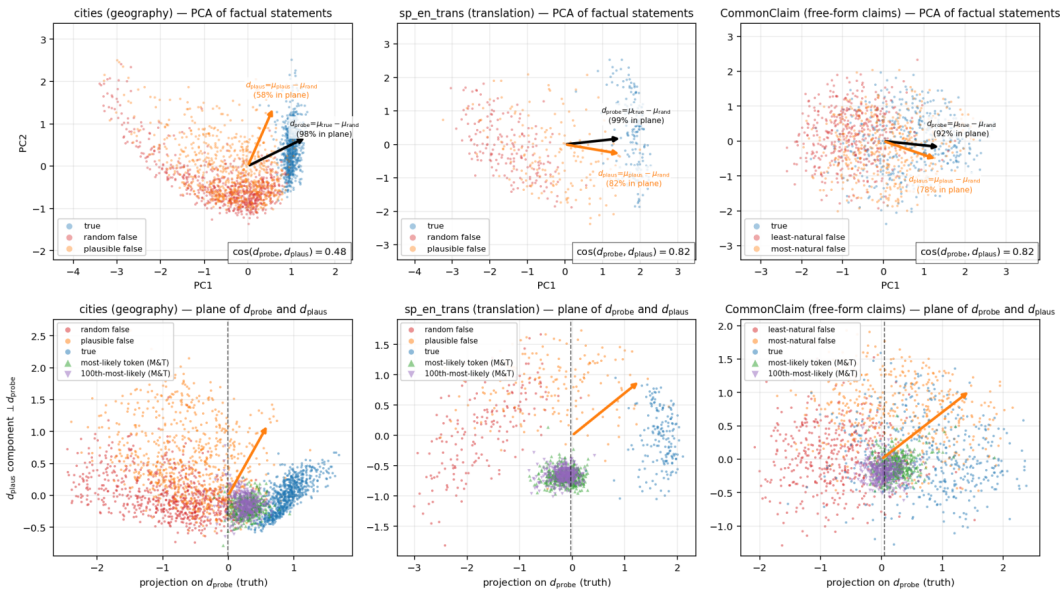


Figure 1: **Factual plausibility and the generic likelihood control.** Columns show **Cities**, **sp_en_trans**, and **CommonClaim**. Top: PCA projections of factual statements, with truth and plausibility directions overlaid. Bottom: projections onto the plane spanned by these directions, including nonfactual continuations ending in the most likely or 100th-most-likely token. For **CommonClaim**, the false-statement groups are labeled least-natural and most-natural in the figure.

Table 8: **Likelihood–truth-score correlations within false statements.** Pearson r between length-normalized statement log-probability and the logistic reader's predicted probability of truth, evaluated on held-out false statements.

Model	**Companies**	**CommonClaim**
Llama-3-8B	0.273	0.420
Llama-3.1-8B	0.325	0.370
Llama-3.1-8B-Instruct	0.260	0.310
Llama-3.1-70B	0.305	0.340

length-normalized statement log-probability. No answer replacements are constructed. The original test split contains 500 statements per dataset; the reported group sizes count false statements in the highest-likelihood third. The Llama-3.1-70B run uses 4-bit quantized weights.

Generated unsupported spans. Table 12 reports all four configurations of the **hallucination-probes** evaluation summarized in Section 4.2. Llama-3.1-8B-Instruct generates long-form responses whose spans are labeled Supported or Not Supported. Logistic readers are fitted using 466 training responses and evaluated on 6,025 labeled spans from 380 held-out responses. Each reader uses either the last-token representation or the mean over span tokens at layer 13 or 30. At threshold 0.5, recall is measured on supported spans and FPR is the fraction of unsupported spans classified as supported. The low/high thirds are formed within unsupported test spans using mean per-token log-probability conditioned on the preceding context. All four configurations show more false positives in the highest-likelihood third.

Table 9: **Ranking and fixed-threshold results on the additional structured datasets.** AUC columns use mass-mean readers; TPR and FPR columns use logistic readers. Values are shown to four decimal places to distinguish near-perfect AUC from 1.

Model	AUC T/R	AUC T/P	TPR	FPR R	FPR P
Cities					
Llama-3-8B	0.9999	0.9935	1.0000	0.0000	0.0669
Llama-2-13B	0.9995	0.9974	1.0000	0.0028	0.0891
Mistral-7B	1.0000	0.9995	1.0000	0.0000	0.0641
OLMo-2-7B	1.0000	0.9918	0.9972	0.0084	0.1226
Qwen2.5-14B	1.0000	0.9898	1.0000	0.0000	0.0446
Gemma-2-9B	0.9982	0.9904	1.0000	0.0000	0.0251
sp_en_trans					
Llama-3-8B	1.0000	0.9939	1.0000	0.0000	0.2442
Llama-2-13B	0.9996	0.9754	1.0000	0.0116	0.3023
Mistral-7B	1.0000	0.9972	1.0000	0.0000	0.1977
OLMo-2-7B	1.0000	0.9774	1.0000	0.0116	0.4186
Qwen2.5-14B	1.0000	0.9943	1.0000	0.0000	0.1279
Gemma-2-9B	1.0000	0.9978	1.0000	0.0000	0.1977

Table 10: **Standard calibration and plausible-negative FPR amplification on CounterFact.** The ratio is computed from unrounded FPRs.

Model	Standard accuracy	Standard ECE	FPR P/FPR R
Llama-3-8B	0.959	0.020	7.625
Llama-2-13B	0.904	0.052	2.643
Mistral-7B	0.934	0.041	4.758
OLMo-2-7B	0.877	0.068	3.184
Qwen2.5-14B	0.920	0.041	3.465
Gemma-2-9B	0.931	0.032	4.267

Table 11: **False-positive rates by likelihood on free-form claims.** Low and high denote the least- and most-likely thirds of held-out false statements. AUC uses all held-out true/false statements.

Model	AUC	n high	FPR low	FPR high
CommonClaim				
Llama-3-8B	0.828	84	0.083	0.452
Llama-3.1-8B	0.830	84	0.071	0.405
Llama-3.1-8B-Instruct	0.862	84	0.095	0.333
Llama-3.1-70B	0.771	84	0.190	0.429
Companies				
Llama-3-8B	0.983	86	0.012	0.070
Llama-3.1-8B	0.982	86	0.000	0.081
Llama-3.1-8B-Instruct	0.983	86	0.000	0.070
Llama-3.1-70B	0.952	86	0.035	0.221

Table 12: **Generated-span results on hallucination-probes.** Low and high FPR refer to the least- and most-likely thirds of unsupported spans.

Layer	Pooling	AUC	Recall	FPR all	FPR low	FPR high
13	Last token	0.748	0.80	0.47	0.42	0.53
13	Mean span	0.781	0.81	0.41	0.35	0.49
30	Last token	0.738	0.79	0.48	0.40	0.57
30	Mean span	0.766	0.82	0.45	0.37	0.57

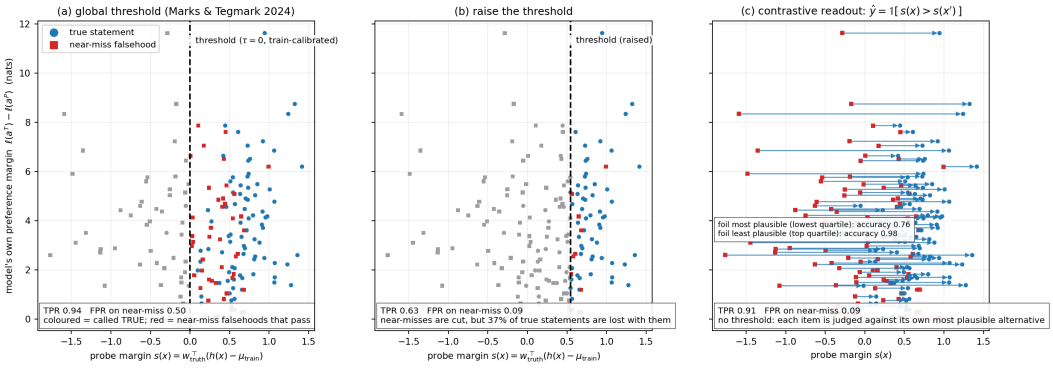


Figure 2: **Global threshold adjustment and local contrastive readout on real data.** The horizontal axis is the centred truth-probe margin; larger values indicate higher truth scores. The vertical axis is the mining model's conditional log-probability margin $\ell_M(a_i^{\mathrm{T}} \mid p_i) - \ell_M(a_i^{\mathrm{P}} \mid p_i)$ in nats: smaller margins mean the selected false answer is nearly as probable as the true answer. This axis exposes variation in relative plausibility and is used only for visualization, not classification. Circles denote true statements and squares plausible falsehoods. (a) The training-calibrated global threshold yields TPR 0.94 and plausible-negative FPR 0.50. (b) Raising the threshold lowers FPR to 0.09 but reduces TPR to 0.63; grey points in (a,b) are classified as false. (c) Local readout compares $s(x)$ with $s(x')$ for an alternative from the same factual context, yielding TPR 0.91 and FPR 0.09. Horizontal arrows connect the paired false and true statements. Rates describe this demonstration, rather than the cross-model means in Table 3.

A.9 Illustration of Global and Local Readouts

Figure 2 illustrates the limitation of applying a global threshold and our idea of local contrastive readout with real data. Plausible false positives concentrate where the mining model's log-probability margin between the true and plausible false answers is small. Raising the global threshold reduces plausible-negative FPR from 0.50 to 0.09, but also lowers TPR from 0.94 to 0.63: moving the threshold does not change the ordering of statements on the global score axis. In contrast, comparing each statement with its own alternative achieves TPR 0.91 at the same plausible-negative FPR of 0.09 in this demonstration. This illustrates how a local comparison can mitigate the interference from plausibility without rejecting as many true statements.