Interpretability and Controllability of Backpack LMs

Stanford CS224N Custom Project, mentored by John Hewitt

Sarah Chen
Department of Computer Science
Stanford University
sachen@stanford.edu

Tae Kyu Kim
Department of Computer Science
Stanford University
taekyu@stanford.edu

Abstract

Backpack language models generate predictions that are log-linear in non-contextual "sense vector" representations of the input tokens. This property enables targeted interventions that can operate consistently across contexts. To lay the groundwork for future control techniques, we perform interpretability analysis on each component of Backpack language models. We qualitatively explore how sense vectors encode grammatical and semantic information, and we characterize how contextualization weights can route responsibility for next-word prediction to specific token senses. We also formulate contrastive saliency scores to investigate how sense vectors and contextualization weights jointly contribute toward predicting a target token over a specific foil token, and we explore the patterns that emerge across arbitrary sets of context sequences via clustering analysis. In the direction of control, we demonstrate that simple interventions on specific sense vectors can interpretably influence verb conjugation, and we propose an updated mechanism for topic control.

1 Introduction

Large language models (LMs) are difficult to interpret and control. Since the Transformer LM is a large monolithic function of the input sequences, any intervention on the input or the word embeddings has complex, non-linear effects that depend on the context. Moreover, since the word embeddings are contextual, there are no guarantees that past analysis of word embeddings in Transformer LMs will generalize globally across all contexts. In cases such as debiasing, we want simple, direct interventions on the language model that reliably reduces bias across all contexts.

Backpack LMs are a recent neural architecture for which predictions are log-linear in a set of non-contextual sense vector representations (Anonymous, 2023). For each vocabulary token $\mathbf{x} \in \mathcal{V}$, we use a sense function C: $\mathcal{V} \to \mathbb{R}^{d \times k}$ to obtain k sense vectors $C(\mathbf{x})_1, \ldots, C(\mathbf{x})_k \in \mathbb{R}^d$. The sense function is parameterized via an embedding matrix $E \in \mathbb{R}^{d \times |\mathcal{V}|}$ and feed-forward network FF: $\mathbb{R}^d \to \mathbb{R}^{d \times k}$ such that $C(\mathbf{x}) = \text{FF}(E\mathbf{x})$. Then, given an input sequence $\mathbf{x}_{1:n}$, we calculate contextualization weights using a standard Transformer $\alpha = A(\mathbf{x}_{1:n}) \in \mathbb{R}^{k \times n \times n}$, and we log-linearly combine the sense vectors according to the contextualization weights to form the final prediction:

$$\mathbf{o_j} = \sum_{i=1}^{n} \sum_{l=1}^{k} \alpha_{lij} C(E\mathbf{x_j})_l, \tag{1}$$

$$p(\mathbf{y}|\mathbf{o_n}) = \text{softmax}(E^\top \mathbf{o}_n). \tag{2}$$

The non-contextual property of tokens' sense vectors provides the opportunity for interventions that operate consistently and interpretably across contexts.

This work explores two primary questions: (1) How do Backpack LMs effectively perform next token prediction? (2) How can this explainability work lead toward interpretable control techniques?

We perform explainability analysis on each component of the Backpack LM architecture. In section 3, we visually examine the grammatical associations embedded in sense vectors of function words such as "and". In section 4, we analyze the role that contextualization weights play in influencing verb conjugation and pronoun agreement.

In section 5, we formalize saliency scores that quantify the joint contribution of sense vectors and contextualization weights to model predictions. We also formulate a contrastive saliency score that compares how much each token sense contributes to the prediction of a target token rather than a foil token. These scores allow us to isolate the individual sense vectors that are relevant to the relationship between target and foil, and reweighting these token senses takes a step toward interpretable control.

In section 6, we follow Yin and Neubig (2022) and use contrastive saliency scores to cluster foil tokens that rely on context across a corpus in similar ways.

Through these analyses, we provide insight into the prediction mechanisms of Backpack LMs, which is a prerequisite to designing more complex interventions on token sense vectors. One approach toward control decomposes the problem into first identifying relevant sense vectors and then applying targeted modifications to achieve a desired result. In this vein, Anonymous (2023) perform interventions by reweighting specific sense vectors, including sense 10 for gender debiasing. Our interpretability methods can facilitate the identification of relevant token tenses, e.g. ranking the relevance of sense vectors via contrastive scores based a set of minimal pairs. Anonymous (2023) also perform experiments that heuristically reweight sense vectors in the setting of topic control, and we normalize the topic re-weighting equation to improve topic-controlled generation in section 7.

All experiments in this paper use a pre-trained 170M parameter backpack model with $k = 16$ senses.

2 Related Work

Explainability. Previous work in explainability has studied the behavior of Transformer LMs from various angles. For example, Hewitt and Manning (2019) and Eisape et al. (2022) study how LMs internally represent syntactic structure. Other methods focus on using saliency scores to explain why models make specific predictions. In particular, Yin and Neubig (2022) establishes methods to interpret LMs via contrastive explanations, which inspires our formulation of contrastive saliency scores for Backpack LMs. Broadly, we extend this line of work to the analysis of Backpack LMs.

Control. Previous work has proposed various methods to intervene on LM predictions. For example, on the level of syntax, explainability analyses often verify their hypotheses by demonstrating that they can modulate model behavior through the mechanisms they identify (Eisape et al., 2022). Additionally, Meng et al. (2022) proposes a causal tracing method for knowledge localization and editing within GPT. However, localization and control of Transformer LMs is a complex challenge, and Hase et al. (2023) demonstrates that causal tracing cannot necessarily determine optimal layers for editing – suggesting that better mechanistic understanding of LLMs may not immediately translate into better mechanisms of control.

The Backpack LM architecture may provide a setting that enables more consistent and interpretable control. Anonymous (2023) explores methods for intervention on Backpack LMs, and our work further lays the groundwork to explore more complex methods of control.

3 Sense vector association visualization

3.1 Methodology

Qualitative inspection suggests that sense vectors for common function words may serve as repositories of grammatical information (Table 1).

We extend the sense visualization analysis in Anonymous (2023). For a token $\mathbf{x}$, we plot the distributions of logits $EC(\mathbf{x})_\ell \in \mathbb{R}^{|\mathcal{V}|}$ for each sense l. The logits $EC(\mathbf{x})_\ell$ measure the associations that the Backpack LM learned between $\mathbf{x}$ and the vocabulary, which have meaningful relative values (i.e. if token y_1 has a higher logit than y_2, then sense vector ℓ of $\mathbf{x}$ will contribute toward upweighting y_1 over y_2 in all contexts). We obtain a histogram of logits for each sense.

" the" sense 5 (-)	" the" sense 2 (-)	"the" sense 9 (-)	" and" sense 9 (-)
were	sightings	few	substituted
Scroll	rises	Less	consulted
concede	clashes	better	drew
admit	phases	worst	worked
querade	uphe	strong	protested
were	foregoing	Few	danced
have	celebrations	another	nodded
remain	collapses	fewer	added
depict	massacres	only	apologised
demonstrate	sighting	unknown	tasted

Table 1: Top 10 vocabulary tokens that are negatively associated with specific token sense vectors. Sense 5 of " the" is strongly negatively associated with plural verbs. Sense 2 is negatively associated with various tokens that appear to contain relevant singular-plural information. Sense 9 is negatively associated with adjectives. Sense 9 of " and" is strongly negatively associated with past tense verbs.

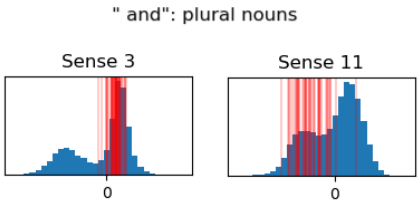

Figure 1: Histograms of logits corresponding to sense associations of " and" for senses 3 and 11. The red lines correspond to logits of 50 plural nouns.

To study the grammatical structures learned by the Backpack LM, we also visualize the logits of representative words from grammatical groups such as adjectives, plural/singular nouns, and past tense verbs. The logits are visually overlaid on each histogram. If logits for words in a given grammatical group are concentrated in one area of the distribution, that sense vector may have grammatical utility within the model (e.g. used to upweight singular/plural verbs).

3.2 Results

Sense 3 and 11 of " and" demonstrate structured assocations with plural nouns (Figure 1). See Appendix Figures 5 and 6 for further results.

We observe that content words such as " sports" and " man" have Gaussian logit distributions while function words like " and" and " the" have several senses with heavily skewed or multimodal logit distributions.

Additionally, for function words, we observe that the 50 highlighted logits concentrate around modes and skews of the logit distributions: the first mode of sense 11 of " and"; the left-skew of sense 9 of " and"; the left-skew of sense 13 of " a". Thus, we hypothesize that the modes and skews of multimodal logit distributions correspond to grammatical associations of the token senses.

4 Contextualization weight analysis

4.1 Methodology

Visualization. We examine whether similar sentences elicit similar Backpack contextualization weights. We consider sentences extracted from the M&L dataset, which is a dataset of minimally different pairs of grammatically correct and incorrect sentences related to different grammatical rules, e.g. pronoun agreement (Marvin and Linzen, 2018). In this work, to maintain consistent interaction with the GPT-2 tokenizer, we only consider examples for which target verbs' singular and plural forms are both a single token. We also omit "is" and "are" as special cases.

Here, we consider sentences from the `vp_coord` category, which focuses on verb conjugation. We extract a set of singular sentences and a set of plural sentences, which maintain identical syntactic structure: "the `<subject>` `<verb1>` and `<verb2>`". For each set, we calculate the average contextualization weights used to predict the second verb in the sequence.

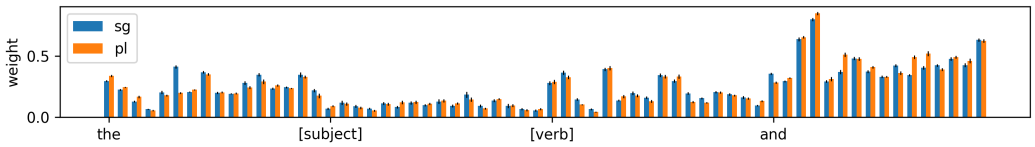

Figure 2: Average contextualization weights for singular (blue) and plural (orange) `vp_coord` sentences. Error bars denote standard deviation. We flatten the contextualization weights so that each context token is represented by $k = 16$ bars, as indicated by the ticks on the x-axis.

	simple agrmt	vp coord	simple reflexives	sent comp (pl)	sent comp (sg)
sg transfer	1 / 20	13 / 20	1 / 140	3 / 120	1 / 120
pl transfer	7 / 20	20 / 20	14 / 140	20 / 120	11 / 120

Table 2: Effect of transferring contextualization weights from a singular sentence to a plural sentence (row 1) and the converse (row 2) for various M&L categories (columns). The table shows the number of samples for which transferring weights from a sg/pl setting causes the model to shift to a sg/pl verb, despite a pl/sg subject. We only report transfer in cases when the model was originally correct.

Transfer. We investigate the role of contextualization weights in model predictions by transferring contextualization weights from singular to plural sentences and vice versa. This transfer also allows us to quantitatively evaluate the variation between sets of contextualization weights. Specifically, we consider pairs of grammatically valid, minimally different singular and plural sentences from the M&L dataset, drawn from subcategories related to verb conjugation and pronoun agreement (e.g. "the author laughs" vs. "the authors laugh"). We transfer weights between the sentences in each pair.

4.2 Results

Visualization. Shared contextualization weight profiles emerge for sentences with shared syntactic structure. Contextualization weights for singular and plural sentences in the `vp_coord` category have low standard deviation (Figure 2). [1] In addition, the weights for singular and plural sentences have slightly different profiles, with the variation primarily located in sense 5 of "the" (Figure 2).

Transfer. Transferring the contextualization weights from plural sentences to singular sentences can lead the model to predict the plural foil verb with higher likelihood than the ground truth singular verb – even though the actual subject is singular (e.g. "the author laugh") (Table 2). This result suggests that the contextualization weights for singular and plural sentences carry significant differences, and these differences play a strong role in verb conjugation.

Our results suggest that a subject's sense vectors are not necessarily the primary determiner of verb conjugation, given that we observe changes even when that token is fixed. Instead, contextualization weights route number responsibility to the token senses of adjacent function words, which are present across all examples. We further explore this result in section 5.2.

However, singular transfer is much less prevalent than plural transfer, suggesting that the subject token still carries significant verb conjugation information (Table 2). Future work could investigate the mechanism for singular conjugation by evaluating the effects of substituting each sense vector of the plural subject with that of the singular subject. [2]

5 Saliency scores and grammatical interventions

5.1 Methodology

To interpret Backpack LM predictions, we must consider (1) how token sense vectors upweight different output tokens and (2) the contextualization weights with which those vectors are combined.

[1] However, we caution that raw contextualization weights may be difficult to interpret. Even small variations may have an impact on model predictions in ways that are not immediately evident.

[2] Note: the baseline model generally has stronger performance in the plural setting than in the singular setting. An overall bias toward predicting plural verbs could in part explain why plural transfer is more prevalent.

We combine these considerations into a saliency score that measures how much each sense vector of each context token contributes to the prediction of a target output token. [3]

Saliency score formulation. When performing next token prediction at index t, we define saliency scores $\mathbf{s} \in \mathbb{R}^{(t-1)\times k}$ such that for token at index $j < t$ and sense vector $\ell < k$, the saliency for a particular target token $\mathbf{x}_t$ is

$$\mathbf{s}(\mathbf{x}_t)_{j\ell} = \left(\alpha_{\ell tj} E(C(\mathbf{x}_j)_\ell)\right)_{\mathbf{x}_t}. \tag{3}$$

$C(\mathbf{x}_j)_\ell$ is the noncontextual vector for sense ℓ of token j, and $E(C(\mathbf{x}_j)_\ell) \in \mathbb{R}^{|\mathcal{V}|}$ is the corresponding distribution of scores across the vocabulary. We scale this by $\alpha_{\ell tj}$, the contextualization weight for sense ℓ of token j, to measure how much that token sense contributes to the logit for each token in the vocabulary. We retrieve the value for token $\mathbf{x}_t$ to quantify the contribution to the target token.

Contrastive saliency score formulation. Contrastive saliency scores explain why a model predicted a target token $\mathbf{x}_t$ rather than a contrastive foil token $\mathbf{x}_f$. We formalize contrastive saliency scores by extending the methodology of Yin and Neubig (2022) to Backpack LMs. In particular, we quantify what token senses are responsible for the prediction of a target token over a foil token:

$$\mathbf{s}'(\mathbf{x}_t, \mathbf{x}_f) = \mathbf{s}(\mathbf{x}_t) - \mathbf{s}(\mathbf{x}_f). \tag{4}$$

This score allows us to pinpoint sources of variation between a target and foil token (e.g. sense 10 of "nurse" is a source of gender bias, as reported in Anonymous (2023) (Appendix Figure 7).

Notably, we can extend this approach beyond single examples. Given a set of minimal pairs, we can identify what token senses are responsible for predicting specified target tokens over foils by accumulating contrastive saliency scores across all pairs. The token senses with large scores reveal what token senses encode the information that distinguishes targets from foils. This approach works on any number of sense vectors. For example, we find that sense 44 for the $k = 64$ Backpack LM model is responsible for gender bias in a manner analogous to sense 10 in $k = 16$.

Intervention. After identifying what token senses are relevant to a given prediction via contrastive saliency scores, we can stage intervention techniques to influence those predictions. Following Anonymous (2023), we upweight or downweight particular sense vectors of particular tokens and quantify the effect on downstream generation. We compare reweighted performance to the baseline performance of the unmodified model. Specifically, we quantify performance by evaluating whether the model can predict the correct verb with higher probability than the incorrect verb, directly comparing the logits at the indices of the verb token.

5.2 Results

Saliency inspection. We consider categories of M&L that test subject-verb agreement for singular and plural verbs: `vp_coord` and `long_vp_coord`. Across the M&L data, sense 5 of "the" has high contrastive saliency during verb conjugation. In the example, "the authors laugh and smile", although "authors" is what necessitates the plural verb, the sense vectors of "the" have the largest contrastive saliency (Figure 3).[4] This effect originates in the underlying sense vectors: sense 5 of "the" has positive association with singular verbs ("smile") and large negative association with plural verbs ("smiles"), making the difference between their saliency scores large (Table 1). Therefore, weighting this token sense highly should increase the likelihood of predicting a singular verb.

Intervention. Reweighting sense 5 of "the" predictably influences verb conjugation. In Figure 4, we multiply the token sense vector by $\lambda \in [-0.5, 2]$) and re-evaluate on M&L. We observe that, as expected, upweighting leads to predicting singular verbs while downweighting leads to plural verbs. This effect is strongly visible on the simpler `vp_coord` dataset and also present in `long_vp_coord`.

Perturbing other senses of "the" also influences conjugation, but to a weaker degree (Appendix for Figure 8). We also found that reweighting sense 10 of " and" leads to strong trends in `vp_coord`, but

[3] Backpacks' contextualization weights provide a natural notion of saliency, but they disproportionately weight the directly preceding token and offer no clear path toward control.

[4] The lack of saliency for the subject is not unexpected. Saliency scores operate on the level of tokens' *sense vectors*. They do not capture how much each *token* is "responsible" for a prediction – which would require deeper analysis of the Transformer that sets contextualization weights.

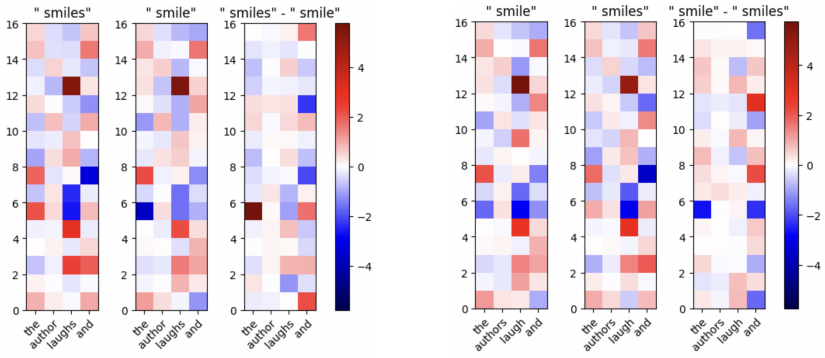

Figure 3: Saliency scores for target words “smile” and “smiles” in the sentences, “the author laughs and X” (left) and “the authors laugh and X” (right). The rightmost subplot in each panel shows the contrastive saliency score. The token senses that contain information related to grammatical number have contrastive saliency scores with large magnitude. The saliency scores on sense 5 of “the” is large and opposite between the two sentences, signaling its grammatical function.

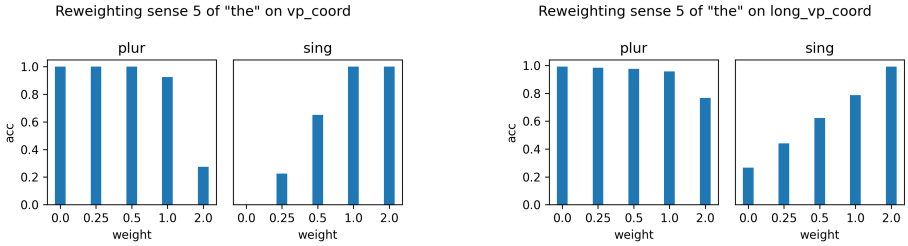

Figure 4: Effect of reweighting sense 5 of “the” on performance on M&L `vp_coord` and `long_vp_coord`. (The bar for weight 1 represents the performance of the unmodified model.)

reweighting the subject of each example leads to only very weak trends in sense 0 and 6. (Plots and further discussion omitted due to space constraints.)

6 Clustering foils

6.1 Methodology

We perform a foil clustering analysis that follows the methodology of Yin and Neubig (2022). Contrastive explanations demonstrate what evidence models use to disambiguate between target and foil tokens. We consider various targets of interest (Table 3) and use the 1,000 most frequent tokens in WikiText-103 (a dataset extracted from verified Wikipedia articles) as foils (Merity et al., 2016).

Consider a target $\mathbf{y}_t$. We sample 500 sentences containing that target from WikiText-103. For each sentence, we calculate the contrastive saliency scores between the target $\mathbf{y}_t$ and each foil $\mathbf{y}_f$, yielding $s'(\mathbf{y}_t, \mathbf{y}_f) \in \mathbb{R}^k \times \mathbb{R}^{d_i}$ where k is the number of sense vectors and d_i is the length of sentence i. Then, for each $\mathbf{y}_f$, we concatenate the saliency score vectors for all sentences to generate an explanation vector in $\mathbb{R}^k \times \mathbb{R}^{\sum_i d_i}$. We separate this into k per-sense explanation vectors of length $\sum_i d_i$, and we consider an additional vector by taking the sum across all senses. These explanation vectors highlight the tokens used to disambiguate the target from the foil in various contexts, per sense.

For each sense $\ell < k$, we apply k-means clustering on the explanation vectors for all foils y_f, using 100 clusters. Foils for which the model uses similar evidence to disambiguate cluster in groups, revealing patterns of context-based grammatical and semantic similarity.

We compare the clusters that emerge to the clusters yielded by analyzing raw foil token sense vectors. We also extract the 5 nearest neighbors for each token sense vector.

target	foil	sense	cluster (truncated)
go	runs	all	runs, was, is, has, works, does
go	went	all	went, were, had, did, began, took, wrote, won, came, died, started, told
man	men	all	men, people, children, women
his	their	all	their, her, its
five	one	all	one, 18, 17, 8, 15, 16, 12, 11, 13, 14
girl	person	9	person, John, James, William, George, David, Michael, Henry, Robert, Paul, Charles, Thomas, Smith, Peter
black	white	0	white, been, American, British, German, French, English, man, Japanese, Australian, black, Black, mother European, director, White, Red, International, Star
black	white	1	white, short, right, black, Black, White, Red

Table 3: We focus on the sum across senses (denoted as "all") that operates on the token level. However, we include some sense-level clusters, e.g. for " girl" to demonstrate that senses pick up on individual patterns that are not evident in the summed analysis – specialization.

6.2 Results

Characterizing contrastive explanations across various contexts allows us to better understand how models disambiguate between targets and foils. We recover clusters roughly consistent with selected findings from Yin and Neubig (2022). For example, the foil "she" is distinguished from the target "he" by similar contextual evidence as other female-gendered pronouns. Distinguishing "runs" from "go" requires similar contextual evidence as other singular verbs. Additional examples can be found in Table 3, and more results with comparison to baseline cluster approaches are in Appendix Table 5.

Foil clusters also capture per-sense specialization for words with multiple meanings. By grouping foils based on the context tokens that distinguish them from a target, our clustering method recovers a context-dependent notion of word sense that is more aligned with traditional linguistic word sense than the raw senses of non-contextual sense vectors. For example, for the target "black", the foil "white" falls into different clusters for senses 0 and 1, which relate to different aspects of the target token's meaning: race/ethnicity descriptor vs. visual color descriptor. In sense 0, "white" clusters with "European" because their disambiguations involve similar context usage even though their raw sense vectors do not demonstrate a significant degree of similarity.

7 Topic control

7.1 Methodology

We also revise the topic-controlled generation method proposed in Anonymous (2023). Upon choosing the re-weighting factors $\delta_{lij} \geq 1$ according to the logits $EC(\mathbf{x}_{topic})$ for the topic $\mathbf{x}_{topic}$, they update the Backpack equation to

$$\mathbf{o}_i = \sum_{j=1}^{n} \sum_{l=1}^{k} \alpha_{lij} \delta_{lij} C(\mathbf{x}_j)_l. \tag{5}$$

However, we observe that setting $\delta_{lij} \to \infty$ turns $p(\mathbf{y}|\mathbf{o}_n)$ into a 1-hot vector, instead of the desired distribution $EC(x_j)_l$. This flaw suggests that Equation (5) is missing a normalization factor. Thus, we repeat the experiments in Anonymous (2023) with the following normalized Backpack equation

$$\mathbf{o}_i = \frac{1}{\sqrt{\frac{1}{nk}\sum_{j,l} \delta_{lij}^2}} \cdot \sum_{j=1}^{n} \sum_{l=1}^{k} \alpha_{lij} \delta_{lij} C(\mathbf{x}_j)_l. \tag{6}$$

The normalizing constant is chosen such that in the baseline case when $\delta_{lij} = 1$ for all l, i, j, Equation (6) reduces to the original Backpack equation.

7.2 Results

We repeat the experiments for topic control in Anonymous (2023) with the normalized Backpack equation (6). We keep the same values of δ with annealing, then apply normalization before generating the predictions.

We generate 500 strings for each of 21 categories in the topic classifier of Antypas et al. (2022) and calculate the percentage of strings that the classifier assigns the correct topic label with at least 0.5

Method	Sem Acc ↑	Toks-in-vocab ↓
Backpack		
Unchanged	7.4%	0.0%
Old_{+1}	12.1%	0.2%
Old_{+2}	24.3%	1.5%
Old_{+3}	35.3%	3.5%
Ours_{+1}	13.9%	0.2%
Ours_{+2}	30.8%	2.4%
Ours_{+3}	42.8%	4.8%

Table 4: Semantic accuracy and frequency of the topic words in the generations for each strength level across the topics in Antypas et al. (2022).

confidence. In addition, we calculate the frequency at which the generated text contains the topic token since we want the language model to discuss the topic without mentioning the topic explicitly. Table 4 shows the performance of the updated topic control method.[5] Although the semantic accuracy significantly increases with normalization, the frequency of topic tokens also increases.

8 Discussion

Our experiments demonstrate that Backpack LMs use a complex combination of sense vectors and contextualization weights to control grammar. Based on input context, Backpack LMs may learn to upweight and downweight specific sense vectors of function words to produce grammatically valid outputs – a hypothesis we tested by evaluating contextualization weight transfer as well as by intervening on the salient function token sense.

Unlike saliency analysis on Transformer LMs, saliency scores on Backpacks LMs provide a direct path to influencing output. As such, our experiments also motivate various lines of future work.

Further foil cluster analysis. Foils that cluster together rely on contexts in a similar fashion, so updating a token sense relevant to one token in a cluster may also impact the other tokens in a cluster (e.g. modifying a certain sense of MacBook will likely affect not only Apple and HP but also any other tokens found in the same foil cluster). Therefore, the foil clustering analysis may be useful to better characterize the impact of proposed interventions.

More generally, studying the mean explanation vector for each foil cluster would allow us to characterize what evidence the model actually uses to make specific distinctions, following how Yin and Neubig (2022) qualitatively analyze what senses of what context tokens are important for specific disambiguations. For Backpack LMs, modifying those relevant token senses may be a path toward effective control.

Contextualization weights and tool use. We can view the trained contextualization Transformer as a model that has learned how to use sense vectors as specialized "tools". High contextualization weights can activate these tools in the relevant settings. For example, Anonymous (2023) found that sense 3 of a given token is often associated with common next word pieces, and we qualitatively observe that the Transformer often learns to activate sense 3 of the immediately preceding token during next token prediction, e.g. "and" in Figure 2. One could examine training a contextualization transformer on a set of frozen sense vectors or vice versa to deepen our understanding of this relationship. A potential experiment could aim to reconstruct a sense like sense 3 that fires for immediately preceding tokens.

Additionally, if we simply subtract the contextualization weights for singular and plural sentences (e.g. "the author smiles", "the authors smile"), we can observe what token senses are being activated differently in different contexts. Future work could look more closely at what token senses the model weights differently in different contexts.

Sense vector relationships. Different sense vectors interface in complex ways that are not directly interpretable. For example, different senses may contribute positively or negatively to the logit for a specific output token, and the manner in which they "cancel out" could be another interesting line of inquiry.

[5]We were unable to generate MAUVE scores due to technical and time constraints.

References

Anonymous. 2023. Backpack language models. *ACL submission*.

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. 2022. Twitter topic classification. *arXiv preprint arXiv:2209.09824*.

Tiwalayo Eisape, Vineet Gangireddy, Roger P Levy, and Yoon Kim. 2022. Probing for incremental parse states in autoregressive language models. *arXiv preprint arXiv:2211.09748*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *arXiv preprint arXiv:2301.04213*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*.

A Appendix

target	foil	sense		
he	she	all	contrastive cluster	her, she
			sense cluster	he, He, she, him, She, John, himself, George, David, Paul, Smith, Peter
			sense neighbors	she, She, he, He, him, they, them, They
man	men	all	contrastive cluster	people, men, children, women
			sense cluster	men, man, children, women
			sense neighbors	men, man, women, people, them, him,ians, others
go	runs	all	contrastive cluster	was, is, has, works, does, runs
			sense cluster	work, released, wrote, built, release, written, published, run, developed, writing, works, worked, working runs, designed, available
			sense neighbors	runs, run,ing,ed,s, with, on,es
go	went	all	contrastive cluster	were, had, did, began, took, wrote, won, came, went, died, started, told
			sense cluster	came, went, go, going, come
			sense neighbors	went, go, going, came, was, were,ed, been
black	white	0	contrastive cluster	been, American, British, German, French, English, man, Japanese, Australian, black, white, Black, mother European, director, White, Red, International, Star
			sense cluster	black, white, Black, White
			sense neighbors	white, White, black, Black,-, (,2, 0
black	white	1	contrastive cluster	short, right, black, white, Black, White, Red
			sense cluster	black, white, Black, White
			sense neighbors	white, black, White, Black, red,-, (,man
his	their	all	contrastive cluster	their, her, its
			sense cluster	the,s, The, his, their, her, its, His, my, Her, My, whose
			sense neighbors	their, its, his, whose, my, her, the, His
girl	person	9	contrastive cluster	John, James, William, George, David, Michael, Henry, Robert, Paul, Charles, Thomas, Smith, Peter, person
			sense cluster	people,ers, members, member, population, others, crew, President,ors, director, church, person,ians
			sense neighbors	person, people, man, others, men,man, member, himself
five	one	all	contrastive cluster	one, 18, 17, 8, 15, 16, 12, 11, 13, 14
			sense cluster	a, an, first, one, this, A,th, This, second, single, another, third, One, fourth,nd, First,one
			sense neighbors	one, One, another,one, each, this, two, that
going	being	all	contrastive cluster	being, following, having, using, making, playing, leading, working, saying, opening, taking, beginning
			sense cluster	was, is, were, had, be, are, have, been, has, being, became, having, become,re, remained, appeared, get
			sense neighbors	being, be, is, been, are, was, were, have

Table 5: Further foil clustering results with comparisons to clustering raw sense vectors for the listed foils, as well as extracting the five nearest neighbors.

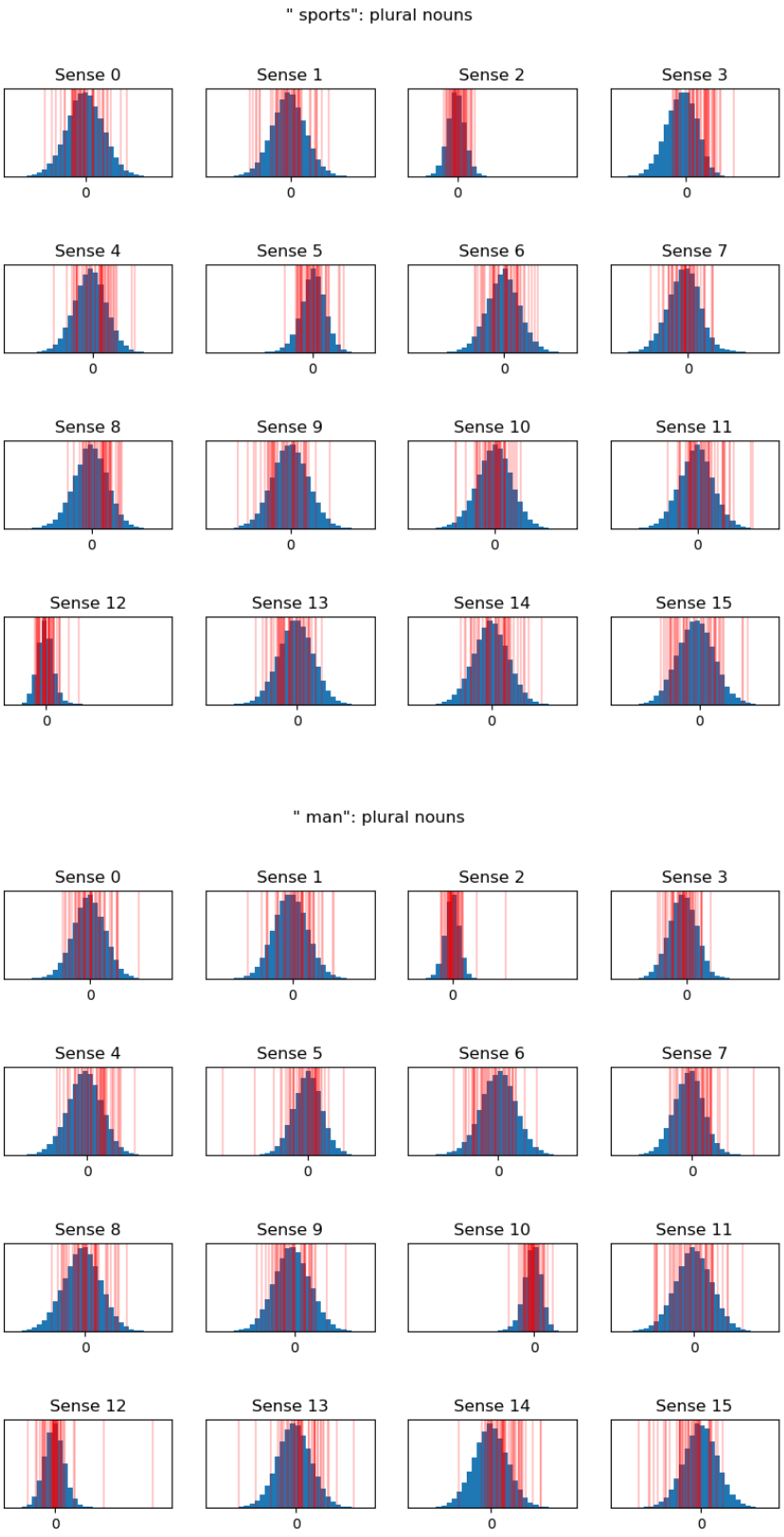

Figure 5: Histograms of logits corresponding to sense associations of content words.

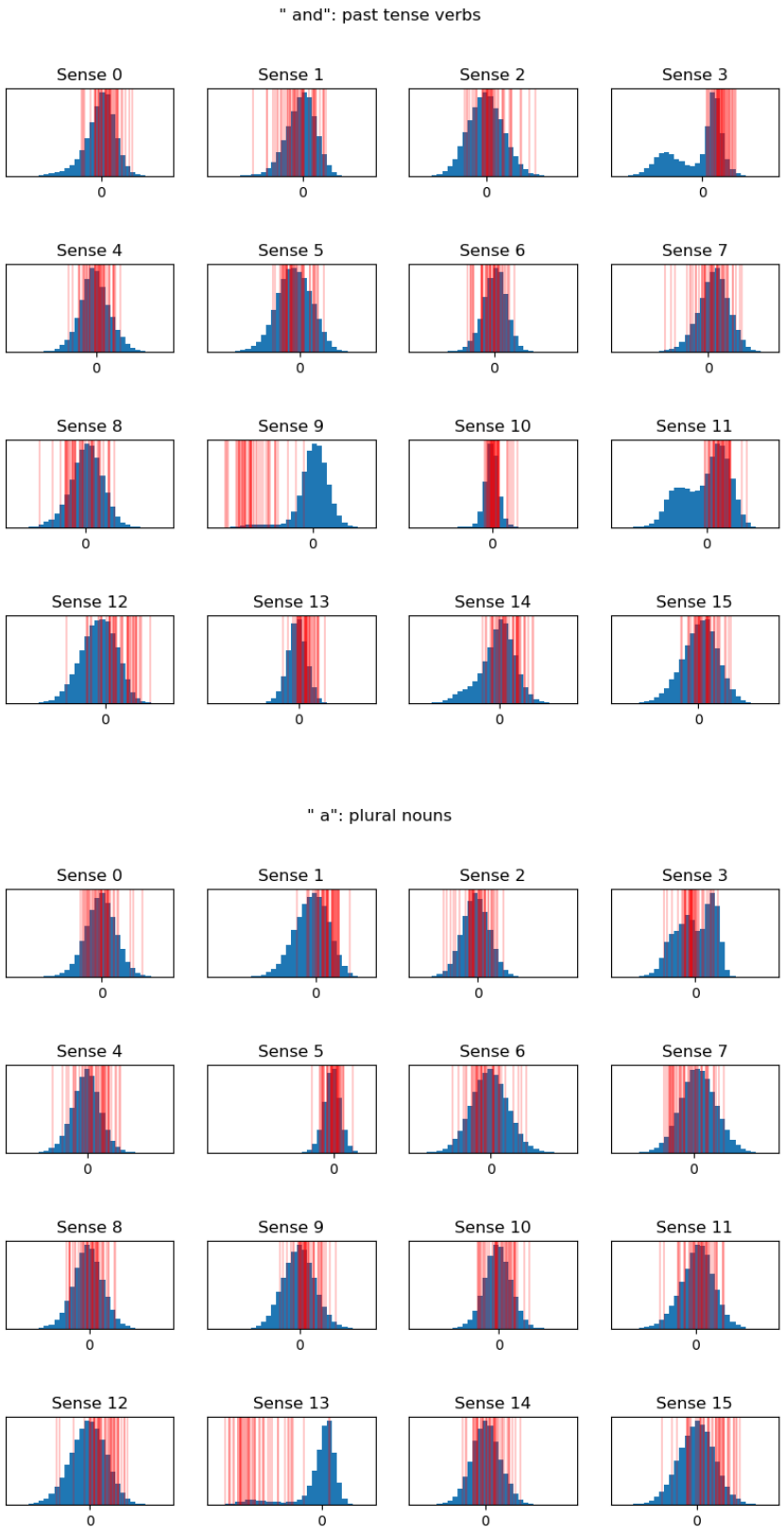

Figure 6: Histograms of logits corresponding to sense associations of function words.

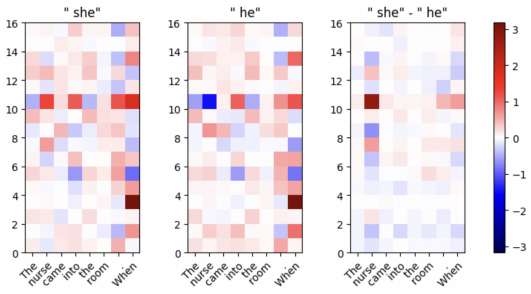

Figure 7: Saliency scores for "The nurse came into the room. When X", where "X" is "she" (left) vs. "he" (middle). The rightmost panel shows the contrastive score between the target "she" and contrastive target "he". Observe the high contrastive saliency scores on the word "nurse". Meanwhile, the non-contrastive method is prone to highlight the directly preceding token. Note: the example sentence is from Anonymous (2023).

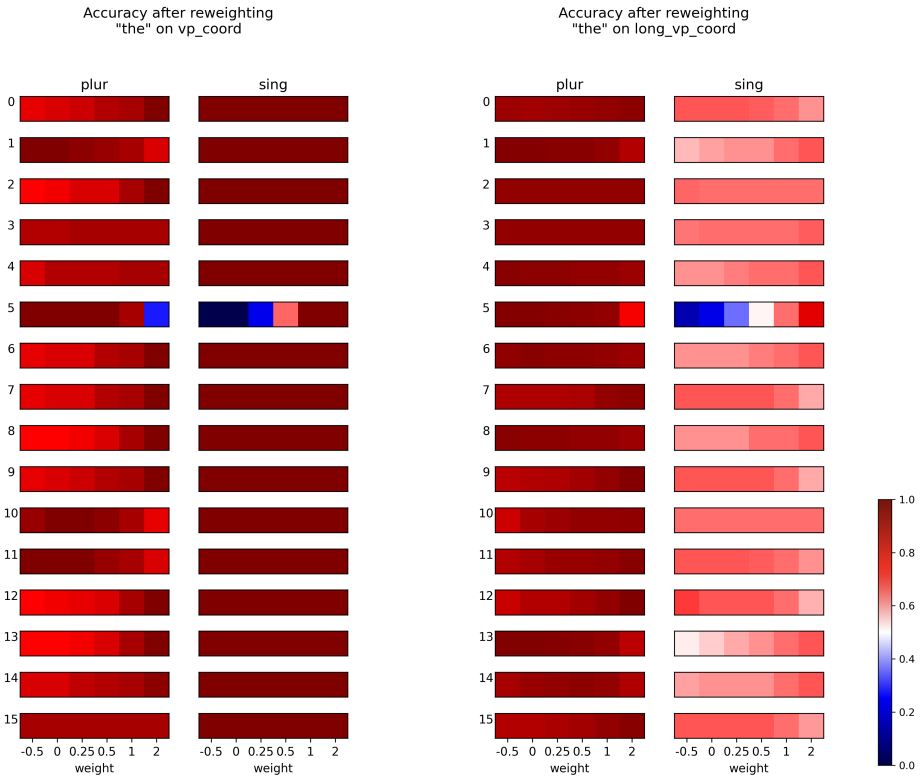

Figure 8: Effect of reweighting all senses of "the" on verb conjugation performance in `vp_coord`(left) and `long_vp_coord` (right). The sense to reweight varies on the y-axis. Accuracy is displayed for the ground truth plural setting and ground truth singular setting in the left and right of each panel.