Propositional compositionality in neural language models

Jane Li, Abhinav Patil, Kyle Rawlins
Department of Cognitive Science
Johns Hopkins University
{sli213, aptil10, kgr}@jhu.edu

1 Introduction

One of the most fundamental representations in linguistic semantics is that of the proposition (McGrath and Frank, 2005), standardly taken as the carrier of truth-conditions. Recent work shows that some form of truth can be decoded from language models (Azaria and Mitchell, 2023; Li et al., 2023), and strikingly, that for some models, truth is even represented linearly in intermediate layers (Marks and Tegmark, 2024, GoT). We take this line of work a step further and argue that neural language models can use propositional representations compositionally (Janssen 2010; Pickel and Szabó 2025 a.o.), drawing from evidence of the behaviour of logical connectives: the *linear compositionality* hypothesis. Specifically, we show (a) that the truth values of individual conjuncts can be decoded independently of the truth value of a complex conjunction, and (b) that we can causally intervene on individual conjuncts in a way that affects the truth value of the whole.

2 Model and datasets

This work analyzes the sentence representations generated by Llama-3.1-8B (Grattafiori et al., 2024). The methods documented in this work generalize to any neural language model, but we take this as a starting point because previous work (GoT) uses models from the Llama series.

We constructed two datasets (4,000 sentences each) consisting of conjoined sentences in two knowledge domains: `cities` and `kinds`. Each dataset is balanced for the truth value of its atomic parts, as illustrated by Table 1. The atomic sentences for the `cities` dataset were recruited from GoT, whereas `kinds` was manually constructed.

3 Decoding logical expressions

Can we detect a linear encoding of truth – of a sentence's parts and globally – in the model's hidden representation of a sentence? The main finding of GoT suggests that this is possible for simple expressions using mass-mean probing, and here we demonstrate that it is also possible to linearly decode for the truth of the conjuncts and the global truth of a complex sentence.

Dataset	Example sentence
`cities`	The city of {Baltimore, Manila} is in the US, and the city of {Vancouver, Paris} is in Canada.
`kinds`	A {guppy, violin} is a fish, and a {stadium, rose} is a building.

Table 1: Example sentences from each of the datasets. Each item consists of four sentences, spanning all possible truth value configurations.

Mass-mean probing is a linear classification method with the probability function $p(\vec{x}) = \sigma(\theta \cdot \vec{x})$. Crucially θ is a difference vector, defined as the difference between the average positive and negative representation (in our case, model representations of true and false sentences). If the probe is found to be efficacious for classification, then we may hypothesize θ to be well-approximating a sentence's truth representation. This allows us to further test how θ affects a model's computation of truth downstream (§5).

At the final token of the sentence, we decode for the following truth-related properties: truth of the first conjunct, second conjunct, and the global truth value of the sentence. Individual probes are fitted for each of these properties, each layer, and each dataset, resulting in 3 (properties) $\times$ 32 (layers) $\times$ 2 (datasets) = 192 total probes. We also report the decoding of truth of the atomic sentences (conjuncts) that make up the dataset.

Each probe is fitted on 1,000 randomly sampled sentences from the dataset and evaluated on the held-out ones. The reported accuracies are the mean of 100 runs, each sampling a different subset. A probe is reported have above-chance signal if

the accuracy distribution is significantly greater than null accuracy distribution, measured from null probes where the labels for truth are shuffled in the fitting sentence set.

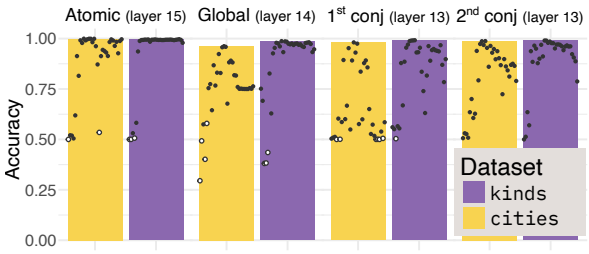

Figure 1: Decoding accuracies (of the layer with maximum signal) on held-out sentences on various truth properties of the sentence. Dots on the bar represent accuracies on each other layer, organized by layer depth (left-to-right). Hollow dots represent decoding accuracies that are at-chance.

We find that beyond early layers, mass-mean probes reliably detect various truth values of both datasets, though with substantial variation in their accuracy across layers and datasets. For each type of truth-related property, we find a layer that classifies held-out sentences at near-perfect accuracies (Fig. 1). PCA plots of the top PCs (Fig. 2) also reflect this separation.

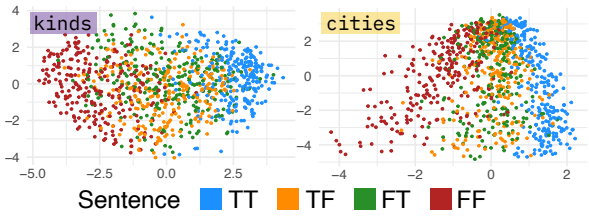

Figure 2: Top two PCs of sentence representations at layer 14, coloured by the truth value combinations of the sentence.

4 Whole state interventions

The previous experiment suggests that the model has some linear encoding of truth, but the linear compositionality hypothesis makes stronger predictions beyond successful decoding: we expect compositional representations of conjunct truth to be relevant for the computation of global truth. In this experiment, we examine a specific manifestation of this prediction: that computations over specific hidden states are responsible for taking conjunct truth as input and determining global truth via their outputs. This experiment demonstrates through hidden state replacement within minimal pairs that these states exist and are relatively localized.

4.1 Prompting for truth

First, we define a query that a model reliably correctly responds to. Here we use a mixture of example and question prompting, embedding each target sentence in the frame: `[ex. true sentence] Is this statement true? Yes. [ex. false sentence] Is this statement true? No. [target sentence] Is this statement true?`. We then measure the probabilities and rank of `Yes` or `No` as a continuation of the prompt. We find that this prompt is 91.0% accurate across all sentences we study, with some variation between datasets and conjunct truth value combinations (Fig. 3). We apply these measurements to truth-altered sentences to observe how the model's credence of the sentence fluctuates as a result of an alteration.

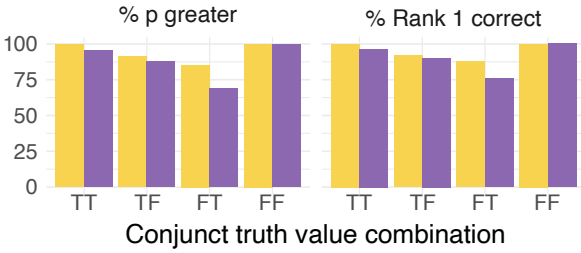

Figure 3: Prompting for `Yes` vs. `No` as a proxy for model-assigned credence to the sentence. (**A**; left) The % of items that correctly assign `Yes`/`No` as its more likely continuation over the alternative. (**B**; right) % of items that assign the correct token as the most likely response.

4.2 Hidden state swapping

We intervene on the truth value of the first conjunct of False-first conjunct sentences (FF and FT) by replacing a hidden state with the matching hidden state of its True-first counterpart. The main measurement is the degree of relative probability increase for `Yes` as a continuation to the prompt.[1] We conducted the swap over all layers and at five increments at the end of the first conjunct: `is`, `in/a`, `country/category`, `,` (comma), and `and`.

The linear compositionality hypothesis predicts that sentences that change in their hypothesized global truth value due to this alteration (i.e., FT → TT; Critical) should elicit greater probability increases relative to their counterparts that do not trigger a global truth value change (e.g., FF → TF; Control). We do not expect any effects of truth value alterations to arise in the first two timepoints

[1] Here we define relative probability (RelP) of `Yes` as $\frac{P(\text{Yes})}{P(\text{Yes})+P(\text{No})}$. The main measurement, relative probability increase, is the difference between RelP of the original sentence and RelP of the altered sentence.

(is, in/a) because the truth of the conjunct is indeterminate at that point of the sentence, thus acting as a methodological control.

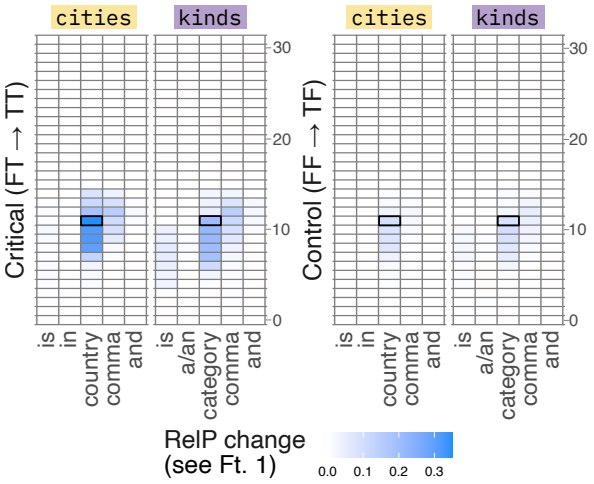

Figure 4: Relative probability change as a function of a state swap, where states vary by layer (x-axis) and token (y-axis). The highlighted grid indicates the hidden state we conduct follow-up analyses on.

We find that sentences in the Critical condition exhibit a greater increase in probability than the Control condition, though the Control condition also elicited positive probability change (Fig. 4). These results are mirrored with change in rank as the dependent variable. The results suggest that a model makes reference to the truth of the first conjunct when computing the truth or falsity of the larger sentence. This reference computation is relatively localizable, concentrated in mid-early layers (8-12) at the country or category tokens (for `cities` and `kinds` respectively). However, probability shifts in the Control condition also indicate that downstream measures such as output probabilities are influenced by conjunct truth, independent of its contribution to global truth. We also find stronger probability shifts for the `cities` dataset than the `kinds` dataset. The `kinds` dataset elicits probability increases in the control timepoints (specifically at the token `is`); we return to this in the discussion.

5 θ-intervention over target states

The findings so far set us up to test the linear compositionality hypothesis directly: if the linear encoding of truth θ (§3) well-approximates an abstract representation of conjunct truth in sentence representation space, then we expect intervening on a causally relevant hidden state by θ should impact model response to queries. At that target hidden state identified in §4, we computed θ for each dataset separately. Instead of replacing the hidden states, we now add or subtract (for False-first and True-first sentences respectively) the value θ from the hidden state. We then conduct the same types of measurements (relative probability change, rank change) as §4 on the altered sentences.

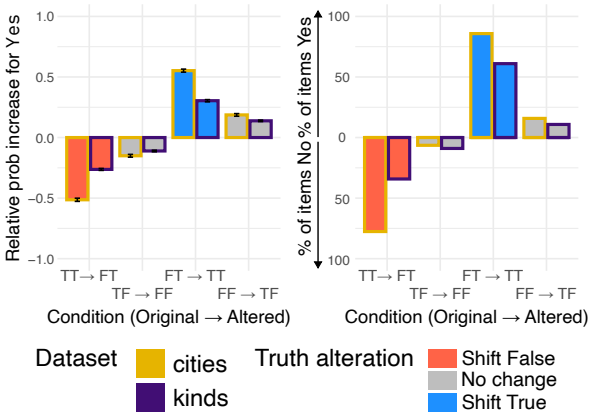

Figure 5: Results from the θ-intervention experiment. All results here are from alterations conducted on layer 11. (**A**; left) The degree of relative probability increase or decrease to the continuation `Yes`. Error bars indicate the 95% CI of the mean. (**B**; right) % of sentences that have changed their ranking to `Yes` (upward) or `No` (downward) as a result of the intervention.

We find that θ-intervention elicits similar results as whole state swapping: greater increases in the Critical condition, while observing some non-negligible increase in the Control condition (Fig. 5A, B). These results are also mirrored in the case of altering True-first sentences, where we observe an increase in the probability of `No` responses post-alteration. Once again, these results support the linear compositionality hypothesis while demonstrating that there is an independent contribution of conjunct truth to downstream tasks. Additionally, for both measures, the effect magnitude on probability (6.02%, $t = 21.035$, $p < 0.01$) are greater relative to whole-state intervention. This further strengthens the argument that the encoding of abstract conjunct truth, as approximated by θ, is the main driver of global truth computations.

6 Discussion

This set of experiments provide preliminary evidence for the linear compositionality hypothesis. We demonstrate that it is possible, for massively distributed systems without an explicit encoding of propositions, truth, or conjunction, to be well-approximated by the compositional, symbolic systems devised by semanticists and logicians. This

hypothesis ideally should extend to other truth-related computations such as other logical connectives, and should be extended to examine the computations that a model is carrying out, not just the input and output representations. In some preliminary work in disjunctive sentences, we find that the hypothesis falls apart. There are many linguistically-interesting hypotheses that may explain the difference between these logical operators, and we plan to explore this in future work.

6.1 Dataset-level variation

Across experiments, we observe differences between datasets: `kinds` shows considerably stronger decoding effects, but this relationship is reversed for the other experiments. Our working hypothesis is that this is due to the two datasets varying in the frequency to which the conjuncts are explicitly mentioned in some form in the training data. <City, country> pairs from `cities` dataset are likely to be stated in the training data frequently, or in the form of structured infoboxes, whereas commonsense facts from `kinds` may be less frequent in the training data even though they represent widely agreed upon knowledge (Gordon and Van Durme, 2013). However, this hypothesis does not explain why these differences cause performance differences between tasks.

6.2 Truth, uncertainty, and plausibility

In this work, we employed a "neat" definition of truth and used examples that are easily categorizable and verifiable under this notion of truth, as our goal was to study the computations of conjunctive sentences. As follow-up work, it would be important to expand to various more graded effects that are known to impact both human and model truth-value judgements, such as uncertainty (Hu et al., 2025a) and sentence plausibility (Hu et al., 2025b).

6.3 Limitations

We highlight two major limitations of this work. Following prior work, we use prompting as a measure of model 'credence'; in doing so we find fluctuations in accuracies across different truth value combinations (Fig. 3), potentially confounding any intervention experiments that also rely on comparisons across conditions. More investigation of the relationship between internal model states and prompting behavior is crucial. A second shortcoming concerns the treatment of token-level interventions, which can sometimes mismatch linguistic units. It remains unclear if failure of intervention at a single token occurs because of a failure to represent truth, or an unexpected positional representation of truth relative to the token sequence.

References

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, AKBC '13, pages 25–30, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Jennifer Hu, Michael A. Lepori, and Michael Franke. 2025a. Signatures of human-like processing in Transformer forward passes. *arXiv preprint*. ArXiv:2504.14107 [cs.AI].

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025b. Shades of zero: Distinguishing impossibility from inconceivability. *Journal of Memory and Language*, 143:104640.

Theo M.V. Janssen. 2010. Compositionality. In Johan van Benthem and Alice G. B. ter Meulen, editors, *Handbook of Logic and Linguistics*. Elsevier.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Samuel Marks and Max Tegmark. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*.

Matthew McGrath and Devin Frank. 2005. Propositions. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Bryan Pickel and Zoltán Gendler Szabó. 2025. Compositionality. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2025 edition. Metaphysics Research Lab, Stanford University.