# Truth-value judgment in language models: 'truth directions' are context sensitive

**Stefan F. Schouten, Peter Bloem, Ilia Markov, Piek Vossen**
Vrije Universiteit Amsterdam
{s.f.schouten,p.bloem,i.markov,p.t.j.m.vossen}@vu.nl

## Abstract

Recent work has demonstrated that the latent spaces of large language models (LLMs) contain directions predictive of the truth of sentences. Multiple methods recover such directions and build probes that are described as uncovering a model's "knowledge" or "beliefs". We investigate this phenomenon, looking closely at the impact of *context* on the probes. Our experiments establish where in the LLM the probe's predictions are (most) sensitive to the presence of related sentences, and how to best characterize this kind of sensitivity. We do so by measuring different types of consistency errors that occur after probing an LLM whose inputs consist of hypotheses preceded by (negated) supporting and contradicting sentences. We also perform a causal intervention experiment, investigating whether moving the representation of a premise along these *truth-value directions* influences the position of an entailed or contradicted sentence along that same direction. We find that the probes we test are generally context sensitive, but that contexts which should not affect the truth often still impact the probe outputs. Our experiments show that the type of errors depend on the layer, the model, and the kind of data. Finally, our results suggest that truth-value directions are causal mediators in the inference process that incorporates in-context information.

## 1 Introduction

As Large Language Models (LLMs) enjoy increasing mainstream adoption, it becomes more important to understand why they fail in some cases, while excelling in others. Recent findings show that LLM latent spaces contain directions predictive of the truth of sentences (Burns et al., 2023; Marks & Tegmark, 2024). Probes that leverage these directions to assign truth values to sentences are accurate even in misleading contexts where prompting fails. When considering simple declarative sentences on their own, it makes sense to evaluate such *truth-value probes* primarily by their accuracy. For example, we might have a hypothesis: *"In New York, days are shortest in December"*, which we would expect an LLM to represent as true. But how should we evaluate if this hypothesis is placed in the context of an incorrect premise, like: *"December is **not** during the winter for New York"*? In that case, we no longer necessarily have the same expectations. For example, if the premise is understood as setting up a counterfactual scenario, then we would expect an LLM to represent the hypothesis as false. In other words, for multi-sentence inputs, evaluations must prioritize *coherence*: the degree of logical consistency in truth-value assignments.

When a truth-value probe assigns a sentence $S$ the label TRUE, it is tempting to refer to the model as *believing* $S$. Especially because such explanations of LLM behavior are highly plausible (convincing to humans, Jacovi & Goldberg, 2020). Explaining LLMs by appealing to attitudes like belief has recently been endorsed as *propositional interpretability* (Chalmers, 2025), and truth-value probes seem like a promising method. Specifically, they could serve to gauge if an LLM—given one or more sentences as input—judges them to be true (actively believes them) or not. But, if a probe's truth-value assignments are not coherent, it has failed to reveal (representations of) judgment or belief (Herrmann & Levinstein, 2025), meaning an interpretation that appeals to those concepts would be unfaithful.

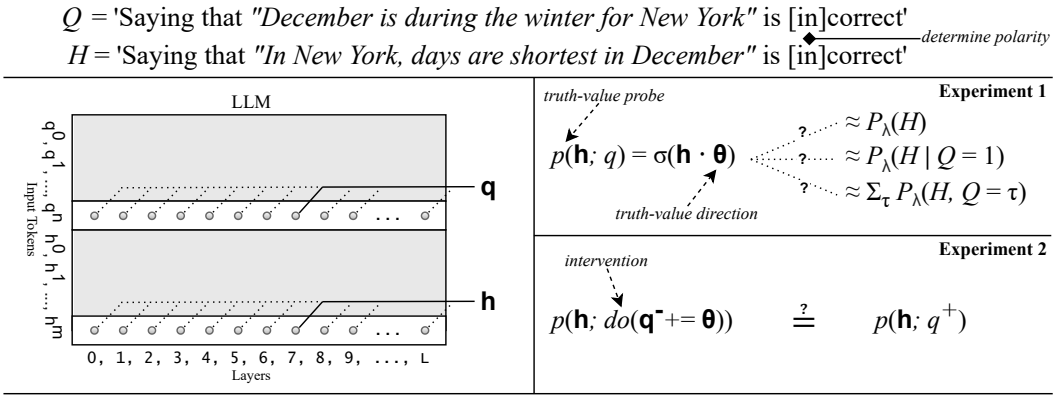



Figure 1: Overview of our setup. LLM representations **q** and **h** for a premise and hypothesis are extracted and used to train truth-value probes. In experiment 1, the probes are evaluated to determine if and how they incorporate context. In experiment 2, we move a premise's representation in the identified truth-value direction, measuring if the probability assigned to the hypothesis changes accordingly.

In this work, we study coherence by probing LLMs on inputs where hypotheses are preceded by premises which appear either affirmed or negated. We take inspiration from the *truth-value judgment* task in language acquisition research where subjects are "asked to make a bipolar judgment about whether a statement accurately describes a particular situation alluded to in some context or preamble" (Gordon, 1996). We find that truth-value assignments are context sensitive, but also sensitive to irrelevant information.

We also investigate if the relevant directions in latent-space causally mediate truth-value judgment, or if they only reflect the outcome of that process. Specifically, we establish if a representation's position along a truth-value direction determines (in part) where subsequent statements are positioned along the same direction. Our results suggest that these directions are mediators in the inference process that incorporates in-context information.

In summary, our contributions are: (1) experiments evaluating the context sensitivity of truth-value probes and the consistency with which they incorporate it; we quantify across layers, model sizes (7 and 13 billion), and type of training (pretrained-only vs. instruction-tuned); and (2) an experiment demonstrating that truth-value directions causally mediate natural language inference. We also propose a new variant of CCS (Burns et al., 2023) for which convergence is more stable, and otherwise behaves and performs similarly. Our code is available at `https://github.com/sfschouten/tvj-in-llms`.

## 2 Related Work

Probing LLM representations for the truth of sentences has recently received much interest. Burns et al. (2023) introduce Contrast Consistent Search (CCS), an unsupervised probing methods based on the representations of contrasting sentence pairs. Their probes often outperform a (zero-shot) prompting approach, even when applied to misleading prompts. Li et al. (2023) shift model activations in the 'truth direction' at inference time, mitigating hallucination. Their interventions use 1) directions from probes trained with logistic regression (LR) and CCS; and 2) a new method (Mass Mean Shift), which finds the direction as the difference between the means of the true and false sentence activations. Marks & Tegmark (2024) take Mass Mean Shift and use its directions to build probes (Mass-Mean Probing, MMP). They show all probes (based on LR, CCS, and MMP) generalize well between datasets, with MMP performing the best. In a causal intervention experiment, they move representations in the identified directions, showing that MMP is the best mediator: it increases the probability of the model calling a false statement true the most.

Previous work has used both data consisting of single facts and longer inputs, such as various NLI datasets, but did not study the impact of the context. We specifically study the

in-context behavior of truth-value probes, analyzing their consistency, and what this means for the way LLMs incorporate contextual information. Like Marks & Tegmark (2024), we also investigate the causal implication of directions in LLM latent space. However, rather than investigate what causes the greatest change in token predictions, we investigate which direction to move a premise in, such that it causes the correct change in the probability of a related hypothesis, as evaluated by the same direction.

Recent work has also criticized this type of probing. Truth-value probes might identify properties that correlate with truth (Levinstein & Herrmann, 2024), especially when truth is not the most salient feature (Farquhar et al., 2023). We evaluate the probes' coherence, which we do not expect spurious correlations to exhibit.

## 3 Methodology

We describe our method in three parts: in 3.1 we cover truth-value probes, the necessary assumptions, notation, and methods to construct them; in 3.2 we describe how to construct samples and the possible quantities measured by probes in our setting; and in 3.3 we introduce error scores with which we measure (different kinds of) coherence.

### 3.1 Truth-value probes

We use several probing methods in our experiments. These methods use datasets of sentences, consisting of both true and false statements. We can turn any true statement into a false statement (and vice versa) by negating it. We use superscript $+$ and $-$ to denote the affirmed ($X^+$) and negated ($X^-$) variants of a sentence, respectively. Their LLM vector representations are given in bold, i.e. $\mathbf{x}^+$, $\mathbf{x}^-$ (see section 4 for how we negate sentences and for how vector representations are extracted). Thus, the dataset used to train probes consist of pairs of hidden states extracted for the positive and negative variants of statements $(\mathbf{x}^+, \mathbf{x}^-, y^+, y^-) \in \mathcal{D}$, and their labels indicating which of the two is true (with $y^+ = 1 - y^-$).

When using truth-value probes, we assume that the truth of sentences is latently modeled by LLMs. We characterize this latent ($\lambda$) model as a probability distribution $P_\lambda(X)$. The probes $p(\mathbf{x})$ are assumed to (approximately) recover this distribution. We believe these assumptions are fair, because previous work (Burns et al., 2023; Marks & Tegmark, 2024) has found directions in latent spaces that suggest LLMs do track sentence truth.

Probes are constructed as: $p(\mathbf{x}) = \sigma(\mathbf{x} \cdot \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the truth-value direction and $\sigma$ is the sigmoid. Because the probes do not have bias terms, all inputs are mean-normalized (in line with previous work): $\mathbf{x}^+ \mathrel{-}= \boldsymbol{\mu}$ and $\mathbf{x}^- \mathrel{-}= \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \frac{1}{2|\mathcal{D}|} \sum_{(\mathbf{x}^+, \mathbf{x}^-) \in \mathcal{D}} (\mathbf{x}^+ + \mathbf{x}^-)$.

**Mass Mean Probing (MMP)** is a supervised method, which defines the truth-value direction as the difference between the average of the correct and incorrect statements:

$$\boldsymbol{\theta}_{\text{mm}} = \mathbb{E}_{\mathbf{x},y}[\mathbf{x} \mid y = 1] - \mathbb{E}_{\mathbf{x},y}[\mathbf{x} \mid y = 0], \tag{1}$$

where $y$ is the truth-value (label) for the statement $X$. We do not include the version of MMP that requires an i.i.d. assumption, because we also evaluate on out-of-distribution data.

**Logistic Regression (LR)** is also used to train a supervised probe. They are trained on $\mathbf{x}' = \mathbf{x}^- - \mathbf{x}^+$, i.e. the difference between the negative and positive statements.

$$\boldsymbol{\theta}_{\text{lr}} = \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{\mathbf{x}',y^+} \left[ y^+ \ln \sigma(\boldsymbol{\theta} \cdot \mathbf{x}') + (1 - y^+) \ln \left(1 - \sigma(\boldsymbol{\theta} \cdot \mathbf{x}')\right) \right], \tag{2}$$

where $y^+$ is the label for the positive variant of the sample, i.e. whether $X^+$ is true.

**Contrast Consistent Search (CCS)** is an unsupervised[1] method with as its objective:

$$\boldsymbol{\theta}_{\text{ccs}} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}^+,\mathbf{x}^-} \left[ [1 - p(\mathbf{x}^+) - p(\mathbf{x}^-)]^2 + \min\{p(\mathbf{x}^+), p(\mathbf{x}^-)\}^2 \right], \tag{3}$$

[1]By unsupervised we mean that no knowledge of which sentences are true or false is given.

which has two terms: the consistency-loss (encouraging solutions where the probabilities add up to one), and the confidence-loss (encouraging non-degenerate solutions, i.e. $p(\mathbf{x}^+) \neq p(\mathbf{x}^-) \neq 0.5$). The objective can be understood as finding a hyperplane with normal $\boldsymbol{\theta}$ that, for each pair: (1) separates $\mathbf{x}^+$ from $\mathbf{x}^-$, and (2) is equidistant to $\mathbf{x}^+$ and $\mathbf{x}^-$.

**Contrast Consistent Reflection (CCR)** is proposed here as a variant of CCS. Rather than finding a hyperplane from which $\mathbf{x}^+$ and $\mathbf{x}^-$ are equidistant, this method requires $\mathbf{x}^+$ and $\mathbf{x}^-$ to be each other's reflection in the hyperplane. It has the following objective:

$$\boldsymbol{\theta}_{\text{ccr}} = \arg\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\mathbf{x}^+,\mathbf{x}^-} \left[ ||\mathbf{x}^+ - \mathbf{P}\mathbf{x}^-||_2 \right], \tag{4}$$

where unit vector $\hat{\boldsymbol{\theta}}$ determines the direction of the Householder reflection $\mathbf{P} = \mathbf{I} - 2\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\intercal$. This objective does not share the degenerate solution of CCS. This is because for $p(\mathbf{x}^+) = p(\mathbf{x}^-) = 0.5$, we need $\boldsymbol{\theta} \cdot \mathbf{x}^+ = \boldsymbol{\theta} \cdot \mathbf{x}^- = 0$, and since $|\boldsymbol{\theta}| = 1$ this would imply that $\boldsymbol{\theta}$ is orthogonal to $\mathbf{x}^+$ and $\mathbf{x}^-$. Thus, while they are equidistant in that scenario (a distance of zero), assuming that $\mathbf{x}^+ \neq \mathbf{x}^-$, they will not be each other's reflection.

On our data CCS does not consistently converge to a good minimum (see Appendix D). CCS finds directions that vary considerably from layer to layer (see Appendix E) making it harder to analyze. We see CCR achieve similar performance with more stable convergence. By including CCR, we can see how unsupervised methods compare to supervised methods, without having to worry about observations being artifacts of CCS's instability.

### 3.2 Probing for truth-value judgment

To probe for truth-value judgment we have a setup as displayed in Figure 1, for example:

| | | |
|---|---|---|
| $Q$ | (premise) | December is [not] during the winter for New York. |
| $H$ | (hypothesis) | In New York, days are [not] shortest in December. |

The bracketed parts in $Q$ and $H$ are omitted or included to produce affirmed $(Q^+, H^+)$ and negated $(Q^-, H^-)$ sentences. We use truth-value probes to see if the model represents $H^+/H^-$ as true or false, and how this changes when preceded by $Q^+$ or $Q^-$.

There are different ways in which a belief could interact with the context of a statement. We define three kinds of beliefs in the following way:

- *prior beliefs*, independent of the context, given by $P_\lambda(H)$;
- *conditional beliefs*, specially where the context is assumed to be truthful, given by $P_\lambda(H|q)$;
- *marginal beliefs*, where the truth of the premise and hypothesis are modeled jointly, with the effect of the premise summed out, given by $\sum_\tau P_\lambda(H, Q{=}\tau)$.[2]

Each of these types of beliefs are candidates for what is measured by $p(\mathbf{h}; q)$, a probe applied to LLM activations for a hypothesis $H$ when preceded by a premise $Q$. Note that there could also be beliefs in between conditional and marginal, where the context's truthfulness is not totally assumed, but still biased (e.g. towards being true rather than false).

### 3.3 Evaluation

To evaluate if probes coherently incorporate in-context information, we include four error scores, each indicating to what degree probe outputs violate desired behavior.

We first define the *premise effect* (*PE*) as the difference in probability assigned to the hypothesis when preceded with an affirmed premise and probability assigned to the hypothesis on its own: $PE = p(\mathbf{h}; q^+) - p(\mathbf{h})$. We call a method's mean absolute premise effect its *premise sensitivity*. A value close to zero for this metric would be consistent with a prior belief.

The effect of adding the in-context premise can differ in magnitude depending on which probing method we use. In order to make the error scores of different methods comparable,

[2]We leave this expression unsimplified to emphasize the dependence on the joint distribution, which is what distinguishes marginal from prior beliefs.

we express the errors in multiples of the premise effect $PE$. This makes the error scores independent of the overall premise sensitivity of the probing method.

The first two error scores, E1 and E2 (see Table 1) are based on the fact that we expect the probabilities to depend only on factors that are actually capable of influencing the truth value of the hypothesis. Thus, these error scores are proportional to the absolute change in probability that occurs after having the hypothesis preceded by either: 1) a corrupted premise $\tilde{Q}$, or 2) an unrelated premise $Q'$. The truth value of both corrupted and unrelated premises are independent of the truth value of the hypothesis, which is why we want the equalities for E1 and E2 in Table 1 to hold.

E3 and E4 measure when probes fail to behave like *conditional* and *marginal* beliefs, respectively. Consider the example: Q="*December is during the winter for New York*" and H="*In New York, days are shortest in December*".

If the model assumes the premise is true (a conditional belief, $p(\mathbf{h}; q) \approx P_\lambda(H|q)$) when determining the probability for H, then either: (1) having the context say 'Q is incorrect' should decrease the probability of H, or (2) having the context say 'Q is correct' should increase the probability. This expectation is captured by E3. For the error score, we have: $(p(\mathbf{h}; q^-) - p(\mathbf{h})) \cdot PE^{-1} = (p(\mathbf{h}; q^-) - p(\mathbf{h}))/(p(\mathbf{h}; q^+) - p(\mathbf{h}))$. We want the effect of a negated premise to be opposite of a positive premise: when the numerator is positive, we want the denominator to be negative and vice versa. Taking $max\{\cdot, 0\}$ of this fraction, we can isolate the cases where the numerator and the denominator have the same sign, which are the errors we want to capture in the score.

If the model instead bases itself on its own evaluation of the truth of Q (a marginal belief, i.e. it models Q and H jointly: $p(\mathbf{h}; q) \approx \sum_\tau P_\lambda(h, Q{=}\tau)$), then having 'Q is incorrect' or 'Q is correct' should not influence the probability of H at all. This expectation is captured by E4. In that case, the probability assigned to the hypothesis should be the same regardless of whether the premise is asserted or denied.

Because low scores for E3 and E4 indicate two equally valid types of truth-value judgment, we do not expect the score to be low for both. See Appendix A for additional details.

Table 1: Expected behavior and corresponding error scores. The subscript $e$ and $c$ indicate hypotheses entailed or contradicted by their premise.

| | (in)equality | error score |
|---|---|---|
| E1 | $P_\lambda(h\|\,\tilde{Q}\,) = P_\lambda(h)$ | $\lvert p(\mathbf{h}; \tilde{q}\ \ ) - p(\mathbf{h})\rvert \cdot \lvert PE^{-1}\rvert$ |
| E2 | $P_\lambda(h\|\,Q') = P_\lambda(h)$ | $\lvert p(\mathbf{h}; q'\ ) - p(\mathbf{h})\rvert \cdot \lvert PE^{-1}\rvert$ |
| E3 | $P_\lambda(h_e\|q^-) \leq P_\lambda(h) \leq P_\lambda(h_e\|q^+)$<br>$P_\lambda(h_c\|q^-) \geq P_\lambda(h) \geq P_\lambda(h_c\|q^+)$ | $\max\{(p(\mathbf{h}; q^-) - p(\mathbf{h})) \cdot PE^{-1},\ 0\}$ |
| E4 | $\sum_\tau P_\lambda(h, Q^-{=}\tau) = \sum_\tau P_\lambda(h, Q^+{=}\tau)$ | $\lvert p(\mathbf{h}; q^-) - p(\mathbf{h}; q^+)\rvert \cdot \lvert PE^{-1}\rvert$ |

## 4 Experiments

In our experiments, we make use of datasets with samples of related sentences whose truth values depend on each other. We use samples from these datasets by creating prompts where the sentences are either affirmed or negated.

We train probes in a `no-prem` and `pos-prem` setting. For `no-prem`, the premise $Q$ is left out, and for `pos-prem` the premise appears in the positive (or affirmed) variant. We include these settings to better understand how truth-values are represented. A direction found in the `no-prem` setting we might expect to represent prior belief. If that direction shows context-sensitivity (when evaluated with premises in-context), that is evidence that the model does not represent the prior and contextual beliefs independently (in orthogonal directions). For `pos-prem`, the direction found is also influenced by what appears in context. If the directions found for `pos-prem` and `no-prem` are different, it suggests there *is* a separate (but possibly related) direction used to represent contextual belief.

The probe inputs $\mathbf{h}$ are the representations of the period following the answer tokens ('correct' / 'incorrect') extracted for each layer. To compare across probing methods we calibrate the probes such that their predictions for the $p(\mathbf{h})$ case have the same variance. We train probes on the following LLMs: Llama2-7b, Llama2-13b (Touvron et al., 2023), and OLMo-7b with and without instruction tuning (Groeneveld et al., 2024).

To measure the premise effect, and error scores described in subsection 3.3, we include the following evaluation cases: $p(\mathbf{h})$ (no premise), $p(\mathbf{h}; q^+)$ (affirmed premise), $p(\mathbf{h}; q^-)$ (negated premise), $p(\mathbf{h}; q')$ (unrelated premise), and $p(\mathbf{h}; \tilde{q})$ (corrupted premise). We evaluate both the `no-prem` and `pos-prem` in all of these cases. The first two cases are 'in distribution' for the `no-prem` and `pos-prem` settings, respectively. The other combinations are out of distribution. When evaluating the probes we use: $p(\mathbf{h}) = \frac{1}{2}(1 - p(\mathbf{h}^-) + p(\mathbf{h}^+))$.

**Data**

We use two existing datasets in our experiments. The first dataset (EntailmentBank, Dalvi et al., 2021) contains hypotheses that are sentences with general world knowledge. These are facts the LLM may have encountered during training and for which it could already have a strong prior belief. The second (SNLI, Bowman et al., 2015) contains statements that describe images, to which an LLM has no access. For both datasets, the corrupted sentences are created by replacing the characters in each word of the base sentence with random characters. The polarity of the premises and hypotheses are determined by switching between sentences that say something is 'correct' and saying that it is 'incorrect'. This style of negation avoids some problems that might otherwise arise.[3]

**EntailmentBank** This dataset contains statements with entailment relationships. The dataset was derived from ARC (Clark et al., 2018), which consists of grade-school level science questions. We combine premises from EntailmentBank with the questions and answers from ARC. The questions are answered correctly or incorrectly to create both entailments and contradictions. For example:

> You are given the following question:
> > In New York, the shortest period of daylight occurs during? (A) December (B) June
>
> $Q_a$ The statement "New York is located in the northern hemisphere." is [in]correct.
> $Q_b$ The statement "December is during the winter for New York." is [in]correct.
> $H$ Answering the question with "(B) June" is [in]correct.

The answer "June" is incorrect, and thus $H$ contradicts the information in $Q_a, Q_b$ (when it is not negated), while in the sample with the correct answer $H$ would be entailed by $Q_a, Q_b$. The dataset contains trees of entailing sentences, but we disregard anything but the first level of supporting premises. For the $p(\mathbf{h}; q)$ case, we use the distractor premises provided in the dataset. These were ranked as potentially relevant, but during annotation were not selected to be part of the entailment tree (Dalvi et al., 2021).

**SNLI** This dataset is a Natural Language Inference dataset, it consists of premise-hypothesis pairs, which are labeled as: entailment, contradiction, or neutral; describing the meaning relation between the sentences. This dataset was created based on the descriptions of images. To avoid ambiguity, we establish a context as follows:

> You are looking at a picture (A) which is placed next to an unrelated picture (B).
> $Q$ Describing picture {A/B} as: "Four children are playing in some water." is [in]correct.
> $H$ Saying (about picture A) that: "The children are wet." is [in]correct.

The neutral sentences for the $p(\mathbf{h}; q')$ case are obtained by taking the premise from a different, randomly sampled premise-hypothesis pair. Furthermore, for this case, the 'A/B' that appears in curly brackets is set to B to ensure that there is a fully neutral relationship. Without it, the fact that the two sentences are about the same picture could make their (simultaneous) truth less likely. It is also set to B for $p(\mathbf{h}; \tilde{q})$, and set to A for all other cases.

[3]For example, negating "four children are playing in some water" as "four children are not playing in some water", still presupposes the existence of four children. Using a negative meta statement leaves open the possibility that the presupposition is false (e.g. the number of children is inaccurate).

Without access to the picture, the model's prior belief should result in 50% accuracy. However, for SNLI it is possible to predict the label solely from the hypothesis (Poliak et al., 2018). This makes for an interesting scenario when it comes to truth-value probing. A probing method might identify a direction that only encapsulates a statistical pattern, rather than a model's truth-value direction. It is also possible that the statistical pattern is absorbed into the model's truth-value direction, as simply another reason to believe the sentence. After the addition of a premise, we do not expect a representation should move (coherently) in a direction which merely encodes a statistical pattern. Thus, if a probe trained only on hypotheses *does* respond coherently to the presence of a premise at test time, it is further evidence of the probe uncovering truth-value judgment, and not just a statistical pattern.

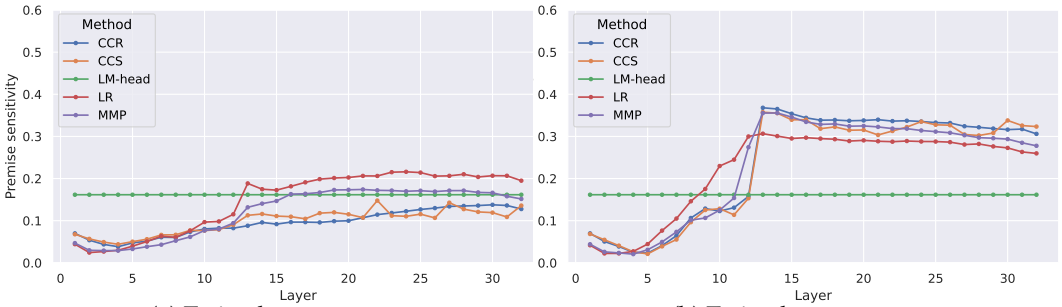
### 4.1 The effects of altering premises

We evaluate the probes on held-out data, including data from all the other variants. We also include an additional baseline, based on the model's LM-head, where the probabilities assigned to the 'correct'/'incorrect' tokens are rescaled to sum up to one.

Table 2 gives an overview of the average probabilities for $p(\mathbf{h}; q^+)$, $p(\mathbf{h}; q^-)$, and $p(\mathbf{h})$, split by whether the premise-hypothesis pair had an entailment or contradiction relation. We observe that the probabilities assigned to hypotheses depend strongly on the presence of relevant premises. When the hypothesis is entailed the probabilities are higher, and when the hypothesis is contradicted they are lower. This is true, even for probes trained without the premises present (`no-prem`), which also achieve good accuracy for the $p(\mathbf{h}; q^+)$ case. Although the premise sensitivity is lower, it is clear that the directions identified in the `no-prem` setting are not encoding specifically *prior* beliefs.

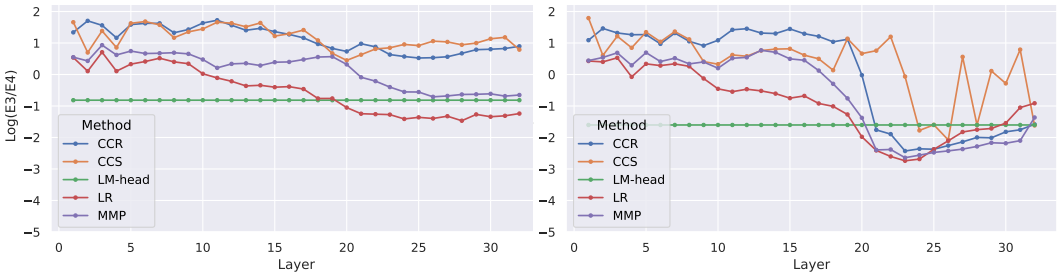
Table 2: Accuracy of $p(\mathbf{h}; q^+)$ (Acc), mean probabilities (orange=0, gray=0.5, blue=1), and trimmed mean errors scores for probes of each method on both datasets for Llama2-7b. The probes are from layers (L) with: (1) the best accuracy; and (2) the overall lowest error scores (by average error rank $E*$). The best scores per dataset are in bold, for E3 and E4 the bold values are based on their sum. CCS omitted, full table in Appendix B.

| | | | | | | Entailment | | | Contradiction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Method | L | Acc | $E*$ | $p(\mathbf{h}; q^+)$ | $p(\mathbf{h}; q^-)$ | $p(\mathbf{h})$ | $p(\mathbf{h}; q^-)$ | $p(\mathbf{h}; q^+)$ | E1 | E2 | E3 | E4 |
| EntailmentBank | | LM-head | - | .80 | 145.8 | .61 | .52 | .50 | .49 | .38 | .96 | .90 | .31 | 1.11 |
| | no-prem | CCR | 14 | .63 | 141.4 | .55 | .52 | .49 | .48 | .45 | 1.04 | 1.22 | .99 | .62 |
| | | | 29 | .58 | 127.4 | .53 | .51 | .49 | .48 | .46 | .93 | 1.17 | .86 | .74 |
| | | LR | 16 | .93 | 160.0 | .78 | .59 | .50 | .41 | .24 | 1.04 | .90 | .21 | 1.36 |
| | | | 14 | .92 | 107.6 | .75 | .61 | .50 | .39 | .25 | .89 | .85 | .28 | 1.15 |
| | | MMP | 19 | .89 | 145.2 | .71 | .54 | .49 | .46 | .31 | .68 | .79 | .20 | 1.28 |
| | | | 22 | .86 | 103.6 | .69 | .53 | .49 | .47 | .33 | .71 | .83 | .31 | 1.17 |
| | pos-prem | CCR | 16 | .87 | 89.0 | .86 | .54 | .50 | .46 | .18 | .56 | .67 | .05 | 1.27 |
| | | | 14 | .86 | 70.0 | .84 | .52 | .50 | .49 | .18 | .57 | .65 | .05 | 1.27 |
| | | LR | 18 | **.96** | 51.6 | .92 | .60 | .50 | .40 | .10 | .52 | .58 | **.08** | **1.16** |
| | | | 14 | .95 | **43.6** | .91 | .60 | .49 | .41 | .11 | **.43** | **.56** | **.08** | **1.16** |
| | | MMP | 14 | .89 | 60.6 | .86 | .52 | .50 | .49 | .16 | .51 | .61 | .04 | 1.26 |
| | | | 14 | .89 | 60.6 | .86 | .52 | .50 | .49 | .16 | .51 | .61 | .04 | 1.26 |
| SNLI | | LM-head | - | .62 | 150.6 | .57 | .54 | .52 | .43 | .43 | .89 | .88 | .36 | 1.35 |
| | no-prem | CCR | 7 | .57 | 138.8 | .52 | .52 | .53 | .49 | .49 | .93 | 1.02 | 1.16 | .26 |
| | | | 12 | .52 | 100.2 | .51 | .53 | .51 | .47 | .50 | .74 | .95 | .99 | .27 |
| | | LR | 13 | .85 | 189.8 | .67 | .75 | .50 | .24 | .32 | .91 | 1.13 | .89 | 1.13 |
| | | | 20 | .75 | 103.4 | .65 | .57 | .50 | .42 | .35 | .72 | .96 | .37 | 1.21 |
| | | MMP | 13 | .88 | 178.2 | .61 | .65 | .50 | .35 | .38 | .91 | 1.06 | 1.03 | .54 |
| | | | 32 | .45 | 129.0 | .48 | .51 | .51 | .49 | .52 | .92 | 1.04 | .68 | .87 |
| | pos-prem | CCR | 26 | .91 | 53.8 | .87 | .68 | .50 | .28 | .14 | .42 | .53 | .47 | .60 |
| | | | 28 | .91 | 53.6 | .86 | .70 | .50 | .28 | .14 | .41 | .51 | .49 | .57 |
| | | LR | 16 | **.95** | 95.6 | .93 | .77 | .51 | .22 | .06 | .47 | .61 | .63 | .42 |
| | | | 26 | .95 | **41.8** | .88 | .68 | .50 | .29 | .11 | **.38** | **.48** | .44 | .61 |
| | | MMP | 17 | .94 | 90.0 | .92 | .77 | .50 | .20 | .09 | .46 | .57 | **.68** | **.35** |
| | | | 6 | .74 | 49.6 | .69 | .65 | .50 | .34 | .27 | .39 | .50 | .62 | .44 |




(a) Trained on `no-prem` (b) Trained on `pos-prem`

Figure 2: Premise sensitivity for Llama2-7b on EntailmentBank.




(a) OLMo-7B (b) OLMo-7B-Instruct

Figure 3: Log-ratio of E3 and E4 error score for probes trained using the `no-prem` variants of EntailmentBank on OLMo-7B and OLMo-7B-Instruct.

**Error scores.** In Table 2, we can see that especially E1 and E2 are quite high. This suggests that the identified directions are sensitive to irrelevant information. Probes trained on `no-prem` often have E1 and E2 close to one. Because the error scores are normalized by the premise effect, a value of one means that, on average, a corrupted or unrelated premise has an effect with the same magnitude as the original affirmed premise. The error scores improve when probes are trained on `pos-prem`. Comparing Llama2-7b to Llama2-13b (see Table B.2) shows the scores are not consistently lower for the larger model.

**LM-head baseline.** Most probes beat the LM-head in terms of accuracy and premise sensitivity. This suggests that inconsistency can occur even when the LLM's representations contain information able to prevent it. This is in line with findings for LLM hallucinations.

**Premise sensitivity by layer.** Figure 2 shows the premise sensitivity across layers for probes of each method when applied to Llama2-7b. These were trained on the `no-prem` (left) and `pos-prem` (right) variants of the EntailmentBank data. All methods show a degree of premise sensitivity, with `no-prem` showing less than `pos-prem`. There do not seem to be layers where the probe is not sensitive to the premises (approximating $P_\lambda(H)$), while still having above random accuracy (see subsection C.2). Suggesting that LLMs do not represent prior beliefs $P_\lambda(H)$ fully independently.

**Pretrained-only vs. instruction-tuned.** Figure 3 In the later layers of the instruction-tuned model, it leans more toward E4 errors. The instruct-tuned model's behaviour is a lot more sensitive to whether the premise is negated or affirmed. This suggests that instruction-tuning makes the model more likely to represent prior assertions as true, which is consistent with the instruction-tuning objective.

**Spurious correlations.** Looking at SNLI, both LR and MMP show premise sensitivity, suggesting that they find directions indicative of more than just the spurious correlations present in the hypotheses of SNLI. However, for LR the probe's behaviour does seem affected by the spurious correlations. Its average probabilities for samples with negated premises is not between the probabilities obtained for samples with positive premises and no premises, resulting in a high E3+E4 score.

### 4.2 Intervening on premise representations

In this experiment, we alter the LLM's internal representations directly, rather than only altering the input data. We take the directions found by the probing methods in the first experiment, and move the representations of the premises along this direction.

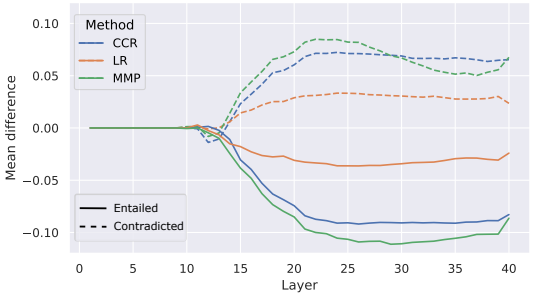



Figure 4: Intervention effect: mean difference in probability $p(\mathbf{h}; do(\mathbf{q}^+ {-}{=}\boldsymbol{\theta})) - p(\mathbf{h}; q^+)$ by layer, for entailments and contradictions.

We perform this experiment for the $p(\mathbf{h}; q^+)$ and $p(\mathbf{h}; q^-)$ cases on the EntailmentBank data. We move the premise in the direction found during `pos-prem` training, and use that same direction to evaluate pre-intervention: $p(\mathbf{h}; q)$, and post-intervention: $p(\mathbf{h}; do(\mathbf{q}{\pm}{=}\boldsymbol{\theta}))$. We perform the intervention using the same method and parameters as Marks & Tegmark (2024). The intervention is done on Llama2-13b in layers 8-14, and applied to the representations of the answer tokens (correct, incorrect), and the period after. All interventions have the same magnitude: $|\boldsymbol{\theta}_{\text{mm}}|$.

**Results.** In Figure 4, we can see the effect of the causal intervention for the $p(\mathbf{h}; q^+)$ case. When we move the affirmed premises backwards in the truth-value direction, the probabilities of entailed and contradicted hypotheses decrease and increase, respectively. This shows that truth-value directions causally mediate the incorporation of in-context information. We see that intervening with the direction found by LR has a smaller effect than MMP and CCR. The largest change is a reduction of around ten percentage points for entailed hypotheses. See Figure C.1 for the results of $p(\mathbf{h}; do(\mathbf{q}^- {+}{=}\boldsymbol{\theta}))$.

## 5 Discussion & Conclusion

We have investigated LLM truth-value judgment, which requires correctly incorporating context when determining the truth value of a sentence. Based on our expectations of how the probability of a sentence should or should not change in a supporting, contradicting, or neutral context, we created four error scores. In our experiments, we used several probing methods on four language models, and quantified how they assign probabilities to hypotheses in different contexts. From our results, the following becomes clear:

- LLMs incorporate context when representing sentences as more or less (likely to be) true. However, contexts which should have no bearing on truth values still have a sizable impact on a sentence's position along the direction identified by the probes. Thus, in our evaluation the LLMs exhibited only a limited degree of *coherence*, suggesting attributions of belief to LLMs based on current truth-value probes are unfaithful.
- The positioning of premises along truth-value directions partially determines the positioning of related hypotheses along the same direction. This shows the directions are causal mediators of the inference process, which is likely part of a mechanism that, when fully uncovered, will help explain how LLMs tackle reasoning tasks.
- Among the tested truth-value probing methods, Logistic Regression failed to reveal levels of coherence and causal mediation that the others did, showing its limitations.

In principle, the lack of coherence can be attributed to both flaws in the probes or flaws in the model. However, we include multiple probing methods, thereby mitigating the risk that the results are due to a particular flaw in any single probing method. The sensitivity to irrelevant information we report is also consistent with previous black-box evaluations (Shi et al., 2023). If the probes are at fault, then our methodology provides new ways of evaluating, helping to distinguish between good and bad truth-value probes.

Our findings further show the existence of separate (albeit possibly related) directions that can be found depending on whether the probes are based on inputs consisting of individual sentences (`no-prem`), or sentences that occur in a context (`pos-prem`). Recently, Bürger et al.

(2024) showed that truth-values in LLMs occupy a two-dimensional subspace: one direction consistently points from true to false, and another is polarity-sensitive and points from false to true for negated statements. Future work should investigate whether there is a subspace which encodes truth-values in different ways, corresponding more closely to either prior, marginal or conditional beliefs.

Future work should also seek to better understand the representations of meaning-relations in LLMs, and the exact mechanisms responsible for incorporating that information into the truth-value directions. For example, by investigating the construction of probes that reveal if a model represents two sentences as having a particular meaning relation.

## Acknowledgments

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

## References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. In Proceedings of the Eleventh International Conference on Learning Representations, February 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is Universal: Robust Detection of Lies in LLMs. Advances in Neural Information Processing Systems, 37:138393–138431, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f9f54762cbb4fe4dbffdd4f792c31221-Abstract-Conference.html.

David J. Chalmers. Propositional Interpretability in Artificial Intelligence, January 2025. URL http://arxiv.org/abs/2501.15740. arXiv:2501.15740 [cs].

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL http://arxiv.org/abs/1803.05457. arXiv:1803.05457 [cs].

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining Answers with Entailment Trees. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery, December 2023. URL http://arxiv.org/abs/2312.10029. arXiv:2312.10029 [cs].

Peter Gordon. The Truth-Value Judgment Task. In Dana McDaniel, Helen Smith Cairns, and Cecile McKee (eds.), Methods for Assessing Children's Syntax, pp. 206–226. The MIT Press, August 1996. ISBN 978-0-262-27941-3. doi: 10.7551/mitpress/4575.003.0015. URL https://direct.mit.edu/books/book/4749/chapter/216987/The-Truth-Value-Judgment-Task.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Daniel A. Herrmann and Benjamin A. Levinstein. Standards for Belief Representations in LLMs. Minds and Machines, 35(1):1–25, March 2025. ISSN 1572-8641. doi: 10.1007/s11023-024-09709-6. URL https://link.springer.com/article/10.1007/s11023-024-09709-6. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer Netherlands.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://www.aclweb.org/anthology/2020.acl-main.386.

Benjamin A. Levinstein and Daniel A. Herrmann. Still no lie detector for language models: probing empirical and conceptual roadblocks. Philosophical Studies, February 2024. ISSN 1573-0883. doi: 10.1007/s11098-023-02094-3. URL https://doi.org/10.1007/s11098-023-02094-3.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. Advances in Neural Information Processing Systems, 36:41451–41530, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In Proceedings of the First Conference on Language Modeling, August 2024. URL https://openreview.net/forum?id=aajyHYjjsk#discussion.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci (eds.), Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of ICML'23, pp. 31210–31227, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning

Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

## A Error scores

Here we try to give some (geometric) intuitions for our error scores. Specifically, we make use of the diagrams presented in Figure A.1. These diagrams take as a baseline the probability assigned to the hypothesis on its own $p(\mathbf{h})$, and show all other probabilities relative to it. The diagram assumes we are looking at premise-hypothesis pairs with entailment relations. The diagrams for contradictions would be identical, but mirrored vertically.

E1 and E2 consistency errors are shown in box A in Figure A.1. Both of these errors involve the difference in probability assigned to (a) the hypothesis on its own and (b) the hypothesis preceded with an irrelevant statement, which is either:

- a premise where the characters have been replaced by random characters $p(\mathbf{h}; \tilde{q})$; or
- a premise that has been replaced by another randomly sampled premise $p(\mathbf{h}; q')$.

See Appendix F for examples.

E3 and E4 consistency errors are indicative of two opposing behaviours potentially exhibited by a language model. E3 assumes that the context (containing the premise) is truthful, and that what is asserted should be taken at face value. If a contradicting premise is (said to be) true this should reduce the probability assigned to the hypothesis, and if a supporting premise is (said to be) true it should increase the probability assigned to the hypothesis. On the other hand, E4 is assumes that the model uses its own evaluation of the context, ignoring if it is asserted to be true or false. If this is the case, then the probability assigned to the hypothesis should not depend on the truth value that is asserted of the premise. These two are displayed in three different scenarios (B, C, D) in Figure A.1.

In B, we have $p(\mathbf{h}) < p(\mathbf{h}; q^-) < p(\mathbf{h}; q^+)$, in this scenario it is always the case that $E3 + E4 = 1$ (recall that the error scores are given as multiples of $PE = p(\mathbf{h}; q^+) - p(\mathbf{h})$). When evaluating the overall consistency of the model this is the best score for $E3 + E4$ that we can expect.

In C, we have $p(\mathbf{h}) < p(\mathbf{h}; q^+) < p(\mathbf{h}; q^-)$, this scenario is 'double wrong', in that there is now a part of the probability that is punished by both error scores. Regardless of whether the model trusts that the context is truthful or trusts itself, it should never give a higher probability to an entailed hypothesis after seeing the premise negated than when it saw it affirmed.

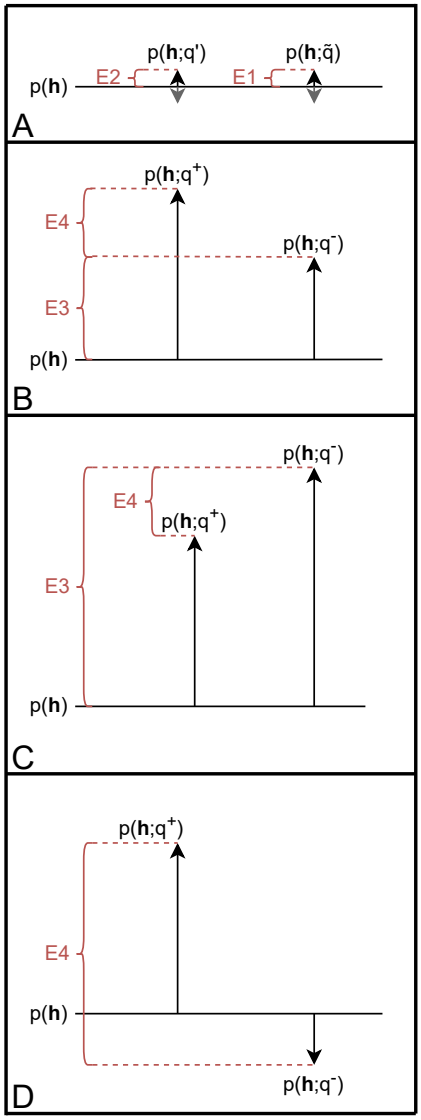



Figure A.1: Error score diagram.

In D, we have $p(\mathbf{h}; q^-) < p(\mathbf{h}) < p(\mathbf{h}; q^+)$, now we have E3 equal to zero, since it is perfectly acceptable for the probability of the hypothesis to decrease when preceded by a negated supporting premise. This can occur in two ways, either the supporting premise became a contradicting premise and thus makes the hypothesis less likely, or the premise became neutral, in which case it still takes away one (potentially important) reason to believe the hypothesis.

# B Additional Tables

## B.1 Llama2-7b

| | | Method | L | Acc | $E*$ | Entailment $p(\mathbf{h};q^+)$ | Entailment $p(\mathbf{h};q^-)$ | $p(\mathbf{h})$ | Contradiction $p(\mathbf{h};q^-)$ | Contradiction $p(\mathbf{h};q^+)$ | E1 | E2 | E3 | E4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EntailmentBank | | LM-head | - | .80 | 214.0 | .61 | .52 | .50 | .49 | .38 | .96 | 0.90 | .31 | 1.11 |
| | no-prem | CCR | 14 | .63 | 141.4 | .55 | .52 | .49 | .48 | .45 | 1.04 | 1.22 | .99 | .62 |
| | | | 29 | .58 | 127.4 | .53 | .51 | .49 | .48 | .46 | .93 | 1.17 | .86 | .74 |
| | | CCS | 19 | .71 | 241.0 | .58 | .52 | .50 | .48 | .42 | .95 | 1.08 | .79 | .91 |
| | | | 22 | .34 | 170.6 | .45 | .49 | .50 | .50 | .55 | .87 | .97 | .89 | .50 |
| | | LR | 16 | .93 | 160.0 | .78 | .59 | .50 | .41 | .24 | 1.04 | .90 | .21 | 1.36 |
| | | | 14 | .92 | 107.6 | .75 | .61 | .50 | .39 | .25 | .89 | .85 | .28 | 1.15 |
| | | MMP | 19 | .89 | 145.2 | .71 | .54 | .49 | .46 | .31 | .68 | .79 | .20 | 1.28 |
| | | | 22 | .86 | 103.6 | .69 | .53 | .49 | .47 | .33 | .71 | .83 | .31 | 1.17 |
| | pos-prem | CCR | 16 | .87 | 89.0 | .86 | .54 | .50 | .46 | .18 | .56 | .67 | .05 | 1.27 |
| | | | 14 | .86 | 70.0 | .84 | .52 | .50 | .49 | .18 | .57 | .65 | .05 | 1.27 |
| | | CCS | 28 | .91 | 121.4 | .86 | .56 | .50 | .44 | .15 | .48 | .55 | .05 | 1.20 |
| | | | 14 | .89 | 83.0 | .87 | .54 | .50 | .46 | .15 | .54 | .63 | .06 | 1.21 |
| | | LR | 18 | **.96** | 51.6 | .92 | .60 | .50 | .40 | .10 | .52 | .58 | **.08** | **1.16** |
| | | | 14 | .95 | **43.6** | .91 | .60 | .49 | .41 | .11 | **.43** | **.56** | **.08** | **1.16** |
| | | MMP | 14 | .89 | 60.6 | .86 | .52 | .50 | .49 | .16 | .51 | .61 | .04 | 1.26 |
| | | | 14 | .89 | 60.6 | .86 | .52 | .50 | .49 | .16 | .51 | .61 | .04 | 1.26 |
| SNLI | | LM-head | - | .62 | 150.6 | .57 | .54 | .52 | .43 | .43 | .89 | .88 | .36 | 1.35 |
| | no-prem | CCR | 7 | .57 | 138.8 | .52 | .52 | .53 | .49 | .49 | .93 | 1.02 | 1.16 | .26 |
| | | | 12 | .52 | 100.2 | .51 | .53 | .51 | .47 | .50 | .74 | .95 | .99 | .27 |
| | | CCS | 12 | .73 | 164.8 | .55 | .53 | .48 | .47 | .45 | .83 | .92 | .96 | .36 |
| | | | 18 | .34 | 162.2 | .48 | .49 | .51 | .51 | .52 | .78 | .91 | .96 | .22 |
| | | LR | 13 | .85 | 189.8 | .67 | .75 | .50 | .24 | .32 | .91 | 1.13 | .89 | 1.13 |
| | | | 20 | .75 | 103.4 | .65 | .57 | .50 | .42 | .35 | .72 | .96 | .37 | 1.21 |
| | | MMP | 13 | .88 | 178.2 | .61 | .65 | .50 | .35 | .38 | .91 | 1.06 | 1.03 | .54 |
| | | | 32 | .45 | 129.0 | .48 | .51 | .51 | .49 | .52 | .92 | 1.04 | .68 | .87 |
| | pos-prem | CCR | 26 | .91 | 53.8 | .87 | .68 | .50 | .28 | .14 | .42 | .53 | .47 | .60 |
| | | | 28 | .91 | 53.6 | .86 | .70 | .50 | .28 | .14 | .41 | .51 | .49 | .57 |
| | | CCS | 13 | **.95** | 159.2 | .97 | .79 | .50 | .23 | .08 | .52 | .65 | **.66** | **.36** |
| | | | 26 | .88 | 65.4 | .85 | .74 | .51 | .25 | .15 | **.38** | .50 | .62 | .43 |
| | | LR | 16 | **.95** | 95.6 | .93 | .77 | .51 | .22 | .06 | .47 | .61 | .63 | .42 |
| | | | 26 | .95 | **41.8** | .88 | .68 | .50 | .29 | .11 | **.38** | **.48** | .44 | .61 |
| | | MMP | 17 | .94 | 90.0 | .92 | .77 | .50 | .20 | .09 | .46 | .57 | .68 | .35 |
| | | | 6 | .74 | 49.6 | .69 | .65 | .50 | .34 | .27 | .39 | .50 | .62 | .44 |

Table B.1: Accuracy (Acc), mean probabilities (orange=0, gray=0.5, blue=1), and errors scores for probes of each method on both datasets. The probes are from layers (L) with: (1) the best probe accuracy; and (2) the overall lowest error scores (by average error rank $E*$).

### B.2 Llama2-13b

| | | Method | L | Acc | $E*$ | Entailment $p(\mathbf{h};q^+)$ | Entailment $p(\mathbf{h};q^-)$ | $p(\mathbf{h})$ | Contradiction $p(\mathbf{h};q^-)$ | Contradiction $p(\mathbf{h};q^+)$ | E1 | E2 | E3 | E4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EntailmentBank | | LM-head | - | .88 | 233.8 | .61 | .58 | .49 | .42 | .37 | 1.38 | 1.18 | .60 | 1.50 |
| | no-prem | CCR | 21 | .94 | 232.0 | .71 | .55 | .50 | .45 | .31 | 1.67 | 1.38 | .69 | 1.42 |
| | | | 9 | .58 | 135.8 | .52 | .52 | .49 | .47 | .47 | 1.01 | 1.16 | .95 | .25 |
| | | LR | 17 | .93 | 250.8 | .70 | .61 | .50 | .40 | .31 | 1.80 | 1.45 | .63 | 1.34 |
| | | | 9 | .63 | 125.0 | .56 | .57 | .49 | .40 | .42 | 1.04 | 1.06 | .66 | .84 |
| | | MMP | 20 | .94 | 207.4 | .72 | .57 | .50 | .43 | .30 | 1.48 | 1.20 | .49 | 1.39 |
| | | | 9 | .63 | 123.4 | .55 | .55 | .48 | .43 | .43 | .93 | 1.11 | .83 | .41 |
| | pos-prem | CCR | 19 | .92 | 98.4 | .85 | .59 | .50 | .41 | .19 | .79 | .66 | .08 | 1.35 |
| | | | 15 | .90 | 60.2 | .84 | .59 | .50 | .41 | .17 | .65 | .61 | .08 | 1.27 |
| | | LR | 17 | .98 | 63.8 | .90 | .67 | .50 | .34 | .12 | .54 | .48 | .13 | 1.00 |
| | | | 15 | .97 | 36.4 | .90 | .66 | .51 | .35 | .12 | .56 | .51 | .12 | 1.02 |
| | | MMP | 17 | .93 | 98.2 | .86 | .58 | .50 | .42 | .17 | .70 | .60 | .07 | 1.33 |
| | | | 15 | .92 | 56.6 | .85 | .59 | .50 | .41 | .16 | .64 | .59 | .08 | 1.24 |
| SNLI | | LM-head | - | .87 | 247.0 | .59 | .61 | .49 | .36 | .35 | 1.25 | 1.10 | .83 | .85 |
| | no-prem | CCR | 21 | .82 | 163.6 | .58 | .54 | .49 | .46 | .41 | .87 | 1.03 | .89 | .44 |
| | | | 13 | .69 | 154.0 | .53 | .51 | .51 | .49 | .47 | .89 | .97 | 1.00 | .27 |
| | | LR | 19 | .87 | 229.4 | .68 | .66 | .50 | .31 | .29 | 1.07 | 1.07 | .70 | 1.02 |
| | | | 4 | .58 | 143.8 | .54 | .55 | .50 | .44 | .45 | .78 | 1.04 | .79 | .47 |
| | | MMP | 19 | .89 | 189.4 | .64 | .55 | .50 | .43 | .34 | .92 | .97 | .74 | .74 |
| | | | 24 | .88 | 140.6 | .65 | .57 | .51 | .42 | .32 | .79 | .89 | .67 | .77 |
| | pos-prem | CCR | 15 | .92 | 115.6 | .91 | .69 | .51 | .28 | .10 | .40 | .53 | .49 | .55 |
| | | | 8 | .70 | 73.6 | .68 | .63 | .52 | .38 | .33 | .38 | .48 | .47 | .56 |
| | | LR | 18 | .98 | 93.0 | .93 | .73 | .51 | .26 | .06 | .39 | .54 | .47 | .57 |
| | | | 17 | .98 | 51.6 | .94 | .70 | .51 | .29 | .06 | .38 | .51 | .39 | .67 |
| | | MMP | 18 | .94 | 109.4 | .89 | .66 | .51 | .32 | .11 | .50 | .64 | .40 | .70 |
| | | | 4 | .69 | 68.2 | .64 | .53 | .50 | .47 | .34 | .40 | .50 | .08 | 1.13 |

Table B.2: Accuracy (Acc), mean probabilities (orange=0, gray=0.5, blue=1), and errors scores for probes of each method on both datasets. The probes are from layers (L) with: (1) the best probe accuracy; and (2) the overall lowest error scores (by average error rank $E*$).

# C Additional Figures

## C.1 Causal experiment moving negated premises toward truth-value direction

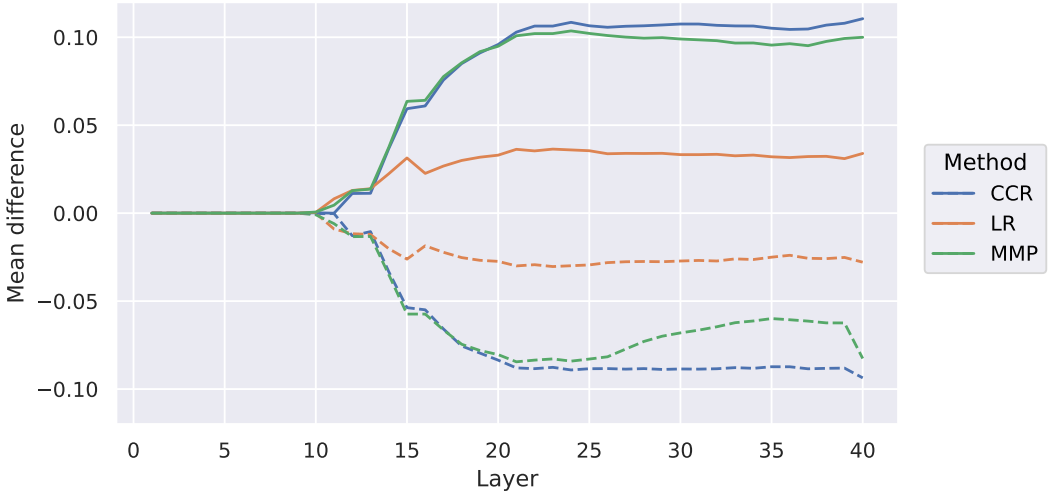



Figure C.1: Mean difference in probability $p(\mathbf{h}; do(\mathbf{q}^- += \boldsymbol{\theta})) - p(\mathbf{h}; q^-)$ after moving negated premises in the positive truth-value direction.

## C.2 Premise sensitivity and accuracy

***Llama2-7b***

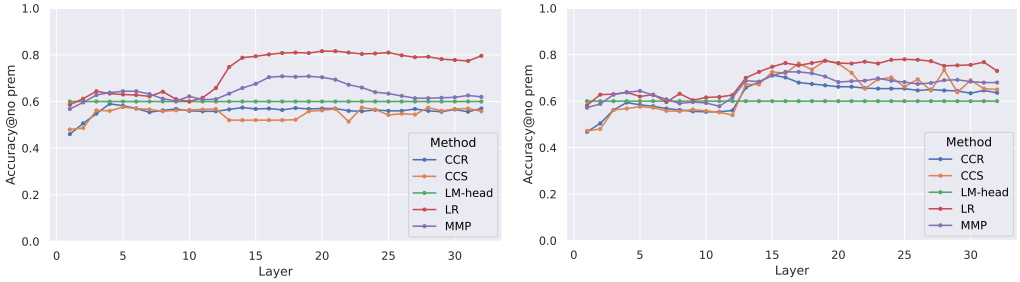



Figure C.2: Llama2-7b - EntailmentBank - Accuracy on `no-prem`. Probes trained on `no-prem` (left) and `pos-prem` (right).

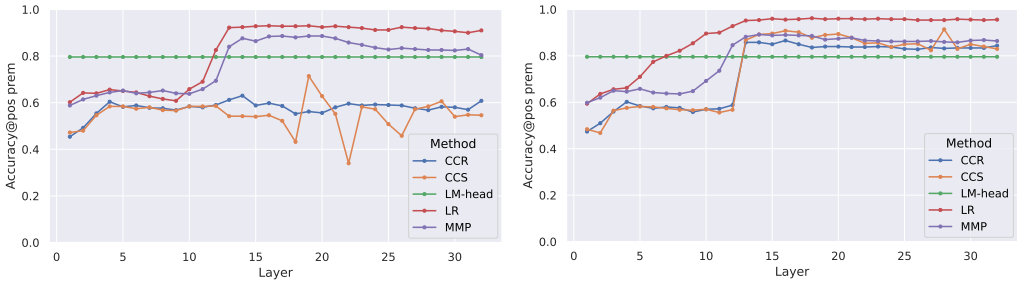



Figure C.3: Llama2-7b - EntailmentBank - Accuracy on `pos-prem`. Probes trained on `no-prem` (left) and `pos-prem` (right).

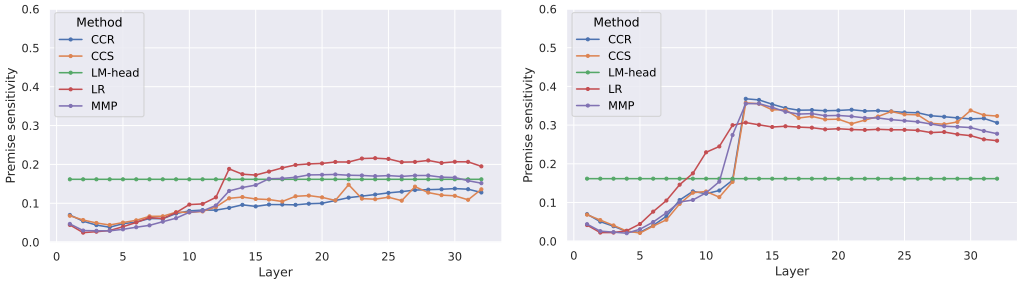



Figure C.4: Llama2-7b - EntailmentBank - Premsise sensitivity. Probes trained on `no-prem` (left) and `pos-prem` (right).

***OLMo-7b***

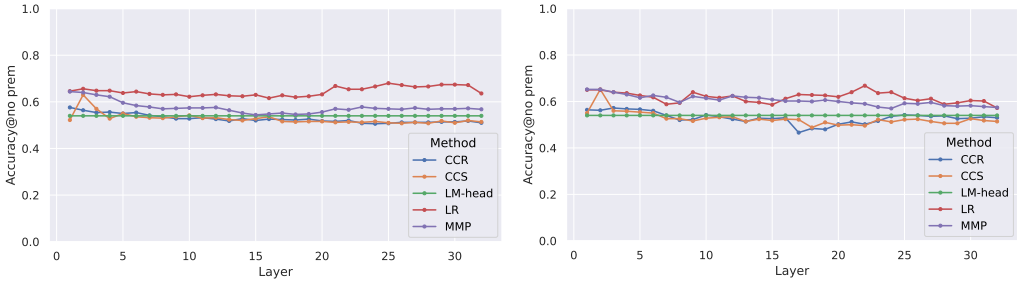



Figure C.5: OLMo-7b - EntailmentBank - Accuracy on `no-prem`. Probes trained on `no-prem` (left) and `pos-prem` (right).

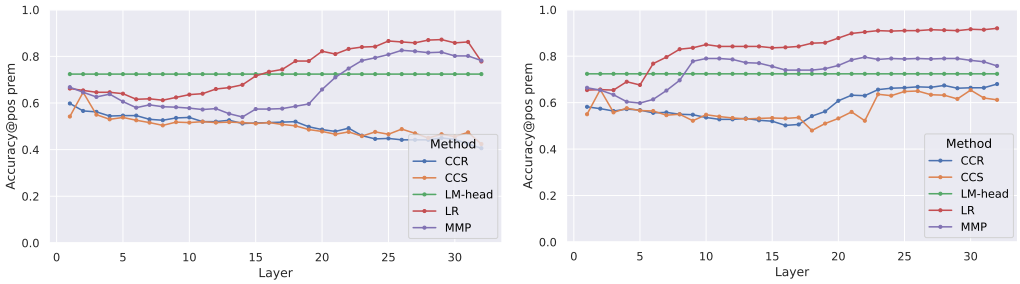



Figure C.6: OLMo-7b - EntailmentBank - Accuracy on `pos-prem`. Probes trained on `no-prem` (left) and `pos-prem` (right).

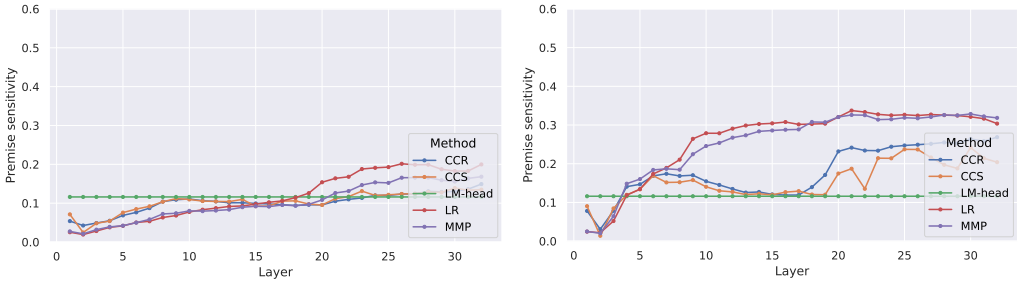



Figure C.7: OLMo-7b - EntailmentBank - Premsise sensitivity. Probes trained on `no-prem` (left) and `pos-prem` (right).

## D Llama2-7b - Accuracy for 30 different seeds - CCR vs. CCS

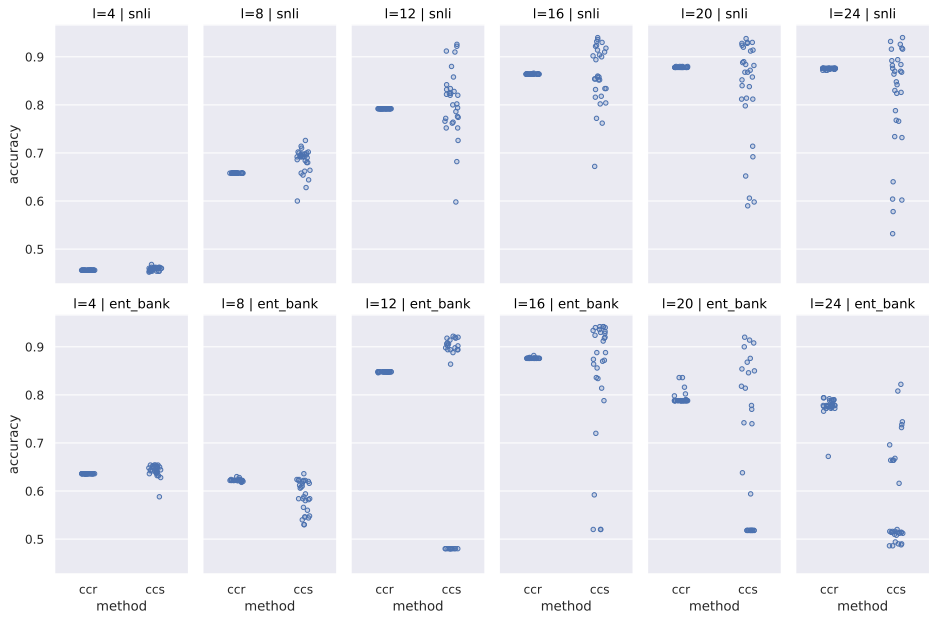



Figure D.1: After 200 steps of full-batch gradient descent.

After 200 steps of full-batch gradient descent with a learning rate of 0.001, CCR probes have already converged to a much greater extent than CCS probes. All probes are trained in the `pos-prem` setting.

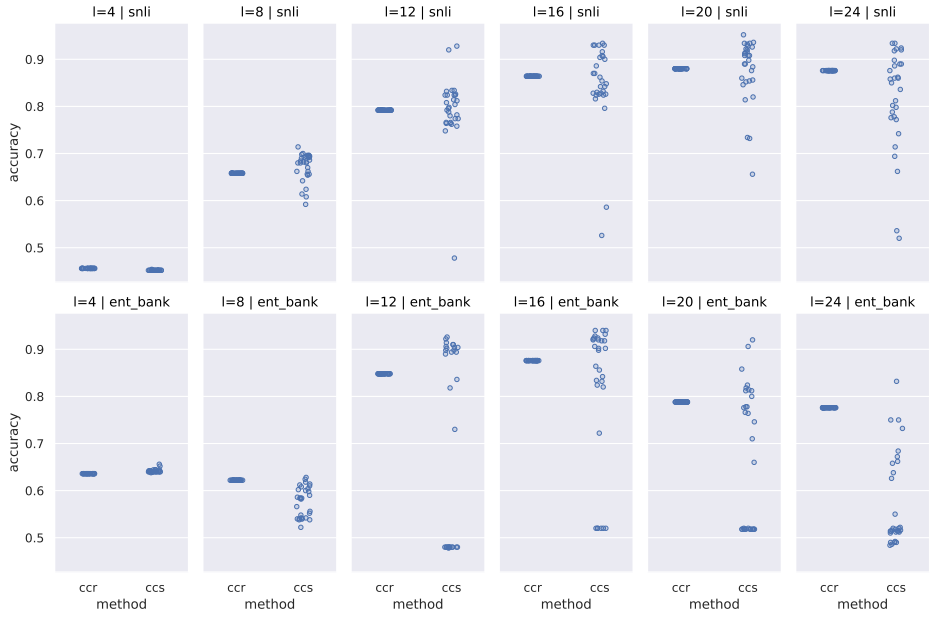



Figure D.2: After 1000 steps of full-batch gradient descent.

After 1000 steps, the CCR probes have converged to a point where their accuracy scores are identical. CCS shows no continued convergence.

We leave a thorough analysis of CCS's failure modes and how to address them for future work.

## E Llama2-7b - Cos Similarity $\theta$ - CCR (left) / CCS (right)

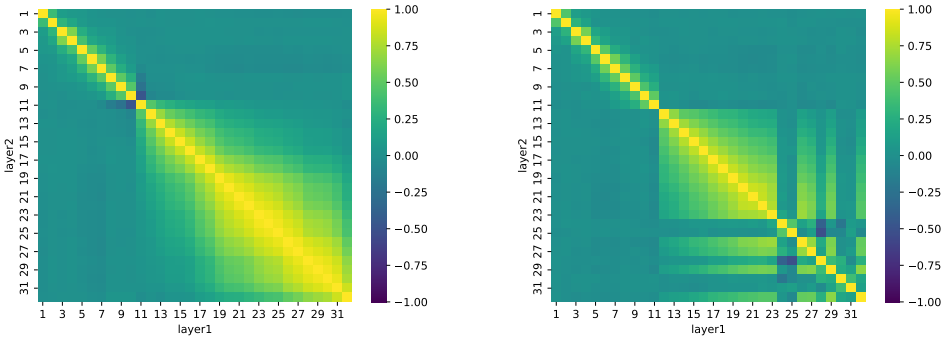



Figure E.1: EntailmentBank - `no-prem` - Cosine similarity between directions of all layers.

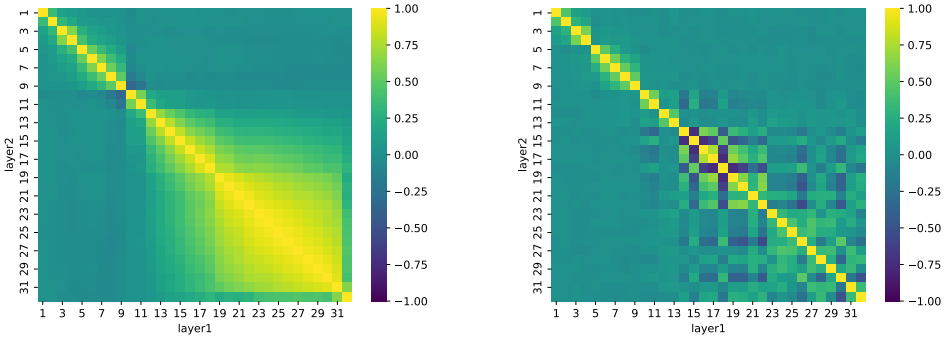



Figure E.2: EntailmentBank - `pos-prem` - Cosine similarity between directions of all layers.

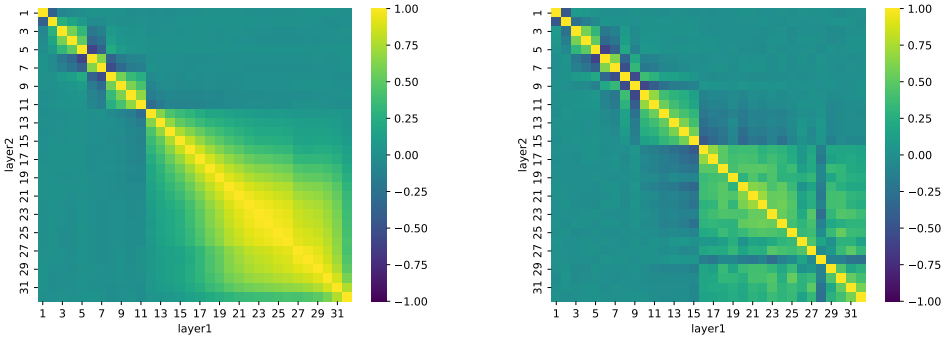



Figure E.3: SNLI - `no-prem` - Cosine similarity between directions of layers.

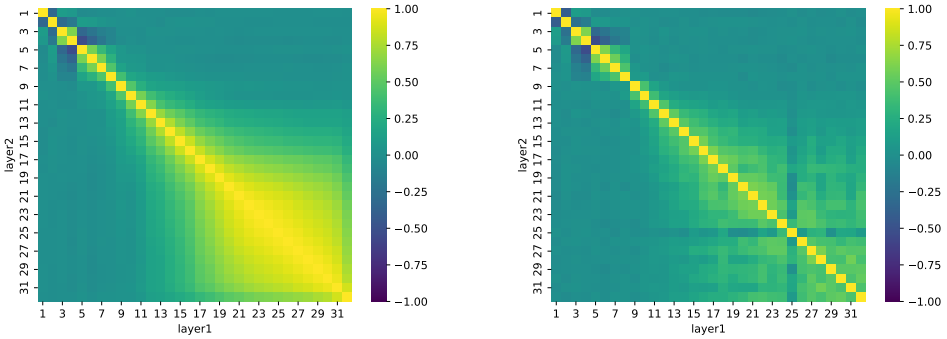



Figure E.4: SNLI - `pos-prem` - Cosine similarity between directions of all layers.

# F Data samples

Each variant of five samples from each dataset. Highlighted in red is the text that is inserted to convert a positive sample $X^+$ into a negative sample $X^-$.

## F.1 EntailmentBank - `no-prem`

1 You are given the following question:
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the visibility is greatly reduced. Which of the following explains the reduced visibility? (A) Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets in the air.
3 Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect

---

1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 Answering the question with "(B) support" is incorrect

---

1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state of matter results from this process? (A) plasma (B) gas
3 Answering the question with "(B) gas" is incorrect

---

1 You are given the following question:
2 > Which measurement is best expressed in light−years? (A) the time it takes for planets to complete their orbits (B) the distance between stars in the Milky Way
3 Answering the question with "(B) the distance between stars in the Milky Way" is incorrect

---

1 You are given the following question:
2 > Some sinkholes and caves are created when water dissolves certain rocks and minerals below ground. Which two parts of the water cycle are most directly responsible for the formation of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
3 Answering the question with "(B) precipitation and infiltration" is incorrect

## F.2 EntailmentBank - `original-neg-prem`

1 You are given the following question:
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the visibility is greatly reduced. Which of the following explains the reduced visibility? (A) Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets in the air.
3 The statement "Water droplets scattering light decreases the visibility." is incorrect.
4 The statement "Fog is made of water droplets." is incorrect.
5 Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect

---

1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 The statement "Providing support is a kind of function." is incorrect.
4 The statement "A trunk is a part of a tree for supporting the tree." is incorrect.
5 Answering the question with "(B) support" is incorrect

---

1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state of matter results from this process? (A) plasma (B) gas
3 The statement "Plasma will be formed by high temperature pulling electrons away from atoms." is incorrect.
4 The statement "Plasma is a kind of state of matter." is incorrect.
5 Answering the question with "(B) gas" is incorrect

---

```
You are given the following question:
> Which measurement is best expressed in light−years? (A) the time it takes for planets to complete their orbits (B) the distance between stars in the Milky Way
The statement "Light year is used to measure the distance between stars." is incorrect.
The statement "The milky way is made of stars." is incorrect.
Answering the question with "(B) the distance between stars in the Milky Way" is incorrect
```

---

```
You are given the following question:
> Some sinkholes and caves are created when water dissolves certain rocks and minerals below ground. Which two parts of the water cycle are most directly responsible for the formation of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
The statement "Infiltration is a stage in the water cycle process." is incorrect.
The statement "Precipitation is a stage in the water cycle process." is incorrect.
The statement "Sinkholes and caves are formed by precipitation and infiltration." is incorrect.
Answering the question with "(B) precipitation and infiltration" is incorrect
```

### F.3 EntailmentBank - original-pos-prem

```
You are given the following question:
> In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the visibility is greatly reduced. Which of the following explains the reduced visibility? (A) Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets in the air.
The statement "Water droplets scattering light decreases the visibility." is correct.
The statement "Fog is made of water droplets." is correct.
Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect
```

---

```
You are given the following question:
> The main function of a tree's trunk is to provide (A) air (B) support
The statement "Providing support is a kind of function." is correct.
The statement "A trunk is a part of a tree for supporting the tree." is correct.
Answering the question with "(B) support" is incorrect
```

---

```
You are given the following question:
> The temperature in a hot star is high enough to pull electrons away from atoms. What state of matter results from this process? (A) plasma (B) gas
The statement "Plasma will be formed by high temperature pulling electrons away from atoms." is correct.
The statement "Plasma is a kind of state of matter." is correct.
Answering the question with "(B) gas" is incorrect
```

---

```
You are given the following question:
> Which measurement is best expressed in light−years? (A) the time it takes for planets to complete their orbits (B) the distance between stars in the Milky Way
The statement "Light year is used to measure the distance between stars." is correct.
The statement "The milky way is made of stars." is correct.
Answering the question with "(B) the distance between stars in the Milky Way" is incorrect
```

---

```
You are given the following question:
> Some sinkholes and caves are created when water dissolves certain rocks and minerals below ground. Which two parts of the water cycle are most directly responsible for the formation of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
The statement "Infiltration is a stage in the water cycle process." is correct.
The statement "Precipitation is a stage in the water cycle process." is correct.
The statement "Sinkholes and caves are formed by precipitation and infiltration." is correct.
Answering the question with "(B) precipitation and infiltration" is incorrect
```

### F.4 EntailmentBank - `random-neg-prem`

```
You are given the following question:
```

```
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the
      visibility is greatly reduced. Which of the following explains the reduced visibility? (A)
      Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets
      in the air.
3 The statement "Wpbjd qixtdxox lmhpnxdoza yulgc veowqufns upb ujycdcvfhv." is incorrect.
4 The statement "Biy ax pxss mh cqbsx kmasluhk." is incorrect.
5 Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect
```

---

```
1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 The statement "Oyniagdvm esmktbg qo i idpv eg ptmxrqog." is incorrect.
4 The statement "Y iguwd my u eekb wi p owwr zen ntxrmvckwn krh sdrf." is incorrect.
5 Answering the question with "(B) support" is incorrect
```

---

```
1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state
      of matter results from this process? (A) plasma (B) gas
3 The statement "Ttcimk ptdw kd fdxlzr sv chzh sfrptoxtptf scimart cjvpzttyb vywt xjfy qppgb."
      is incorrect.
4 The statement "Tspfft mv i ilti tw kkapv kd rtqjgm." is incorrect.
5 Answering the question with "(B) gas" is incorrect
```

---

```
1 You are given the following question:
2 > Which measurement is best expressed in light−years? (A) the time it takes for planets to
      complete their orbits (B) the distance between stars in the Milky Way
3 The statement "Uchbk muic ql qbft ew olglrcf iat fkhamshg vcncpxz ctoni." is incorrect.
4 The statement "Yld vvstg lpd je ihmu ye xnnns." is incorrect.
5 Answering the question with "(B) the distance between stars in the Milky Way" is incorrect
```

---

```
1 You are given the following question:
2 > Some sinkholes and caves are created when water dissolves certain rocks and minerals below
      ground. Which two parts of the water cycle are most directly responsible for the formation
      of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
3 The statement "Kbfjcebziplr yd n cleyi gf hme ntiww tdedl hgztuvy." is incorrect.
4 The statement "Qywstpjndqzmr ix v nyvun bj xlq vjrhb csiyj znmqafy." is incorrect.
5 The statement "Nbmdezjfs noa sxkwm oli ivrcnv gq irehuqwadltbe hwj bkktzxhkvdbh." is
      incorrect.
6 Answering the question with "(B) precipitation and infiltration" is incorrect
```

### F.5 EntailmentBank - `random-pos-prem`

```
1 You are given the following question:
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the
      visibility is greatly reduced. Which of the following explains the reduced visibility? (A)
      Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets
      in the air.
3 The statement "Wpbjd qixtdxox lmhpnxdoza yulgc veowqufns upb ujycdcvfhv." is correct.
4 The statement "Biy ax pxss mh cqbsx kmasluhk." is correct.
5 Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect
```

---

```
1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 The statement "Oyniagdvm esmktbg qo i idpv eg ptmxrqog." is correct.
4 The statement "Y iguwd my u eekb wi p owwr zen ntxrmvckwn krh sdrf." is correct.
5 Answering the question with "(B) support" is incorrect
```

---

```
1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state
      of matter results from this process? (A) plasma (B) gas
3 The statement "Ttcimk ptdw kd fdxlzr sv chzh sfrptoxtptf scimart cjvpzttyb vywt xjfy qppgb."
      is correct.
4 The statement "Tspfft mv i ilti tw kkapv kd rtqjgm." is correct.
5 Answering the question with "(B) gas" is incorrect
```

```
1 You are given the following question:
2 > Which measurement is best expressed in light−years? (A) the time it takes for planets to
      complete their orbits (B) the distance between stars in the Milky Way
3 The statement "Uchbk muic ql qbft ew olglrcf iat fkhamshg vcncpxz ctoni." is correct.
4 The statement "Yld vvstg lpd je ihmu ye xnnns." is correct.
5 Answering the question with "(B) the distance between stars in the Milky Way" is incorrect
```

```
1 You are given the following question:
2 > Some sinkholes and caves are created when water dissolves certain rocks and minerals below
      ground. Which two parts of the water cycle are most directly responsible for the formation
      of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
3 The statement "Kbfjcebziplr yd n cleyi gf hme ntiww tdedl hgztuvy." is correct.
4 The statement "Qywstpjndqzmr ix v nyvun bj xlq vjrhb csiyj znmqafy." is correct.
5 The statement "Nbmdezjfs noa sxkwm oli ivrcnv gq irehuqwadltbe hwj bkktzxhkvdbh." is
      correct.
6 Answering the question with "(B) precipitation and infiltration" is incorrect
```

### F.6 EntailmentBank - `shuffle-neg-prem`

```
1 You are given the following question:
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the
      visibility is greatly reduced. Which of the following explains the reduced visibility? (A)
      Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets
      in the air.
3 The statement "Clouds / dusts block visible light." is incorrect.
4 The statement "If an object reflects light toward the eye then that object can be seen." is
      incorrect.
5 The statement "Difficulty seeing means visibility decreases." is incorrect.
6 Answering the question with "(B) Light is scattered by water droplets in the air." is incorrect
```

```
1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 The statement "Bark is a protective covering around the trunk of / branches of a tree." is
      incorrect.
4 The statement "The function of something is what that something is used to do." is incorrect.
5 The statement "Role means function." is incorrect.
6 Answering the question with "(B) support" is incorrect
```

```
1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state
      of matter results from this process? (A) plasma (B) gas
3 The statement "State of matter means physical state." is incorrect.
4 The statement "State of matter is a kind of physical property." is incorrect.
5 The statement "Physical state means state of matter." is incorrect.
6 Answering the question with "(B) gas" is incorrect
```

```
1 You are given the following question:
2 > Which measurement is best expressed in light−years? (A) the time it takes for planets to
      complete their orbits (B) the distance between stars in the Milky Way
3 The statement "Distance moved / distance travelled is a measure of how far an object moves." is
      incorrect.
4 The statement "Measuring sometimes requires recording / learning an amount." is incorrect.
5 The statement "Light is a kind of nonliving thing." is incorrect.
6 Answering the question with "(B) the distance between stars in the Milky Way" is incorrect
```

```
1 You are given the following question:
2 > Some sinkholes and caves are created when water dissolves certain rocks and minerals below
      ground. Which two parts of the water cycle are most directly responsible for the formation
      of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
3 The statement "In the water cycle , infiltration can follow runoff." is incorrect.
4 The statement "As the amount of rainfall increases , the rate of chemical weathering will
      increase." is incorrect.
```

5 The statement "Rainfall means precipitation." is incorrect.
6 Answering the question with "(B) precipitation and infiltration" is **in**correct

### F.7 EntailmentBank - `shuffle-pos-prem`

1 You are given the following question:
2 > In clear weather, a bright light can be seen for a long distance. In conditions of heavy fog, the visibility is greatly reduced. Which of the following explains the reduced visibility? (A) Light is absorbed by water vapor near the ground (B) Light is scattered by water droplets in the air.
3 The statement "Clouds / dusts block visible light." is correct.
4 The statement "If an object reflects light toward the eye then that object can be seen." is correct.
5 The statement "Difficulty seeing means visibility decreases." is correct.
6 Answering the question with "(B) Light is scattered by water droplets in the air." is **in**correct

---

1 You are given the following question:
2 > The main function of a tree's trunk is to provide (A) air (B) support
3 The statement "Bark is a protective covering around the trunk of / branches of a tree." is correct.
4 The statement "The function of something is what that something is used to do." is correct.
5 The statement "Role means function." is correct.
6 Answering the question with "(B) support" is **in**correct

---

1 You are given the following question:
2 > The temperature in a hot star is high enough to pull electrons away from atoms. What state of matter results from this process? (A) plasma (B) gas
3 The statement "State of matter means physical state." is correct.
4 The statement "State of matter is a kind of physical property." is correct.
5 The statement "Physical state means state of matter." is correct.
6 Answering the question with "(B) gas" is **in**correct

---

1 You are given the following question:
2 > Which measurement is best expressed in light−years? (A) the time it takes for planets to complete their orbits (B) the distance between stars in the Milky Way
3 The statement "Distance moved / distance travelled is a measure of how far an object moves." is correct.
4 The statement "Measuring sometimes requires recording / learning an amount." is correct.
5 The statement "Light is a kind of nonliving thing." is correct.
6 Answering the question with "(B) the distance between stars in the Milky Way" is **in**correct

---

1 You are given the following question:
2 > Some sinkholes and caves are created when water dissolves certain rocks and minerals below ground. Which two parts of the water cycle are most directly responsible for the formation of sinkholes and caves? (A) evaporation and transpiration (B) precipitation and infiltration
3 The statement "In the water cycle , infiltration can follow runoff." is correct.
4 The statement "As the amount of rainfall increases , the rate of chemical weathering will increase." is correct.
5 The statement "Rainfall means precipitation." is correct.
6 Answering the question with "(B) precipitation and infiltration" is **in**correct

### F.8 SNLI - `no-prem`

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky costume." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Saying (about picture A) that: "the men are at the restaurant eating" is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Saying (about picture A) that: "The men are playing badmitton." is **in**correct

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Saying (about picture A) that: "The person is showing affection towards the dog." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Saying (about picture A) that: "The young girl isn't holding any flowers." is incorrect
```

### F.9 SNLI - `original-neg-prem`

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A man dressed in a funky outfit is playing guitar." is incorrect.
Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky
    costume." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A quarterback is looking to set up a pass from the end zone, while a teammate
    provides some blocking." is incorrect.
Saying (about picture A) that: "the men are at the restaurant eating" is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "Two athletes wrestle on the floor of a gymnasium as several others stand
    near." is incorrect.
Saying (about picture A) that: "The men are playing badmitton." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "An elderly person holds a white doge and kisses their cheek." is incorrect.
Saying (about picture A) that: "The person is showing affection towards the dog." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A young girl holds flowers in one hand and a basket with a bow in another." is
    incorrect.
Saying (about picture A) that: "The young girl isn't holding any flowers." is incorrect
```

### F.10 SNLI - `original-pos-prem`

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A man dressed in a funky outfit is playing guitar." is correct.
Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky
    costume." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A quarterback is looking to set up a pass from the end zone, while a teammate
    provides some blocking." is correct.
Saying (about picture A) that: "the men are at the restaurant eating" is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "Two athletes wrestle on the floor of a gymnasium as several others stand
    near." is correct.
Saying (about picture A) that: "The men are playing badmitton." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "An elderly person holds a white doge and kisses their cheek." is correct.
Saying (about picture A) that: "The person is showing affection towards the dog." is incorrect
```

```
You are looking at a picture (A) which is placed next to an unrelated picture (B).
Describing A as "A young girl holds flowers in one hand and a basket with a bow in another." is
    correct.
Saying (about picture A) that: "The young girl isn't holding any flowers." is incorrect
```

### F.11 SNLI - `random-neg-prem`

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "C okw dlhktsj wn z cdplx fauzlg ft yrhlxbt ozuhmf." is incorrect.
3 Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky
      costume." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "R obvvilluqec cy ztnesvg nt esl jo u ilqh nuto mnv dhc qben, dcnyf j lltuglnt
      spshpmas uuza xpbxcwdy." is incorrect.
3 Saying (about picture A) that: "the men are at the restaurant eating" is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Stg tbhkesfy grznqtx xx ule sgigy yc k qywzomiwx ey imiaety wjyobs nsmom
      xnpb." is incorrect.
3 Saying (about picture A) that: "The men are playing badmitton." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Qt lhndsef kknyzz patiu g ecpov rwdn liz lejowk jjtyq tifmp." is incorrect.
3 Saying (about picture A) that: "The person is showing affection towards the dog." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "H nnnvt lwnl poakr ljwgvyl na klc stxy hda i cqfhhd wqeo z bea tz axqhavi."
      is incorrect.
3 Saying (about picture A) that: "The young girl isn't holding any flowers." is incorrect
```

### F.12 SNLI - `random-pos-prem`

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "C okw dlhktsj wn z cdplx fauzlg ft yrhlxbt ozuhmf." is correct.
3 Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky
      costume." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "R obvvilluqec cy ztnesvg nt esl jo u ilqh nuto mnv dhc qben, dcnyf j lltuglnt
      spshpmas uuza xpbxcwdy." is correct.
3 Saying (about picture A) that: "the men are at the restaurant eating" is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Stg tbhkesfy grznqtx xx ule sgigy yc k qywzomiwx ey imiaety wjyobs nsmom
      xnpb." is correct.
3 Saying (about picture A) that: "The men are playing badmitton." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Qt lhndsef kknyzz patiu g ecpov rwdn liz lejowk jjtyq tifmp." is correct.
3 Saying (about picture A) that: "The person is showing affection towards the dog." is incorrect
```

---

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "H nnnvt lwnl poakr ljwgvyl na klc stxy hda i cqfhhd wqeo z bea tz axqhavi."
      is correct.
3 Saying (about picture A) that: "The young girl isn't holding any flowers." is incorrect
```

### F.13 SNLI - `shuffle-neg-prem`

```
1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "A bald man wearing black using a fan made of feathers, walking down the
      street." is incorrect.
3 Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky
      costume." is incorrect
```

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Children all dressed the same are standing outside a building." is incorrect.
3 Saying (about picture A) that: "the men are at the restaurant eating" is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "There is one man in the foreground with a hammer, another is in the background, possibly doing the same work as the man in the foreground." is incorrect.
3 Saying (about picture A) that: "The men are playing badmitton." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Man walking by a corner market with graffiti." is incorrect.
3 Saying (about picture A) that: "The person is showing affection towards the dog." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Two men by the lake one dressed in a penguin costume while his friend runs along side of him." is incorrect.
3 Saying (about picture A) that: "The young girl isn't holding any flowers." is **in**correct

### F.14 SNLI - `shuffle-pos-prem`

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "A bald man wearing black using a fan made of feathers, walking down the street." is correct.
3 Saying (about picture A) that: "A man is rocking out on his guitar, while wearing a funky costume." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Children all dressed the same are standing outside a building." is correct.
3 Saying (about picture A) that: "the men are at the restaurant eating" is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "There is one man in the foreground with a hammer, another is in the background, possibly doing the same work as the man in the foreground." is correct.
3 Saying (about picture A) that: "The men are playing badmitton." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Man walking by a corner market with graffiti." is correct.
3 Saying (about picture A) that: "The person is showing affection towards the dog." is **in**correct

---

1 You are looking at a picture (A) which is placed next to an unrelated picture (B).
2 Describing B as "Two men by the lake one dressed in a penguin costume while his friend runs along side of him." is correct.
3 Saying (about picture A) that: "The young girl isn't holding any flowers." is **in**correct