IDENTIFYING TOKEN AND TASK-RELEVANT REPRESENTATIONAL SUBSPACES IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how language models predict particular tokens and perform specific tasks requires identifying representational subspaces that selectively support these behaviors across contexts. We develop a unified framework based on gradients of token scores and behavioral losses with respect to hidden states. Averaging gradient outer products yields a sensitivity matrix whose leading eigenvectors capture directions of greatest local sensitivity. At the token level, we separate the context-averaged gradient direction from a residual subspace that captures context-dependent sensitivity orthogonal to it. We extend this framework to task selectivity by contrasting sensitivity matrices for target and comparison tasks, yielding Contrastive Task Directions. We find that token-relevant residual subspaces in Qwen and Llama generalize across independent contexts, with sensitivity increasingly concentrated in later layers. On held-out multihop reasoning prompts, ablating subspaces associated with an intermediate concept selectively reduces the probability of the correct downstream answer, while counterfactual interventions can redirect generation toward the answer implied by a substituted concept. Extending this analysis to task-level behavior, we find that ablating contrastive subspaces across seven Qwen and Llama models produces selective effects on syntactic, world-knowledge, and commonsense-reasoning benchmarks, with less off-target disruption on average than non-contrastive subspaces. Together, these results connect local gradient geometry to selective causal effects at both token and task scales.

1 INTRODUCTION

Large language models perform diverse computations within high-dimensional hidden representations. We investigate the hypothesis that computations supporting particular token predictions and tasks preferentially rely on specific subspaces within these representations. A conceptual motivation comes from population neuroscience: Amematsro et al. (2025) find that motor subskills engage distinct, partially overlapping dimensions of activity in macaque motor cortex. This suggests that a high-dimensional system can support diverse behaviors through selective use of its dimensions.

Identifying such organization in language models requires establishing which representational dimensions contribute to a behavior and whether their contributions persist across contexts. Linear probes reveal that many linguistic properties can be decoded from hidden states (Hewitt & Manning, 2019; Belinkov, 2022), but successful decoding does not establish whether the model uses those properties in making a prediction. Causal tracing, model editing, and representation engineering examine this relationship by intervening on internal states and measuring changes in model outputs (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2024). A central challenge is to identify subspaces whose contributions to a particular token prediction or task persist across contexts while remaining selective relative to other behaviors.

Recent work on J-Lens identifies shared directions in hidden representations associated with predicting particular tokens (Gurnee et al., 2026). J-Lens estimates these directions by measuring how small changes to hidden states affect a token's output score and averaging these effects over different input contexts. Such averages can obscure additional directions whose effects vary across contexts. It

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

therefore remains unclear whether these additional directions consistently contribute to predicting the same token and whether they support downstream computation.

We address these questions through a unified framework based on gradients of token scores and behavioral losses with respect to hidden states. Averaging gradient outer products across contexts yields a sensitivity matrix whose leading eigenvectors identify directions with the greatest average squared first-order effect on the chosen score or loss. For an individual token, the normalized mean gradient defines the J-Lens direction J. We project this direction out of each gradient and take the leading eigenspace of the resulting second moment to define the *residual-F* subspace. This subspace captures context-dependent sensitivity orthogonal to the shared direction.

The same framework extends to tasks by applying the sensitivity construction to a behavioral loss. Its leading eigenspace defines a raw F subspace, but high sensitivity to one task may also reflect computations shared with other tasks. To identify selective directions, we contrast the target sensitivity matrix with those of comparison tasks. The resulting generalized eigenvectors, which we call *Contrastive Task Directions* (CTDs), prioritize sensitivity to the target relative to the comparison tasks. Token subspaces and task-selective directions thus follow from the same gradient geometry.

We evaluate the token-level framework in Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. We find that residual sensitivity becomes increasingly concentrated in a small subspace in later layers; independent context splits in Qwen further show that both the mean direction and residual subspace become more reproducible with depth. To test their functional relevance, we intervene on subspaces associated with intermediate concepts in held-out multihop reasoning prompts. Ablation selectively reduces the probability of the correct downstream answer, while counterfactual patches can redirect generation toward the answer implied by a substituted intermediate concept.

Extending the analysis to task-level behavior, we evaluate subspaces across seven Qwen and Llama models on BLiMP syntax, EWoK world knowledge, and physical and social commonsense reasoning in PIQA and SocialIQA (Warstadt et al., 2020; Ivanova et al., 2024; Bisk et al., 2020; Sap et al., 2019). We find that CTD ablations preferentially impair the target behavior and reduce off-target disruption on average relative to raw F. The results support partially separable subspaces for these behaviors.

Our main contributions are:

- A common first- and second-moment gradient framework for identifying shared token directions, residual token subspaces, and task-selective contrastive directions.
- Evidence that token sensitivity contains reproducible structure beyond the mean direction, with causal interventions linking these subspaces to intermediate-concept use in multihop reasoning.
- An evaluation across seven models showing that contrastive subspaces support selective task-level interventions on held-out data while reducing off-target disruption on average.

2 METHOD

2.1 GRADIENT SENSITIVITY

For a scalar token score or behavioral loss ϕ, let $q = \nabla_z \phi$ be its gradient with respect to hidden state z. Across contexts or examples, we define the mean $\mu = \mathbb{E}[q]$ and uncentered second moment $M = \mathbb{E}[qq^\top] = \mu\mu^\top + \mathrm{Cov}(q)$. For a unit direction v, $v^\top M v = \mathbb{E}[(q^\top v)^2]$ measures average squared first-order sensitivity. The leading eigenvectors therefore span the directions of greatest local sensitivity. We use this construction to identify both token- and task-relevant subspaces.

2.2 TOKEN-RELEVANT SUBSPACES

Let $z_{\ell,s}$ be the output of block ℓ at position s, and $o_{w,t}$ the unnormalized score of token w at a later position t. For corpus context c, we collect $q^{(w)}_{\ell,c,s,t} = \nabla_{z_{\ell,s}} o_{w,t}$ with $s < t$. Suppressing context and position indices, the normalized mean $j^{(w)}_\ell = \mathbb{E}[q^{(w)}_\ell]/\|\mathbb{E}[q^{(w)}_\ell]\|_2$ defines the J-Lens direction (Gurnee et al., 2026) (it need not be an eigenvector of the full second moment although empirically we find it is very close to the top eigenvector). We remove this direction and compute

$r_\ell^{(w)} = (I - j_\ell^{(w)} j_\ell^{(w)\top}) q_\ell^{(w)}$ and $F_{\ell,\mathrm{res}}^{(w)} = \mathbb{E}[r_\ell^{(w)} r_\ell^{(w)\top}]$. Its leading k eigenvectors form the residual-F basis $U_{\ell,k}^{(w)}$, capturing context-dependent sensitivity orthogonal to the mean direction.

To measure concentration, we compute a validation residual matrix $F_{\ell,\mathrm{val}}^{(w)}$ on independent contexts using the fitted $j_\ell^{(w)}$. The held-out sensitivity captured by the fitted basis is $C_k = \mathrm{tr}(U_{\ell,k}^{(w)\top} F_{\ell,\mathrm{val}}^{(w)} U_{\ell,k}^{(w)}) / \mathrm{tr}(F_{\ell,\mathrm{val}}^{(w)})$. We also fit directions independently on two equal-sized context sets and measure the absolute cosine between J-Lens directions and the residual energy captured by the other split's subspace. Rank-matched isotropic subspaces provide a chance baseline.

2.3 Task-relevant subspaces

Behavioral objectives. We estimate task sensitivity using losses that compare scores for correct and incorrect alternatives. For context x and candidate sequence $y = (y_1, \ldots, y_T)$, we compute $s(x, y) = \rho(\{\log p_\theta(y_t \mid x, y_{<t})\}_{t=1}^T)$, where ρ denotes a sum or mean over tokens. For a correct alternative with score s_+ and an incorrect alternative with score s_-, the pairwise logistic loss is $\mathcal{L}_{\mathrm{pair}}(s_+, s_-) = \log(1 + \exp(s_- - s_+))$. The losses are used to estimate directions and accuracy is evaluated from the same candidate scores.

Sensitivity estimation. For item loss $\mathcal{L}_i$, let $g_{\ell,i,t} = \nabla_{z_{\ell,t}} \mathcal{L}_i$. Applying the same second-moment construction to domain d gives $F_\ell^{(d)} = N_d^{-1} \sum_{(i,t)\in d} g_{\ell,i,t} g_{\ell,i,t}^\top$, where the sum includes every valid token position in every candidate sequence and N_d counts these gradients. The leading r eigenvectors define the non-contrastive raw-F subspace.

Contrastive Task Directions. To prioritize target sensitivity relative to other tasks, we trace-normalize the target matrix $F_\ell^{(d)}$ and the sum of comparison matrices $\sum_{d' \neq d} F_\ell^{(d')}$ to obtain $\bar{F}_\ell^{(d)}$ and $\bar{F}_\ell^{(-d)}$. CTDs maximize the sensitivity ratio $R(v) = (v^\top \bar{F}_\ell^{(d)} v)/(v^\top (\bar{F}_\ell^{(-d)} + \gamma I) v)$ and solve $\bar{F}_\ell^{(d)} v = \lambda (\bar{F}_\ell^{(-d)} + \gamma I) v$ with $\gamma = 10^{-4}$. We orthonormalize the leading r generalized eigenvectors in the Euclidean metric to obtain $U_{\ell,r}$. This connects generalized eigenvalue methods (Ghojogh et al., 2019) with Fisher-like sensitivity geometries (Martens, 2020; Kirkpatrick et al., 2017).

2.4 Causal interventions and selectivity

Ablation removes an orthonormal basis U_ℓ fitted at each selected layer: $z'_{\ell,t} = z_{\ell,t} - U_\ell U_\ell^\top z_{\ell,t}$.

Token-level ablation. We ablate the J-Lens direction, a rank-k residual-F subspace, or their orthogonal direct sum over a contiguous layer band, intervening on the final query. Selectivity is the decrease in the correct final-answer token's log probability after ablating the correct intermediate concept's subspace, minus the mean decrease for rank-matched wrong-token subspaces. We compare against global activation PCA and isotropic random subspaces to control for nonspecific disruption (the PCA basis is token-independent and has zero correct-minus-wrong selectivity by construction).

Counterfactual patching. For each multihop item, we construct two *donor prompts* that supply activations for the intervention. The source donor explicitly states the original first-hop answer w_{src}; the target donor replaces only this answer with a counterfactual intermediate w_{tgt}. Both contain the same final query, and neither provides the downstream answer. We call an item *donor-capable* if the unmodified model generates the prespecified downstream answer implied by w_{tgt} when given the target donor. This subset is fixed before comparing interventions.

Let $z_{\ell,p}^{\mathrm{src}}$ and $z_{\ell,p}^{\mathrm{tgt}}$ be the residual-stream activations from unmodified donor passes at corresponding final-query positions p. We compute $\delta_{\ell,p} = z_{\ell,p}^{\mathrm{tgt}} - z_{\ell,p}^{\mathrm{src}}$ and, during a new forward pass on the original prompt, patch every final-query position over the selected contiguous layer band $\mathcal{L}$ as $\tilde{z}_{\ell,p} = z_{\ell,p} + \alpha P_\ell^{(w_{\mathrm{tgt}})} \delta_{\ell,p}$ for $\ell \in \mathcal{L}$. Here α controls patch strength, and $P_\ell^{(w_{\mathrm{tgt}})}$ projects orthogonally onto the target intermediate's J-Lens direction, rank-k residual-F subspace, or their orthogonal direct sum. Rank-matched activation-PCA and isotropic-random controls use projected patches rescaled to

match the root-mean-square magnitude of the corresponding residual-F patch. Conversion requires generating the prespecified downstream answer in full, including all tokens of a multi-token answer.

Task-level ablation. We ablate CTDs at every token position and compare against raw-F subspaces and ten independently sampled, rank-matched random orthonormal subspaces. Let $\mathcal{A}$ be the set of evaluation domains and $\Delta\mathrm{Acc}(d \to e)$ the ablated minus baseline accuracy on held-out domain e after ablating directions fitted for d. We report matched damage $D_{\mathrm{match}}(d) = -100\,\Delta\mathrm{Acc}(d \to d)$, mean off-target damage $D_{\mathrm{off}}(d) = -\frac{100}{|\mathcal{A}|-1}\sum_{e \neq d}\Delta\mathrm{Acc}(d \to e)$, and selectivity $S(d) = D_{\mathrm{match}}(d) - D_{\mathrm{off}}(d)$, all in percentage points. Positive S indicates greater impairment of the target behavior. Accuracies are computed over all held-out examples in each broad domain.

3 Experiments

3.1 Benchmarks and splits

BLiMP. BLiMP evaluates grammatical acceptability using pairs of grammatical and ungrammatical sentences (Warstadt et al., 2020). We use their mean token log probabilities as s_+ and s_-, respectively, and apply $\mathcal{L}_{\mathrm{pair}}$. We group minimal pairs into anaphor agreement, determiner-noun agreement, subject-verb agreement, binding, filler-gap dependencies, negative-polarity item (NPI) licensing, quantifiers, and argument structure. Each domain uses 1,000 randomly selected items for direction estimation and 500 disjoint items for evaluation.

EWoK. EWoK evaluates world knowledge using two contexts and two candidate statements, each matched to one context (Ivanova et al., 2024). We score each context-candidate pair by mean token log probability. The comparison margin is the summed score of the two correct assignments minus that of the two swapped assignments. We use material dynamics, social relations, physical relations, and agent properties. We assign each concept pair to estimation or evaluation (40% held out) split. We use up to 300 examples per domain in each partition; the smaller physical-relations pool provides 255 estimation and 160 evaluation examples.

PIQA and SocialIQA. PIQA represents physical commonsense (Bisk et al., 2020) and SocialIQA social commonsense (Sap et al., 2019). We score each candidate answer by summed token log probability and average $\mathcal{L}_{\mathrm{pair}}$ between the correct answer and each incorrect alternative. Direction and evaluation partitions have disjoint manually assigned subdomains. For example, PIQA estimation uses tools/repair, containers/liquids, cleaning/washing, and cooking/food, while evaluation includes force/motion/support, heat/cold, cutting/piercing, clothing/body, paper/writing, and a residual category. We use up to 2,000 estimation and 1,000 evaluation examples per broad domain.

3.2 Token-level datasets and splits

Vocabulary geometry. We study token-relevant subspaces in Qwen2.5-7B-Instruct (Qwen Team, 2024) and Llama-3.1-8B-Instruct using a set of 118 English words spanning abstract nouns, concrete nouns, function words, proper names, verbs and adjectives. Each word is encoded as a single token by both models. For each model, we fit subspaces using 800 pairs of positions (s, t) and validate them using 400 independent pairs. Here s specifies the hidden-state position and $t > s$ the later position at which the word's output score is evaluated. Fitting and validation pairs are drawn from disjoint documents in FineWeb and WikiText-103 (Penedo et al., 2024; Merity et al., 2017).

Intermediate-concept ablation. We select MQuAKE-CF-v2 prompts that each model answers correctly before intervention (Zhong et al., 2023), yielding 122 prompts from 46 independent cases for Qwen-7B and 129 prompts from 46 cases for Llama-8B. Each item has a known intermediate concept and a final answer, both represented by single tokens. For example, "The color of the planet fourth from the Sun is" requires the unspoken intermediate *Mars*. We estimate token subspaces from 800 source-target position pairs ($s < t$) sampled from 400 documents and use a separate set of 250 pairs for geometric validation. Neither corpus contains the behavioral evaluation prompts.

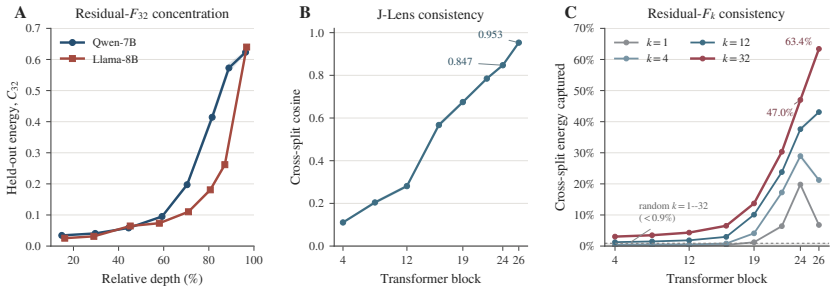


Figure 1: **Token sensitivity becomes concentrated and reproducible with depth. (A)** Held-out residual sensitivity captured by the top 32 directions (C_{32}) in Qwen-7B and Llama-8B (118 shared words, 800 fitting and 400 validation pairs). **(B-C)** Qwen-7B J-Lens cosine agreement (B) and cross-split residual-F_k energy capture (C), using 250 tokens and 400 pairs per independent split. The dashed line is the rank-matched isotropic baseline. Curves show token means; bands are 95% bootstrap intervals over tokens.

Counterfactual patching. We use MQuAKE prompts that each model answers correctly before intervention: 88 prompts from 37 cases for Qwen-7B and 100 prompts from 38 cases for Llama-8B. The donor-capable subsets contain 48 prompts from 22 cases and 71 prompts from 31 cases, respectively. Conversion rates are computed over these subsets using the criterion defined above.

3.3 Models and implementation

Our model set includes Qwen2.5-Instruct at 0.5B, 1.5B, 3B, and 7B parameters, Llama3.2-Instruct at 1B and 3B, and Llama3.1-Instruct at 8B. Selected contiguous mid-to-late layers are 17-23 (Qwen-0.5B), 20-27 (Qwen-1.5B), 26-35 (Qwen-3B), 19-26 (Qwen-7B), 11-15 (Llama-1B), 20-27 (Llama-3B) and 22-30 (Llama-8B). The representative heatmaps use Qwen2.5-1.5B-Instruct.

For token-level experiments, we fit J-Lens and residual-F bases separately at each block and intervene jointly at blocks 19-26 (Qwen) and depth-matched blocks 22-30 (Llama). Ablation ranks are $k \in \{1, 4, 12, 32\}$; J+F has total rank $k + 1$. Counterfactual projected patches use $\alpha = 1.5$. Counterfactual PCA and three random controls are patch-RMS matched to the F method.

Task-level experiments use gradients from all valid tokens and ten random subspaces. Rank sweeps use $r \in \{1, 2, 5, 10, 25\}$, with 50 additionally available for PIQA/SocialIQA.

4 Results

4.1 Do token subspaces generalize across contexts?

We first ask whether token sensitivity contains reproducible structure beyond the mean direction, and how its dimensionality changes across layers. We compare directions fitted on independent context sets and measure how much residual sensitivity their subspaces capture on held-out data.

In Qwen-7B, directions fitted on independent contexts become more consistent with depth (Fig. 1B–C). Mean J-Lens cosine similarity rises from 0.111 at block 4 to 0.847 at block 24 and 0.953 at block 26. A rank-32 residual-F subspace fitted on one split captures 47.0% of the other split's residual energy at block 24 and 63.4% at block 26, compared with less than 0.9% for isotropic random subspaces. The held-out concentration analysis shows that a small number of dimensions captures an increasing fraction of residual sensitivity in both models (Fig. 1A). Mean C_{32} increases from 0.034 at block 4 to 0.623 at block 26 in Qwen, and from 0.025 at block 5 to 0.640 at block 30 in Llama. Token sensitivity therefore contains a reproducible, multidimensional component beyond the shared mean direction, with a larger fraction captured by small subspaces in later layers.

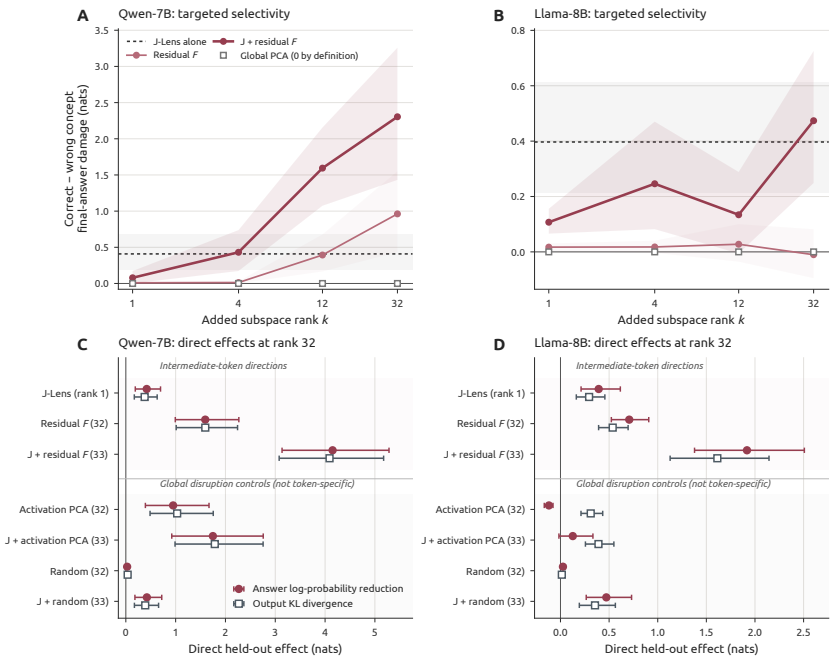


Figure 2: **Intermediate-concept ablation in Qwen-7B and Llama-8B. (A–B)** Selectivity (correct-minus-wrong answer damage) across residual ranks. J+F_k includes J-Lens; dotted lines show J-Lens alone. **(C–D)** Rank-32 correct-answer log-probability reduction (filled circles) and output KL divergence (open squares). PCA and random measure nonspecific disruption; PCA's shared basis has zero selectivity. Qwen and Llama use 122 and 129 held-out MQuAKE prompts, respectively, from 46 cases each. Intervals are 95% bootstrap intervals over 10,000 resamples.

4.2 Does the ablation of token subspaces selectively impair multihop reasoning?

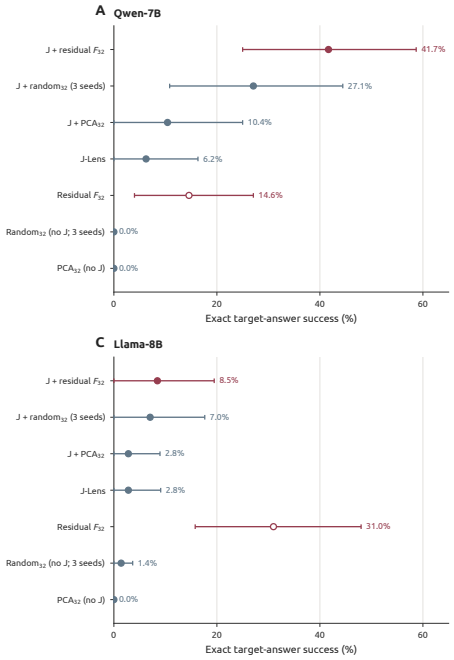
We next ask whether these subspaces selectively support reasoning that depends on the corresponding intermediate concept. On held-out multihop prompts, we compare the reduction in the correct answer's log probability after ablating the correct intermediate's subspace with that produced by rank-matched subspaces for incorrect concepts.

In Qwen, removing the correct concept's residual-F subspace causes greater answer damage than removing wrong-concept subspaces (Fig. 2A). At rank 32, this selectivity is 0.962 nats (95% CI $[0.433, 1.552]$). Combining J-Lens with F_{32} yields 2.304 nats ($[1.431, 3.259]$), compared with 0.409 for J-Lens alone ($[0.181, 0.677]$). Llama shows a narrower range of selective residual directions (Fig. 2B). Standalone F_1 yields 0.017 nats of selectivity ($[0.003, 0.030]$), whereas standalone F_{32} is not selective. The combined J+F_{32} intervention remains selective at 0.474 nats ($[0.249, 0.726]$), compared with 0.397 for J-Lens alone ($[0.213, 0.613]$). Comparison against PCA shows that while PCA can also change the output distribution substantially, it does so in a nonselective manner and its token-independent basis has zero correct-minus-wrong selectivity.

4.3 Can patching token-level subspace redirect the downstream answer?

For token-level analysis, finally, we ask whether these subspaces can carry information about a substituted intermediate concept and redirect the downstream answer. We project donor activation differences onto the target concept's subspace and patch the original prompt, measuring conversion to the prespecified answer among donor-capable prompts.

In Qwen, the combined J+F_{32} patch converts 20/48 prompts (41.7%, 95% CI $[25.0, 58.7]$), compared with 3/48 for J-Lens, 5/48 for J+PCA, and a three-seed mean of 27.1% for J+random (Fig. 3A). Standalone F_{32} converts 7/48, while standalone PCA and random patches convert none.



B

Question and intermediate edit	Clean	Comparator outputs	J + residual F_{32}
1. Capital What is the capital city of the country to which James Marsh belongs? *Intermediate: UK → Portugal*	London	J-Lens: London J + PCA$_{32}$: Reykjavik J + random$_{32}$: London / Reykjavik / London	Lisbon (target)
2. Continent Which continent does the nationality of Seán McLoughlin belong to? *Intermediate: Ireland → Venezuela*	Europe	J-Lens: Europe J + PCA$_{32}$: Europe J + random$_{32}$: Europe (3/3)	South America (target)
3. Official language Which language is recognized as the official language in Htin Kyaw's country of citizenship? *Intermediate: Myanmar → Netherlands*	Burmese	J-Lens: Burmese J + PCA$_{32}$: Burmese J + random$_{32}$: Burmese (3/3)	Dutch (target)

D

Question and intermediate edit	Clean	Comparator outputs	Residual F_{32}
1. Capital What is the capital of the country of citizenship of John Gorton? *Intermediate: Australia → China*	Canberra	J-Lens: China J + PCA$_{32}$: China J + random$_{32}$: China (3/3)	Beijing (target)
2. Continent In which continent is the country of citizenship of Johann Jakob Froberger located? *Intermediate: Germany → Ecuador*	Europe	J-Lens: Ecuador J + PCA$_{32}$: Ecuador J + random$_{32}$: Ecuador (3/3)	South America (target)
3. Official language In which language are the official affairs conducted in the country where Liam Fox holds citizenship? *Intermediate: UK → Turkey*	English	J-Lens: Turkey J + PCA$_{32}$: Turkey J + random$_{32}$: Turkey (3/3)	Turkish (target)

Figure 3: **Counterfactual intermediate-concept patching on MQuAKE. (A,C)** Target-answer conversion in Qwen-7B and Llama-8B using 48 and 71 donor-capable prompts from 22 and 31 cases, respectively. PCA and random controls match rank and RMS patch magnitude; random results average three seeds. **(B,D)** Example conversions using J+F_{32} (Qwen) and F_{32} (Llama), where J-Lens, J+PCA, and all J+random controls fail. Intervals are 95% bootstrap intervals over 10,000 resamples.

In Llama, standalone F_{32} is the most effective projection, converting 22/71 prompts (31.0%, 95% CI $[15.8, 48.0]$), compared with 2/71 for J-Lens, 6/71 for J+F_{32}, 0/71 for PCA, and 3/213 across three random seeds (Fig. 3C). Adding the mean direction therefore affects the two models differently.

Donor/subspace mismatch controls test whether these conversions depend on the intended intermediate concept. In Qwen, the matched donor and J+F_{32} subspace convert 41.7%, whereas the three mismatch conditions convert at most 1.4%. In Llama, the matched standalone F_{32} patch converts 31.0%, while wrong donors projected through their own F_{32} subspaces convert 1/213 (0.47%).

Within this donor-capable subset, residual-F subspaces can transfer information about a substituted concept and redirect downstream generation. The best projection differs across models, with Qwen favoring the combined mean and residual subspace and Llama the residual subspace alone.

4.4 BLiMP SYNTAX

We ask whether ablating a subspace associated with one syntactic phenomenon selectively impairs judgments involving that phenomenon. We compare CTD, raw-F, and rank-matched random ablations on held-out examples from the target and comparison BLiMP domains.

In Qwen2.5-1.5B, CTDs produce their largest target effects for NPI licensing, anaphor agreement, and subject-verb agreement (Fig. 4A). Accuracy drops by 28.2, 24.8, and 24.4 pp, respectively, with selectivities of 27.0, 23.1, and 24.0 pp. Quantifiers and binding also show selective effects (15.0 and 11.0 pp). Rank-matched random ablations generally cause much smaller target drops.

Across seven models and eight domains, mean CTD selectivity is 10.3 pp, versus 9.5 pp for raw F (Fig. 4D). CTDs produce smaller target drops on average (11.1 versus 16.5 pp), but also much less off-target damage (0.8 versus 7.0 pp). These results suggest syntactic judgments can be selectively disrupted through domain-associated subspaces (though this depends on how subspaces are estimated).

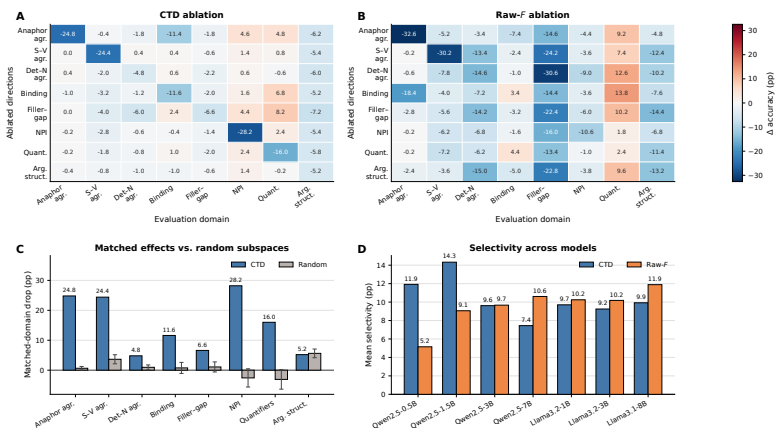


Figure 4: **BLiMP, example-held-out evaluation. (A–B)** Accuracy change after CTD and raw-F ablations (Qwen2.5-1.5B, rank 10). Rows are estimation domains; columns are evaluation domains. **(C)** Matched CTD drops and mean±SD of ten rank-matched random ablations. **(D)** Domain-mean selectivity for the seven common models. All panels use example-weighted broad-domain accuracy.

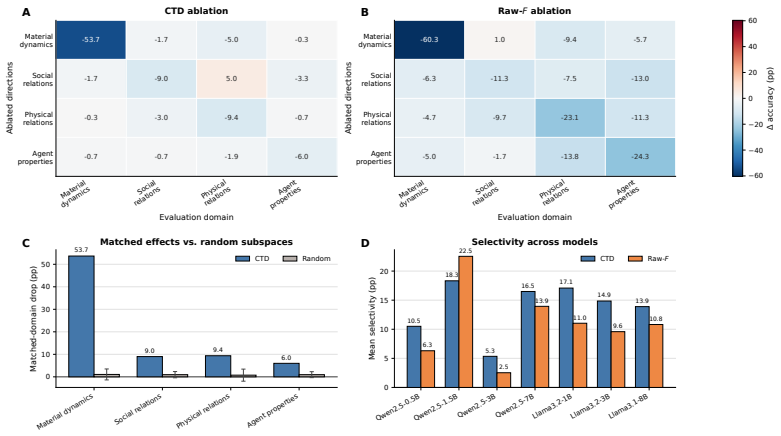


Figure 5: **EWoK concept-held-out evaluation.** Panels follow Fig. 4, using Qwen2.5-1.5B at rank 25. Estimation and evaluation use disjoint concept pairs; concept-group accuracies are not averaged.

4.5 EWoK CONCEPT HOLDOUT

We next ask whether subspaces associated with a world-knowledge domain retain selective effects on unseen concepts. We estimate directions and evaluate ablations on disjoint concept pairs in EWoK.

In Qwen2.5-1.5B, material-dynamics CTDs reduce target accuracy by 53.7 pp while causing a mean off-target drop of 2.3 pp, yielding 51.3 pp selectivity (Fig. 5A). Physical-relations, social-relations, and agent-properties CTDs yield smaller target drops of 9.4, 9.0, and 6.0 pp, with selectivities of 8.0, 9.0, and 4.9 pp. Their corresponding random ablations reduce target accuracy by only 0.8–1.1 pp.

Across seven models and four domains, CTD selectivity averages 13.8 pp versus 11.0 pp for raw F, with mean off-target damage of 0.7 versus 4.5 pp (Fig. 5D). Six of the seven models favor CTD. The exception is Qwen2.5-1.5B, where raw F has higher mean selectivity (22.5 versus 18.3 pp) and is particularly effective for material dynamics and agent properties. Thus, the large selective effects in the representative model are also obtainable with non-contrastive subspaces.

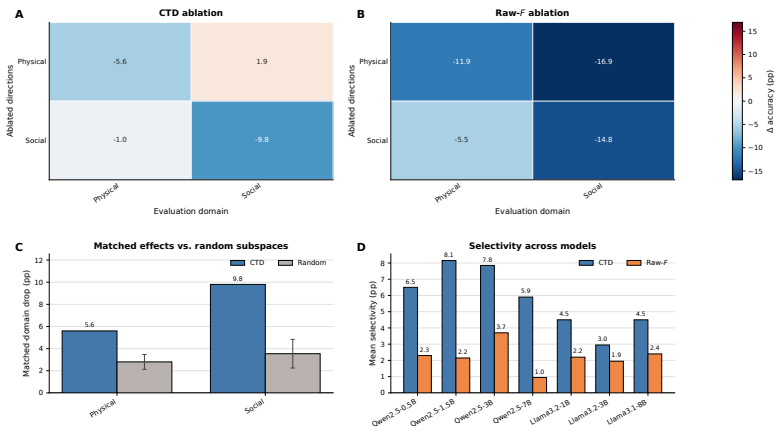


Figure 6: **PIQA/SocialIQA subdomain-held-out evaluation.** Panels follow Fig. 4, using Qwen2.5-1.5B at rank 50. Estimation and evaluation use disjoint physical or social subdomains. Each broad-domain cell is computed directly over its held-out examples; subdomain means are not averaged.

4.6 PIQA AND SOCIALIQA SUBDOMAIN HOLDOUT

Finally, we ask whether physical and social commonsense reasoning can be selectively affected on new types of problems. We estimate subspaces using one set of physical and social subdomains and evaluate ablations on disjoint subdomains in PIQA and SocialIQA.

In Qwen2.5-1.5B, physical CTDs reduce held-out physical accuracy by 5.6 pp while improving social accuracy by 1.9 pp, giving 7.5 pp selectivity (Fig. 6A). Social CTDs reduce social accuracy by 9.8 pp and physical accuracy by 1.0 pp, giving 8.8 pp selectivity. Random ablations produce smaller target drops (2.8 and 3.5 pp). Here, raw F consistently showed lower mean selectivity than CTD (Fig. 6D).

5 DISCUSSION

We introduce a novel framework for understanding token prediction and task performance through the geometry of representational sensitivity. These findings support the hypothesis that specific computations depend preferentially on subspaces within high-dimensional model representations. At the token level, we identify reproducible structure beyond the mean sensitivity direction and show that intervening on this structure can selectively impair or redirect downstream answers.

This token-level sensitivity raises a broader question: to what extent different language model capabilities can be functionally dissociated within a shared representational space? Our task-level ablations address this question by measuring how strongly a subspace affects one behavior relative to others. Our results show that contrasting sensitivities yields subspaces whose ablation causes less off-target disruption on average than non-contrastive alternatives. This observed selectivity supports partially separable computations within high-dimensional representations. Connecting these subspaces to the components that construct and read them could reveal which mechanisms are shared across capabilities and which selectively support particular behaviors.

Our approach has several limitations. First, F and CTDs summarize first-order sensitivity around the model's existing states, whereas multilayer ablation is a finite intervention that can move activations off manifold. Second, selectivity depends on the chosen objectives, comparison tasks, and intervention settings (e.g., chosen rank) and its robustness should be tested across them. Within these limits, our results suggest that the second-moment structure of sensitivity, beyond its mean direction and contrasted across tasks, identifies representational subspaces that particular behaviors rely on. Much as motor cortex deploys distinct subskills through partially overlapping dimensions of a shared population (Amematsro et al., 2025), language models may support diverse capabilities through selective use of a common representational space.

AI USE STATEMENT

AI tools assisted with manuscript editing for clarity and concision and with literature searches. The author takes responsibility for the final text including scientific claims and references.

REFERENCES

Elom A. Amematsro, Eric M. Trautmann, Najja J. Marshall, L. F. Abbott, Michael N. Shadlen, Daniel M. Wolpert, and Mark M. Churchland. Motor cortex flexibly deploys a high-dimensional repertoire of subskills. *bioRxiv*, 2025. doi: 10.1101/2025.09.07.674717. URL https://doi.org/10.1101/2025.09.07.674717. Preprint.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. Eigenvalue and generalized eigenvalue problems: Tutorial. *arXiv preprint arXiv:1903.11240*, 2019.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026. URL https://arxiv.org/abs/2607.15495.

John Hewitt and Christopher D. Manning. 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 (NAACL-HLT)*, pp. 4129–4138, 2019.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U. Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. Elements of world knowledge (EWoK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. SocialIQA: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023. doi: 10.18653/v1/2023.emnlp-main.971.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A CROSS-MODEL RESULTS

Table 1 gives the exact domain-mean selectivity values plotted in the main figures. Dashes denote incomplete BLiMP multilayer runs; we do not substitute single-layer results.

	BLiMP, $r = 10$		EWoK, $r = 25$		PIQA/SocialIQA, $r = 50$	
Model	CTD	Raw F	CTD	Raw F	CTD	Raw F
Qwen2.5-0.5B	**11.92**	5.15	**10.51**	6.31	**6.50**	2.30
Qwen2.5-1.5B	**14.32**	9.06	18.33	**22.53**	**8.15**	2.15
Qwen2.5-3B	9.61	**9.67**	**5.35**	2.53	**7.85**	3.70
Qwen2.5-7B	7.45	**10.61**	**16.49**	13.94	**5.90**	0.95
Llama3.2-1B	9.70	**10.25**	**17.08**	11.03	**4.50**	2.20
Llama3.2-3B	9.25	**10.18**	**14.85**	9.59	**2.95**	1.95
Llama3.1-8B	9.93	**11.89**	**13.89**	10.82	**4.50**	2.40
Common-model mean	**10.31**	9.54	**13.79**	10.96	**5.76**	2.24
Gemma-2-2B	–	–	10.05	**11.80**	**6.15**	2.10
Mistral-7B	–	–	**13.02**	6.95	**5.95**	2.25

Table 1: Mean selectivity (pp) across target domains. Every value is computed from example-weighted broad-domain accuracies. The "common-model mean" uses the seven models with complete results for all three benchmarks.

B RANK SENSITIVITY

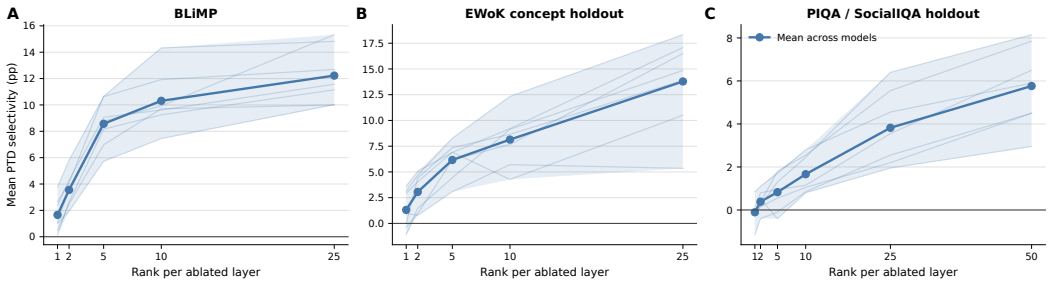


Figure 7: **CTD selectivity across ranks.** Thick lines are means over the seven common models and target domains; faint lines are model means and shading spans the minimum to maximum model mean. Selectivity generally increases with rank over the evaluated range, although the between-model spread also grows.