

The Geometry of Political Abstraction in Large Language Models

Noé Durandard^{1,2*}, Saurabh Dhawan^{3*}, Thierry Poibeau²,

¹Ecole Normale Supérieure - PSL, Paris, France,

²LATTICE, ENS, PSL, U. Sorbonne Nouvelle Paris 3, CNRS, Montrouge, France,

³Technical University of Munich, Munich, Germany,

noe.durandard@psl.eu, saurabh.dhawan@tum.de, thierry.poibeau@ens.psl.eu

Abstract

Large language models encode high-level concepts with heterogeneous domain-specific manifestations, but it remains unclear whether their internal representations are organized in a form that supports cross-domain generalization. We investigate this question by adapting representational geometry methods from systems neuroscience and using political ideology as a naturalistic test case for a candidate latent variable spanning diverse policy domains. Across three open-weight LLMs, we analyze hidden states elicited by matched left- and right-leaning statements from ten policy domains. Within each domain, while political polarity is linearly decodable, lexical features achieve comparable performance. By contrast, cross-domain generalization substantially exceeds the lexical baseline and peaks in intermediate layers. This peak coincides with a marked reduction in participation ratio, suggesting that domain-specific political distinctions, instead of being maintained as independent axes, become aligned within a shared low-dimensional subspace. These findings suggest that LLMs transiently organize heterogeneous inputs into a common representational geometry that supports cross-domain abstraction, offering insight into how models structure complex social knowledge.¹

1 Introduction

Large language models (LLMs) exhibit a remarkable capacity to generalize across tasks, domains, and linguistic contexts (Brown et al., 2020; Conneau et al., 2020). Their success suggests that they do not merely memorize statistical regularities, but instead construct internal representations that capture higher-order structure underlying language and

the world. A central question for mechanistic interpretability is therefore not simply what information can be decoded from model activations, but how that information is organized and used within the models distributed representational space (Elhage et al., 2022; Geiger et al., 2025). Existing interpretability research has made substantial progress in identifying decodable concepts, localized features, and mechanistic circuits (Belinkov, 2022; Conmy et al., 2023; Templeton et al., 2026), yet comparatively little is known about the representational geometry that enables models to generalize across semantically distinct situations. Understanding this geometry is essential for explaining how neural networks transform diverse observations into abstract concepts that support flexible reasoning (Halford et al., 2010; Kemp and Jern, 2009; Kriegeskorte and Kievit, 2013).

Recent work in systems neuroscience has proposed that a capacity for abstraction depends on the geometric organization of neural state space instead of merely being a property of information content (Bernardi et al., 2020; Courellis et al., 2024). In this view, a representation is considered abstract if it organizes neural activity such that a decoder trained under one set of conditions generalizes to previously unseen combinations of other task variables. These analyses have revealed that populations of neurons in the hippocampus and prefrontal cortex encode hidden task variables in formats that directly support inference and flexible behavior. Together, these studies suggest that abstraction is not simply a matter of information content, but of how information is structured within a distributed representational space (Chung and Abbott, 2021).

Whether analogous representational principles emerge in LLMs remains underexplored. Most existing analyses have focused on identifying semantic features, concept directions, or individual computational components (Bricken et al., 2023; Gurnee and Tegmark, 2024; Marks and Tegmark,

* These authors contributed equally to this work.

¹Data and code are available on GitHub at [d-noe/llm_political_abstraction](https://github.com/d-noe/llm_political_abstraction).

2024; Templeton et al., 2026; Wang et al., 2022). While these approaches establish that particular concepts are encoded, they do not distinguish between representations that merely contain information and those that organize information into latent variables capable of supporting structural generalization. This distinction is particularly important for high-level socially constructed concepts, whose instances often differ dramatically in surface form while sharing a common conceptual organization.

Political ideology (Feldman and Johnston, 2014) provides a naturalistic test case for addressing this question. Unlike laboratory tasks in which latent variables are explicitly designed by the experimenter, political ideology emerges from long-term social, historical, and institutional processes. It is expressed across diverse policy domains, including taxation, healthcare, immigration, climate policy, etc. that share little vocabulary or semantic content, yet humans readily recognize them as instantiations of a common left-right distinction. Consequently, successful representation of ideology cannot rely solely on lexical similarity or topic-specific associations, but instead requires extracting a latent conceptual variable that spans heterogeneous semantic contexts. Political ideology therefore constitutes a demanding benchmark for abstraction: if a model represents ideological stance in a common format across policy domains, it provides evidence that the model has organized diverse observations around an abstract conceptual coordinate rather than a collection of independent topical features.

Here we investigate whether LLMs represent political ideology as an abstract cross-domain variable using geometric analyses inspired by systems neuroscience. We analyze hidden activations elicited by matched pairs of ideologically opposed political statements spanning multiple policy domains and characterize their representational geometry across model layers. Specifically, we test whether classifiers trained on one set of policy domains generalize to unseen domains, and whether ideological representations occupy low-dimensional subspaces while remaining sufficiently expressive to preserve rich semantic structure. Rather than evaluating political bias or ideological preference, our objective is to characterize the internal organization of distributed representations that supports abstraction.

2 Related Work

Interpretability research has shown that linguistic, semantic, factual, and behavioral variables are often recoverable from model activations using linear probes, sparse autoencoders, and circuit-level analyses (Belinkov, 2022; Bricken et al., 2023; Templeton et al., 2026; Wang et al., 2022). These methods primarily establish what information is encoded, rather than how different variables are geometrically organized within the models representational space. The global organization of distributed representations that enables abstraction and structural generalization remains underexplored.

Some recent work has started to examine the geometry of high-level concepts in language models. Marks and Tegmark (2024) showed that truth and falsehood are encoded as approximately linear directions in activation space, while Kim et al. (2025) demonstrated that political perspective forms a linear representational axis that generalizes across politicians and media outlets. These studies provide evidence that important conceptual variables may be encoded geometrically. However, they primarily establish the existence of linear representational directions rather than asking whether these directions participate in an abstract geometry that supports structural generalization across semantically distinct domains.

A complementary perspective originates in systems neuroscience (Chung and Abbott, 2021), where abstraction has been characterized geometrically rather than informationally. Bernardi et al. (2020) introduced cross-condition generalization performance (CCGP) and parallelism score as operational measures of abstract neural representations, demonstrating that variables are represented abstractly when their coding geometry supports generalization across previously unseen task conditions. More recently, Courellis et al. (2024) showed that similar geometric signatures emerge in human hippocampal neurons during inference, suggesting that abstract representations naturally emerge in biological neural systems.

Political ideology in LLMs has been studied from the perspectives of bias evaluation, and value alignment (Bang et al., 2024; Röttger et al., 2024; Faulborn et al., 2025). In the Political Science literature, prior studies have shown multiple ways to estimate ideological positions from political text. Early text-scaling methods such as Word-scores (Laver et al., 2003) and Wordfish (Slapin and

Proksch, 2008) demonstrated that latent political positions can be recovered from patterns of word usage, while recent work shows that LLMs can likewise position political texts along ideological and policy dimensions with high accuracy (Le Mens and Gallego, 2025). Whereas this literature seeks to infer political positions from observed text, we examine the internal representations of LLMs and ask what representational organization inside the model makes this possible? We treat ideology not as an object of political analysis but as a naturalistic latent variable spanning multiple policy domains, that can act as a demanding test case for a wider class of complex social ideas, and ask whether it is represented in internal model activations as an abstract cross-domain concept. This allows us to evaluate whether LLM representations exhibit the geometric signatures of abstraction previously identified in systems neuroscience, thus extending these methods to socially grounded concepts in LLMs.

3 Methodology

The structure of political information representation in LLMs’ latent spaces is analysed through the activation patterns induced by politically charged texts. This section presents the different steps of the experimental procedure, designed to run principled layer-wise evaluation of LLMs’ representation spaces. It outlines data curation and preparation, models selection, activation extraction, and the mathematical tools, inspired by neuroscience, used to operationalise and quantify political abstraction. A schematic overview of the experimental pipeline is illustrated in Figure 1.

3.1 Experimental Framework

This study relies on representations extracted from the activation spaces of three medium-sized LLMs in response to politically loaded stimuli.

3.1.1 Dataset

As the base material for the experiments, the Twin-Views dataset (Fulay et al., 2024) was selected because it aggregates a large-scale set of politically opposed stances, labelled by political domains.

This dataset originally contains nearly 14K pairs of left- and right- leaning political statements covering a broad range of policy issues. Each pair consist of short, one- to two- sentence long stances generated with ChatGPT (GPT-Turbo 3.5) to express starkly diverging perspectives on a given topic. The

dataset was further “[extensively audited] to ensure ideological balance and topical relevance”².

To allow for domain-related analyses, Twin-Views was further refined through label-smoothing and filtering. First, topic-labels are unified through hierarchical clustering on the original fine-grain topic tags using BERTopic (Grootendorst, 2022) to merge highly overlapping semantic categories. Then, only the top ten topics are retained to filter out sparse or ambiguous topics.

This refinement process results in a final dataset of 7,886 stance pairs, or $N_{\text{data}} = 15,772$ political statements in total, associated with ten policy issues, namely: *tax, regulation, education, gun control, energy, healthcare, diversity and inclusion, immigration, social welfare, and climate change*. Additional details and examples are provided in Appendix A.

3.1.2 Models

The experimental framework is applied identically to three medium-sized LLMs, chosen to control for architectural dimensions, alignment paradigms and training strategies: Llama-3.1-8B, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and OLMo-2-1124-7B-Instruct (OLMo et al., 2024).

From a structural perspective, all three models share an identical geometry with $L = 32$ computational layers and a latent space dimension of $D = 4096$. This design enables direct comparisons between models in the ensuing analyses, where the extracted latent representations are the core material.

These specific models were selected to allow to differentiate potential variations along two comparative axes. First, the inclusion of both base and instruction-tuned versions of Llama-3.1-8B isolates the impact of the alignment-tuning. Then, studying models from Llama-3.1 and OLMo-2 suites, contrasts model families through representatives of proprietary and open-science ecosystems. Also, the inclusion of OLMo-2-1124-7B-Instruct, a fully transparently trained LLM, may open avenues for further investigations into the origins of the measured behaviors, from training data, to pre-training checkpoints, until alignment tuning.

²While the relevance of binary political left (l) and right (r) tags can certainly be discussed, they serve as an operationalisation of polarized, competing viewpoints within specific domains, reflecting the contemporary partisan cleavage in the United States’ political landscape (e.g. Levendusky, 2009).

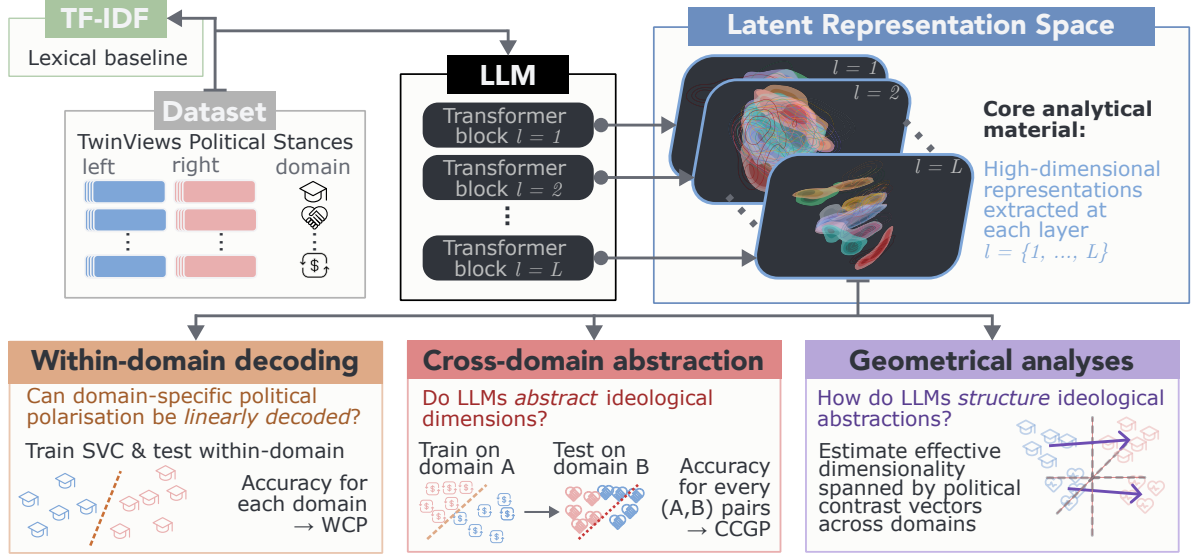


Figure 1: Schematic representation of the experimental framework. The first row illustrates the data acquisition part (related to subsection 3.1), and the second row introduces the analytical pipeline (outlined in 3.2).

3.1.3 Representation Extraction

To investigate the latent representations of LLMs, internal hidden states are systematically extracted across the networks depth during a single forward pass.

The raw textual stances from TwinViews are directly fed into the models without any prompt engineering, instruction wrapping, or specialized formatting templates. Respecting the dynamic nature of information processing in feed-forward neural networks, for each input text fed to a model m constituted of L_m layers, we extract the sequence of hidden state tensor H_l from the residual stream at the output of every Transformer block $l \in \{0, 1, \dots, L_m\}$.

Then, because causal LMs process text autoregressively, the activation vector corresponding to the final token in the sequence is assumed to aggregate the contextual information of the entire input. Consequently, and for analytical convenience, we employ a last-token pooling strategy, discarding all intermediate token vectors and retaining only the final token’s hidden state at each layer.

This extraction pipeline yields an activation matrix $\mathbf{X}_m \in \mathbb{R}^{N_{\text{data}} \times L_m \times D_m}$, for each model m , where N_{data} is the number of samples in the dataset, L_m the number of layers in the model, and D_m the dimensionality of the model’s latent representation space. The recorded $\mathbf{X}_m$ vectors are the core material for the ensuing analyses.

3.2 Analytical Framework

The analytical framework largely relies on methodologies established in system neuroscience by [Bernardi et al. \(2020\)](#), who set the geometric hallmarks of feature abstraction in neural populations. According to their definition, an abstract representation requires that variables are encoded such that they allow for generalisation across context-dependent conditions.

3.2.1 Probing Decoding Directions

In line with the aforementioned concept of abstraction, the analysis is fundamentally driven by domain-constrained linear Support Vector Classifiers (SVCs) trained to retrieve separating hyperplanes between opposing political views. These probes are evaluated along two primary dimensions: within conditions and across conditions.

Within Condition Performance (WCP) WCP verifies whether the studied feature is linearly decodable within individual contexts. Formally, given a set of K domains $\mathcal{D} = \{d_1, d_2, \dots, d_K\}$ (here, $K = 10$ policy issues), let $\theta_{i \rightarrow i}$ denote the mean cross-validation classification accuracy (here, binary classification between opposing political stances) of linear probes trained and evaluated strictly within the data partition of domain d_i . Then, for each model, and at each layer, the overall WCP is given by the macro-average of these

within-domain decoding accuracies:

$$\text{WCP} = \frac{1}{K} \sum_{i=1}^K \theta_{i \rightarrow i} \quad (1)$$

A low WCP indicates that the probes fail to retrieve domain-specific hyperplanes separating left- from right-leaning views, while high values demonstrate that such directions exist on an issue-basis, outlining that the model represents local polarized structure.

Cross-Condition Generalisation Performance (CCGP)

This metric formalises and tests semantic abstraction within latent representation spaces. The performance of probes trained on single domains are assessed on every other domain. Adapting the framework of [Bernardi et al. \(2020\)](#), the structural abstraction is quantified through the cross-condition generalisation performance (CCGP). Here, CCGP measures whether linear hyperplanes trained to separate views within a given context generalise to other domains. Formally, letting $\theta_{i \rightarrow j}$ be the accuracy of a probe trained on representations from domain d_i and evaluated in domain d_j , the macro CCGP is the average performance across all distinct domain pairs:

$$\text{CCGP} = \frac{1}{K(K-1)} \sum_{1 \leq i \neq j \leq K} \theta_{i \rightarrow j} \quad (2)$$

Geometrically, a high CCGP requires that the coding directions separating left- from right-leaning statements are sufficiently aligned to allow for functional classification generalisation, indicating the existence of an overarching ideology-discriminant dimension. In contrast, if the models treat policy domains as structurally unrelated, the cross-domain directions are expected to be orthogonal in the latent space, forcing CCGP toward a chance level of 50%.

In practice, to take into account data variability, all linear probes are trained using a 5-fold cross-validation scheme. WCP is computed on the 20% holdout set within the same domain, while for CCGP, the accuracy is calculated on every other domain using the entirety of the available data. All reported performance scores reflect the mean classification accuracy across these random folds. The results related to these analyses are reported respectively in [section 4](#) and [section 5](#).

Lexical Baseline To substantiate that the decoded geometry reflects conceptual abstraction beyond surface-level lexico-syntactic patterns, the probing results are contrasted against a Term Frequency Inverse Document Frequency (TF-IDF) baseline. The textual features extracted are then treated identically as the latent vector representations, using linear SVCs for classification across 5 data folds. This method provides a strict lower bound for performance achievable solely through surface-level lexical features.

If latent representations yield higher separability than the TF-IDF, especially across domains, it provides evidences for semantic abstraction within the models’ internal spaces that transcends mere vocabulary differences.

3.2.2 Latent Geometry and Effective Dimensionality

Further, the structural alignment of domain-specific political directions within the latent geometry is investigated using the *contrast vectors*. Those are defined as the directional axes of partisan divide for each individual policy issue. Contrast vectors are computed as normalised centroids’ differences. In practice, for each domain $d_i \in \mathcal{D}$, stance-wise centroids $\mu_{i,r}$ and $\mu_{i,l}$ are computed as the mean of the latent vectors of the right- and left- stances, conditioned by the domain d_i . Normalised contrast vectors are then given by:

$$v_i = \frac{\mu_{i,r} - \mu_{i,l}}{\|\mu_{i,r} - \mu_{i,l}\|} \quad (3)$$

Contrary to [Bernardi et al. \(2020\)](#) and traditional neuroscience frameworks which typically rely on max-permutation-based parallelism scores (PS) to explore whether parallel structures exist, the hypothesis-driven nature of our work demands a slightly different approach. A natural adaptation of PS to our experiments could involve averaging parallelism scores across domain pairs, but instead we opt for a mathematically related metric (as discussed in [Appendix C](#)): the **participation ratio** (PR).

PR provides a continuous estimation of the effective number of orthogonal dimensions required to fully characterize the geometry of political polarisations. The set of contrast vectors $\{v_i\}_{i=1}^K$ is used to construct a $K \times K$ Gram matrix G , where each entry $G_{i,j} = v_i \cdot v_j$ represents the cosine similarity between the polarisation axes of domain i and domain j . PR is then derived from the eigenvalues

$\{\lambda_k\}_{k=1}^K$ of the Gram matrix G :

$$\text{PR} = \frac{\left(\sum_{k=1}^K \lambda_k\right)^2}{\sum_{k=1}^K \lambda_k^2} \quad (4)$$

Through eigenmodes, PR acts as a continuous estimator of the number of orthogonal directions needed to reconstruct the distribution of original contrast vectors. This score theoretically ranges from 1 to K , where 1 indicates a single global axis (here, all domains’ first modes are perfectly collinear), and K means that all contrast vectors are strictly orthogonal (here, completely independent, domain-specific features). It was preferred over PS to capture global organisation patterns of the contrast vectors while mitigating local pairwise dependencies. Empirical results are presented in [section 6](#).

The subsequent sections map out the findings along the pipeline described above. First, [section 4](#) establishes whether local domain polarization is linearly decodable from LLMs’ latent representations. Then, [section 5](#) expands this to cross-domain evaluations to test for overarching ideological axes. Finally, [section 6](#) seeks to characterize the intrinsic dimensionality and geometric layout of these shared policy representation spaces.

4 Policy-Based Political Leaning is Linearly Encoded in Latent Space

The first step of the experimental pipeline establishes whether opposing political stances can be linearly decoded from LLMs’ internal representations when conditioning on individual policy domains. This analysis serves as a critical sanity check: before evaluating whether a LLM *abstracts* a concept of ideology, one must verify that it encodes local, domain-specific polarization. Hence, WCP scores ([Equation 1](#)) are computed for all layers of the tested models.

As illustrated on [Figure 2](#), for each model, WCP reaches high scores throughout the layers. In the earliest layers ($l \leq 3$), WCP increases rapidly from $\sim 95\%$ – 96% to near-perfect classification accuracy ($\sim 99\%$ – 100%) by layer $l = 4$. Then the values plateau at ceiling level performance until the last layer of the models.

This rapid saturation aligns with the TF-IDF lexical standpoint, which also achieve near-perfect WCP. This suggests that, at local scales, the vocabulary choices and lexical tokens characteristics

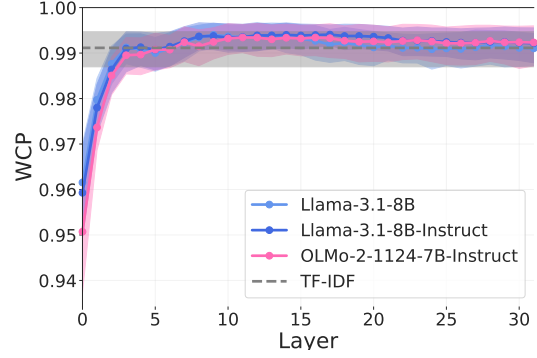


Figure 2: Within-Condition Performance across layers with 95% bootstrapped confidence intervals. Horizontal dashed line shows the lexical baseline performance.

of partisan stances within TwinViews are sufficient to cleanly separate views from opposing sides of the political spectrum.

If the results confirm that linear directions encoding domain polarization are highly prominent in LLMs latent space, WCP alone does not differentiate between two competing structural hypotheses:

1. The model represents political divides in disjoint, local clusters on an issue-basis, independently from one domain to another, or
2. The model maps all domains onto a unified political abstraction direction.

5 Political leaning is represented abstractly across policy domains

Having established, in [section 4](#), that opposing views can be consistently separated within individual domains, a more fundamental question arises: *Does this organisation reflect a collection of independent domain-specific features, or do LLMs construct a unified overarching ideological abstraction?* These hypotheses are tested through the evaluation of cross-condition performance of within-domain trained probes. CCGP ([Equation 2](#)) is used to assess potential structural abstractions.

The empirical behavior of CCGP across network depth exposes a distinct architectural phenomenon, as shown in [Figure 3](#). Unlike WCP, which saturates since early layers, CCGP follows a non-monotonic curve, common to each tested model. It first increases in the early layers, then peaks in the middle layers, before slowly decaying in the later layers.

Indeed, in accordance with most findings on Transformer-based models, the early layer seem

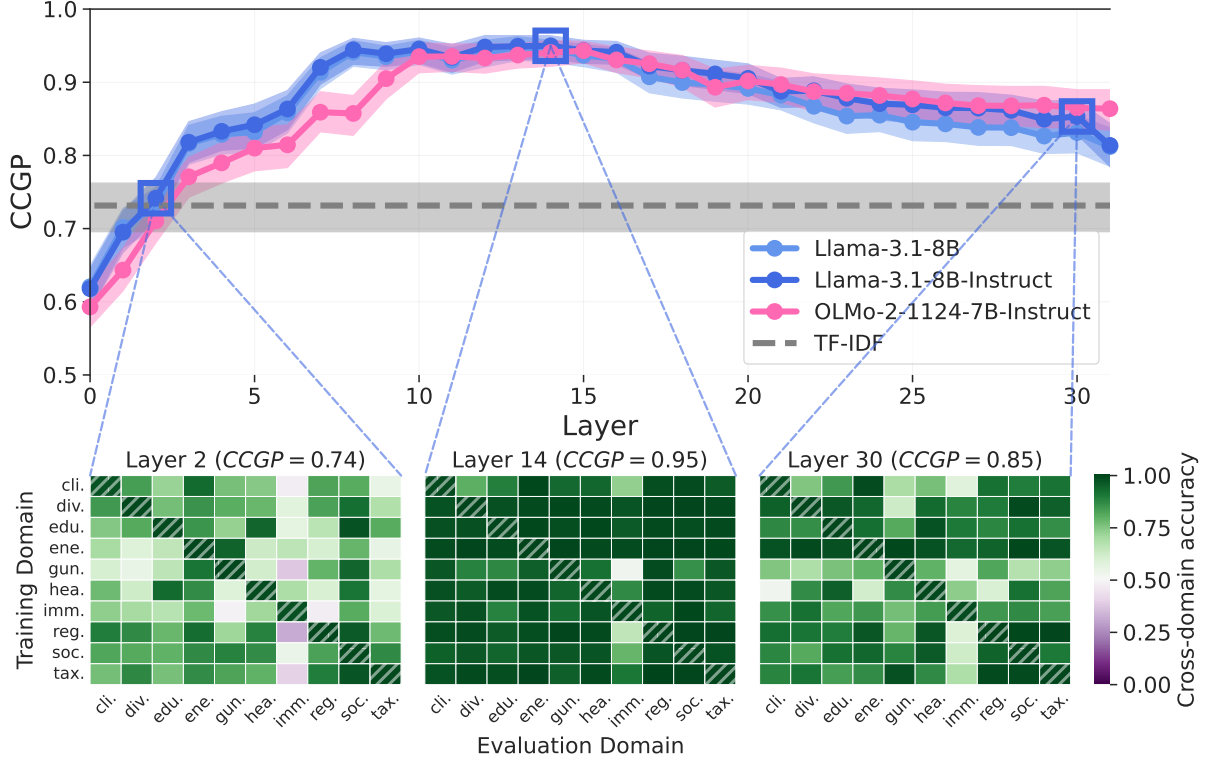


Figure 3: CCGP, or Cross-Condition Generalization Performance, against layer depth with 95% bootstrapped confidence intervals. Detailed cross-condition performance for Llama-3.1-8B-Instruct are presented at specific layers. Diagonal elements represent within-domain accuracy, evaluated on holdout cross-validated folds (used to compute WCP); off-diagonal elements are cross-domain performance, evaluated on the whole domain-constrained data subsets. Additional detailed results for all models at every layers are provided in [Appendix D](#).

to work as lexical build-up. Here, in the first layers, CCGP steadily increases from $\sim 60\%$ at the input, until reaching values above 90% by the tenth layer. This pattern aligns with semantic abstraction by the model, here resulting in high domain generalisation, likely indicating that shared ideological axes emerge during processing. Finally, after a short plateau with constantly high values during approximately 7 to 9 consecutive layers, CCGP slowly decays from the late middle layers to the end of the network, as the model prepares for next token generation.

The contrast between lexical performance and high mid-network latent generalisation, with CCGP significantly above the lexical baseline, provides robust empirical evidence that LLMs construct an latent, conceptual geometry representing political ideology.

6 The Geometry of Political Abstraction is Low-Dimensional

The alignment and cross-domain generalization observed in [section 5](#) suggests a high degree of struc-

tural organisation and implies shared geometric arrangements in the latent space. To test that this generalisation is driven by a compression of political concepts rather than decentralized, or high-dimensional layout, the investigation turns to quantifying the effective intrinsic dimensionality of the latent policy representation subspaces.

As plotted on [Figure 4](#), the layerwise trajectory of PR inversely mirrors the CCGP trends detailed in [section 5](#). Note that all observed participation ratios (except for the two first layers of OLMo-2-1124-7B-Instruct) are significantly lower than those obtained under domain-wise label permutation null models (with $p \leq 0.01$), indicating substantially greater alignment of contrast vectors than expected by chance.

In the early layers of every tested models, the PR sits at a relatively high value (≈ 7). This indicates that the models initially treat political stances within highly distinct, domain-bounded subspaces. Yet, as the information flows to the middle layers of the networks, PR continuously decreases until reaching global minima with $PR \leq 3$ in the

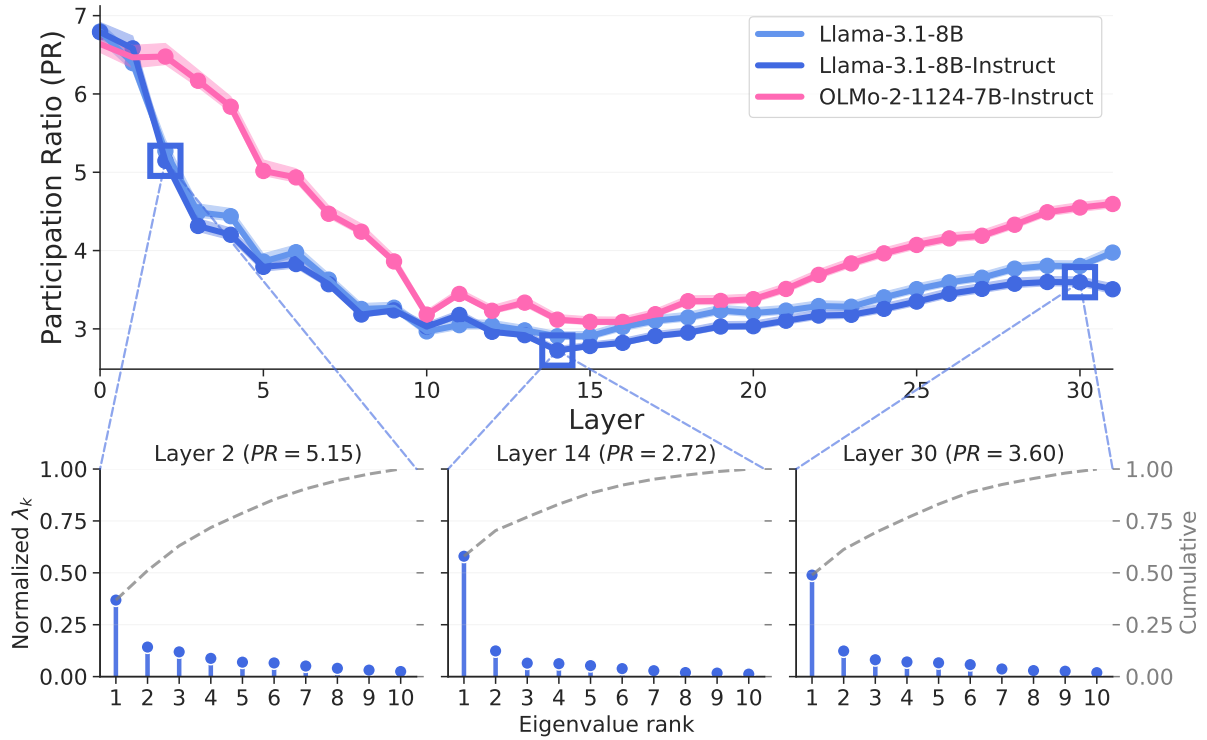


Figure 4: PR, or Participation Ratio, against layer depth with 95% bootstrapped confidence intervals. Detailed normalised eigenvalues spectrum are presented at specific layers for Llama-3.1-8B-Instruct. Detailed cosine similarity scores between contrast vector pairs, for all models, at every layers, are provided in [Appendix D](#).

same range of layers where CCGP peaks. This systematic reduction in effective dimensionality provides additional evidence of structural compression and the abstraction of political information by LLMs. At such layers, spectral analyses of the Gram matrices reveal that the leading eigenvalues λ_1 alone regularly account for 54-58% of total variance across models (over five times the uniform baseline). While this finding confirms that the geometry is anchored by a dominant global political axis, further analyses would be required to characterize the factors preventing PR from reaching 1. In the late layers of the model, PR begins a gradual upward ascent. Once again, this structural decompression aligns with the functional demands of next-token prediction heads. To generate contextualised text, the models need to unpack the low-dimensional ideological abstraction back into a more expressive, domain-rich semantic space. To complement on the relationship with cross-condition generalisation, layerwise PR and CCGP exhibit a strong negative correlation across all models (Pearson $r \approx -0.94$ to -0.95 , $p < 0.001$). As shown in [Figure 5](#), their trajectories reveal a clear co-evolution, with CCGP peaking

precisely as effective dimensionality collapses.

Together, these scores suggest that, through dynamic processing of information, LLMs construe a low-dimensional conceptual representation space that abstracts complex social ideas, while retaining sufficient information to allow for re-contextualisation.

7 Discussion & Conclusion

This work investigates whether and how LLMs represent political ideology as an abstract variable that generalizes across substantively different policy domains. Our study extends geometric approaches to abstraction from systems neuroscience to LLMs, demonstrating that abstract representational geometry can be investigated in LLMs using naturalistic socially grounded concepts instead of experimentally constructed tasks.

Across three open-weight models, political leaning is highly linearly decodable within individual domains from early layers onward ([section 4](#)). However, this local separability is also largely captured by lexical features (TF-IDF baseline), indicating the distinction between the existence of a decodable concept and the presence of an ab-

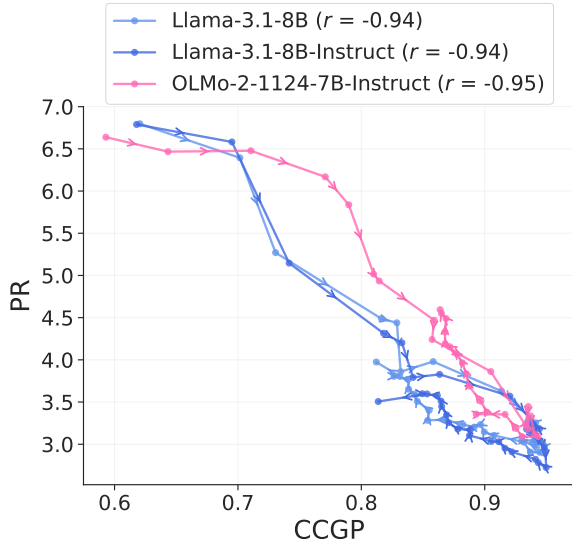


Figure 5: Participation ratio (PR) against cross-condition generalisation performance (CCGP) showing strong negative Pearson correlation r . Arrows show the flow throughout the models layers.

abstract representation. A different pattern emerges in cross-domain evaluation (section 5). Classifiers trained on one policy domain generalize strongly to unseen domains, with performance peaking in the middle layers and substantially exceeding the baseline. This functional generalization coincides with a marked reduction in the participation ratio of domain-specific contrast vectors (section 6), suggesting that policy-specific ideological distinctions become organized within a shared, low-dimensional subspace. In later layers, cross-domain performance decreases while participation ratio rises, consistent with a partial re-expansion into more domain-specific representations.

Together, these findings indicate that LLMs transiently compress heterogeneous political stances into a common representational coordinate during intermediate processing. More broadly, they support the view that abstraction in LLMs is reflected in the geometry that allows a representation learned in one context to transfer to another. This shifts the interpretability question from locating conceptual information to characterizing the representational organization that enables generalization. Political ideology, in this respect, provides a useful naturalistic test case for examining how LLMs structure diverse training data on high-level socially grounded concepts into coherent conceptual structures.

Limitations

Through the use of analytical methods inspired by system neuroscience, the findings described in this paper provide robust evidences of abstract low-dimensional geometry of political ideology in LLMs. However, several limitations bind the scope of this work, while opening directions for future research.

Correlational Methods A key limitation of this study resides in the choice of a correlational, rather than causal, research design. Indeed, WCP, CCGP, and PR demonstrate that a low-dimensional, generalised political subspace is systematically constructed in the tested LLMs’ middle layers. Yet, these analytical tools do not allow to draw any causal link between this specific geometry and the models’ downstream behavior.

Future works could push toward causal claims through intervention techniques, such as activation patching or steering, along the domain contrast vectors, to verify whether manipulating these axes directly alters the political stances of the generated outputs. Further, taking into account the observed high entanglement of political contrast vectors in the models’ middle layers, it would be interesting to study cross-domain steering impacts with respect to the altered layers, or to research decoupling strategies to allow for more granular, domain-specific political steering.

Political Simplification Besides the correlational research design, the experiments rely heavily on the latent representation induced by stances from the TwinViews dataset (Fulay et al., 2024), which comes with conceptual limitations. While the use of a large-scale set of short political statements annotated with policy issue tags and representing opposing political poles, provides a clean methodological framework for our experiments, it may also distort the real nature of the political landscape through oversimplification and stereotypical framing.

First, including only binary political “left-” and “right-” leaning labels introduces a highly truncated, if not deceptive, representation of ideology. While such simplifications are necessary for operationalisation and experimental design, they fail to reproduce the inherent multi-dimensional nature of politics and convey perspectives that might misrepresent even highly polarized political landscapes. This limitation is partially addressed by

cross-coding political leaning and policy issue labels in our experimental setting. By doing so, we consider the “left” and “right” tags primarily as proxies for ‘support’ and ‘oppose’ labels related to specific views on each domain. Then, the cross-domain analyses reveal whether or not the traditional left/right divide encoded in the dataset directly resurfaces from LLMs’ activation space.

Second, the use of heavily marked synthetically generated stances may lead the models to respond to highly stereotypical formulations of US-centric partisan views (sample examples of the dataset stances can be found in [Appendix A’s Table 1](#)). Thus, it would be interesting to replicate the analytical pipeline using real-world politically loaded texts, also extending to other horizons than solely US politics.

Technical Limitations Finally, the scope of the study is also limited by the granularity used for analysis and the selection of models.

The representation extraction pipeline is limited to the residual stream of the model, pooled at the final token. While it acts as the information highway of the model, this coarse level of analysis hinders more granular conclusions. A complete mechanistic account for model processing of information should likely include other units of analysis (such as multi-layer perceptron sublayers, attention heads, or circuits), so that it could also isolate the specific contributions of the different components.

Also, this paper includes results obtained with standard, widely adopted mid-sized open-weight LLMs. Yet, the findings cannot be confidently generalised to the whole pool of available LLMs. Besides the extension to other models of comparable size and architecture, it remains an open question how scale or routing paradigms actually affect the ‘abstraction’ of such concepts. For instance, it could be hypothesized that larger models possess the capacity to retain highly nuanced, disentangled representations of individual policy issue polarisation throughout their layers.

Acknowledgements

Parts of this research were carried within the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 945304 - Cofund AI4the-Sciences hosted by PSL University.

This work benefited from funding from the French State, managed by the Agence Nationale

de la Recherche, under the France 2030 program (grant reference ANR-23-IACL-0008).

SD was supported by a grant from IEAI-TUM.

References

- Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. [Measuring Political Bias in Large Language Models: What Is Said and How It Is Said](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.
- Yonatan Belinkov. 2022. [Probing Classifiers: Promises, Shortcomings, and Advances](#). *Computational Linguistics*, 48(1):207–219.
- Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. 2020. [The geometry of abstraction in the hippocampus and prefrontal cortex](#). *Cell*, 183(4):954–967.e21.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, and Amanda Askell. 2023. [Towards monosemanticity: Decomposing language models with dictionary learning](#). *Transformer Circuits Thread*, 2(5):6.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language Models are Few-Shot Learners](#). *arXiv preprint. ArXiv:2005.14165 [cs.CL]*.
- SueYeon Chung and L. F. Abbott. 2021. [Neural population geometry: An approach for understanding biological and artificial neural networks](#). *Current Opinion in Neurobiology*, 70:137–144.
- Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. [Towards Automated Circuit Discovery for Mechanistic Interpretability](#). *arXiv preprint. ArXiv:2304.14997 [cs.LG]*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#). *arXiv preprint. ArXiv:1911.02116 [cs.CL]*.
- Hristos S. Courellis, Juri Minxha, Araceli R. Cardenas, Daniel L. Kimmel, Chrystal M. Reed, Taufik A. Valiante, C. Daniel Salzman, Adam N. Mamelak, Stefano Fusi, and Ueli Rutishauser. 2024. [Abstract representations emerge in human hippocampal neurons during inference](#). *Nature*, 632(8026):841–849.

- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. [Toy Models of Superposition](#). *arXiv preprint*. ArXiv:2209.10652 [cs.LG].
- Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. [Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704, Vienna, Austria. Association for Computational Linguistics.
- Stanley Feldman and Christopher Johnston. 2014. [Understanding the Determinants of Political Ideology: Implications of Structural Complexity](#). *Political Psychology*, 35(3):337–358. [eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pops.12055](#).
- Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayana, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*. Association for Computational Linguistics. ArXiv:2409.05283.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. 2025. [Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability](#). *arXiv preprint*. ArXiv:2301.04709 [cs.AI].
- 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](#). *Preprint*, arXiv:2407.21783.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Wes Gurnee and Max Tegmark. 2024. [Language Models Represent Space and Time](#). *arXiv preprint*. ArXiv:2310.02207 [cs.LG].
- Graeme S. Halford, William H. Wilson, and Steven Phillips. 2010. [Relational knowledge: the foundation of higher cognition](#). *Trends in Cognitive Sciences*, 14(11):497–505.
- Charles Kemp and Alan Jern. 2009. [Abstraction and Relational learning](#). In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.
- Junsol Kim, James Evans, and Aaron Schein. 2025. [Linear Representations of Political Perspective Emerge in Large Language Models](#). *arXiv preprint*. ArXiv:2503.02080 [cs].
- Nikolaus Kriegeskorte and Rogier A. Kievit. 2013. [Representational geometry: integrating cognition, computation, and the brain](#). *Trends in Cognitive Sciences*, 17(8):401–412.
- Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.
- Gaël Le Mens and Aina Gallego. 2025. [Positioning political texts with large language models by asking and averaging](#). *Political Analysis*, 33(3):274282.
- Matthew Levendusky. 2009. *The partisan sort: How liberals became Democrats and conservatives became Republicans*. University of Chicago Press.
- Samuel Marks and Max Tegmark. 2024. [The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets](#). *arXiv preprint*. ArXiv:2310.06824 [cs].
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. [2 olmo 2 furious](#).
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Jonathan B. Slapin and Sven-Oliver Proksch. 2008. [A scaling model for estimating time-series party positions from texts](#). *American Journal of Political Science*, 52(3):705–722.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Calum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, and 7 others. 2026. [Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet](#). *arXiv preprint*. ArXiv:2605.29358 [cs.AI].
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. [Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small](#). *arXiv preprint*. ArXiv:2211.00593 [cs.LG].

A Data Preparation

This section describes the preparation of the ‘domain’ labels in [subsection A.1](#), and gives concrete examples of the political stances included in the dataset in [subsection A.2](#).

A.1 Policy Issue Domain Labels

The original format contains 1024 unique topics (spread across the 14K pairs), which tend to be redundant and lexically noisy. They include same topics under different formatting (e.g; ‘Climate Change’, ‘Climate change’, ‘Climate Change Policies’, ‘Climate change policy’), synonyms and paraphrases (e.g; ‘Gun Regulations’, ‘Firearm Regulations’), inconsistent (and sensitive) granularity (e.g; ‘Education’, ‘Arts in Education’, ...), etc.

To address those, the original labels are merged into more general categories. This procedure is done automatically and relies on the clustering of the textual inputs using BERTopic ([Grootendorst, 2022](#)). This unsupervised pass on the data allows to retrieve a reduced number of clusters (here 52) based on the inferred topics. The original topic tags are then hierarchically merged into these categories while ensuring that none of the original topics are splitted into different clusters. Finally, only the ten most common clusters are selected, filtering down the data to 7886 statements pairs.

The number of stance pairs per grouped domain, as well as the five most common original domain tags in the clusters are listed below:

- **tax** (1070): Taxation (536), Income Inequality (195), Economy (35), Corporate Taxes (27), Wealth Inequality (25), ...
- **regulation** (1043): Government Regulations (175), Government Regulation (113), Financial Regulation (93), Environmental Regulations (83), Regulation (77), ...
- **education** (945): Education (529), Higher Education (170), Education Funding (89), College Education (38), Public Education (20), ...
- **gun control** (900): Gun Control (787), Gun control (56), Gun Control Measures (11), Gun Control Laws (9), Gun Rights (7), ...
- **energy** (791): Renewable Energy (502), Energy (49), Energy Policy (45), Clean Energy (33), Renewable energy (31), ...

- **healthcare** (723): Healthcare (637), Universal Healthcare (27), Healthcare System (12), Healthcare Systems (7), Healthcare Access (6), ...
- **diversity and inclusion** (674): Affirmative Action (383), Diversity and Inclusion (35), Social Justice (33), Diversity (32), Affirmative action (27), ...
- **immigration** (623): Immigration (548), Immigration Reform (65), Immigration Policy (5), Immigration policy (5), Immigration Policies (5), ...
- **social welfare** (588): Social Welfare (171), Social Programs (127), Welfare (83), Social Safety Net (71), Social Safety Nets (29), ...
- **climate change** (509): Climate Change (390), Environment (79), Climate change (25), Environmental Policy (6), Environmental Issues (5), ...

Note that while this unsupervised procedure reduces noise at the topic-tag level and allows to reduce it to an analysable amount, it may introduce additional within-topic noise and group together statements that may be less directly semantically consistent compared to the original taxonomy.

A.2 Examples

Randomly selected stances from the dataset for each policy issue domain and opposing views are provided in [Table 1](#).

B Computational Details

All computational experiments were executed on a workstation equipped with two NVIDIA RTX 6000 Ada Generation GPUs, each featuring 48 GB of VRAM. During the representation extraction process, models were 8-bit quantized and inference was run on batches, with 32 samples per batch. To accommodate varying sequence lengths within each batch, sequences were dynamically padded to the maximum length of the batch, with attention masks applied to ensure padding tokens did not influence the latent representations of the texts.

C Discussion on Parallelism Score and Participation Ratio

This section discusses our choice of using the participation ratio rather than proper parallelism scores

Domain	Left-leaning stance	Right-leaning stance
climate change	<i>Climate change is a serious threat that requires immediate action to reduce carbon emissions.</i>	<i>Climate change is a natural occurrence and should not be a major concern.</i>
diversity and inclusion	<i>Racial and ethnic diversity should be embraced and celebrated in all aspects of society.</i>	<i>Affirmative action is reverse discrimination and undermines the principles of meritocracy.</i>
education	<i>Higher education should be more affordable and accessible to all, with reduced tuition fees and increased financial aid.</i>	<i>The free market should drive education, and school choice and competition would lead to better outcomes for students.</i>
energy	<i>The government should invest in renewable energy sources to reduce dependence on fossil fuels.</i>	<i>The market should determine the viability of renewable energy without government intervention.</i>
gun control	<i>Gun control measures are necessary to prevent mass shootings and ensure public safety.</i>	<i>Gun ownership is a constitutional right and any regulations infringe upon that right.</i>
healthcare	<i>The government should provide free or subsidized healthcare for low-income individuals and families.</i>	<i>Private healthcare system allows for competitive market and better quality of care.</i>
immigration	<i>Comprehensive immigration reform is necessary to provide a pathway to citizenship for undocumented immigrants and address the broken immigration system.</i>	<i>Strict immigration policies are necessary to protect jobs and national security.</i>
regulation	<i>Protecting the environment should be a top priority, even if it means imposing stricter regulations.</i>	<i>Internet service providers should have the freedom to offer different service plans and speeds.</i>
social welfare	<i>Social safety net programs are essential to support the most vulnerable members of society and reduce poverty.</i>	<i>Reducing government intervention and promoting self-reliance promotes individual responsibility and economic growth.</i>
tax	<i>Higher taxes on the wealthy can help reduce income inequality.</i>	<i>Lower corporate tax rates stimulate economic growth and encourage investment and job creation.</i>

Table 1: Random samples for each of the policy issue domains and opposing views.

in [section 6](#). First, we revisit the original definition of PS and its relation to the goals of our experiments in [subsection C.1](#). Then we establish the formal mathematical links between PR and an alternative, mean-based, definition of PS in [subsection C.2](#).

C.1 Maximum permutation- and mean-defined parallelism scores

Originally, the parallelism score is defined as a maximization of cosine similarities over all possible label permutations ([Bernardi et al., 2020](#)). This

definition positions the parallelism score as an exploratory tool to search for parallel directions.

Formally, this quantity can be expressed as follows. Let S_K be the set of all possible bijective transformations (permutations) between domain indices $\{1, \dots, K\}$. Also, consistently with our notations from [Equation 3](#), let’s define normalised cross-domain contrast vectors as:

$$v_{i \rightarrow j} = \frac{\mu_{i,r} - \mu_{j,l}}{\|\mu_{i,r} - \mu_{j,l}\|}$$

Then, the average cosine similarity between cross-domain contrast vectors for a given permutation

$\sigma \in S_K$ is:

$$\langle \cos \theta_\sigma \rangle = \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{j \neq i} v_{i \rightarrow \sigma(i)} \cdot v_{j \rightarrow \sigma(j)}$$

Based on these definitions, the parallelism score is defined as the maximised average cosine similarity across all possible permutations:

$$\text{PS}_{\text{max-perm}} = \max_{\sigma \in S_K} \langle \cos \theta_\sigma \rangle. \quad (5)$$

This permutation-maximisation formulation acts as a search engine to discover any abstractly represented variable within the population, without assuming knowing "the 'correct' way of pairing up conditions a priori". However, our experimental setting is closer to hypothesis testing, with a fixed, predefined contrast structure. Because the polarization axes are imposed by the experimental design, considering the whole set of permutations would break the data structure and provide less insightful results to our original research question. Instead, a stricter setting which tests alignment solely between the contrast vectors spanned by the different categories is more meaningful to our hypothesis.

On this basis, a natural candidate would be a mean-defined parallelism score, computed between the within-domain coding vectors of each domain pairs:

$$\text{PS}_{\text{mean}} = \frac{1}{K(K-1)} \sum_{i=1}^K \sum_{j \neq i} v_i \cdot v_j \quad (6)$$

However, while PS_{mean} respects the semantic boundaries of the data domains, it still acts as a binary-adjacent detector of alignment. It is not suited to capture global, effective dimensionality spanned by this polarized latent geometry. Thus, we decided to focus our analysis on the participation ratio, which holds strong mathematical ties with PS_{mean} (denoted simply as PS in the following section to ease readability).

C.2 Mathematical relationship between participation ratio and mean-defined PS

The participation ratio PR (Equation 4) and the mean-defined parallelism score PS (Equation 6) have strong mathematical ties. Indeed, we can show that PR is upper-bounded by an inverse quadratic function of PS, scaling inversely with its square (see below, subsubsection C.2.1). Also, while PR knows no simple lower bound depending

solely on PS (other than the trivial $\text{PR} \geq 1$), it is fully characterised as a function of PS and the its variance σ_{PS}^2 , as described in subsubsection C.2.2.

C.2.1 Upper bound $\text{PR} \leq h(\text{PS})$.

Recall that G is the $K \times K$ Gram matrix of the normalised contrast vectors $\{v_i\}_{i=1}^K$, so that $G_{ij} = v_i \cdot v_j$. As the contrast vectors are normalised, the diagonal elements are $G_{ii} = v_i \cdot v_i = 1 \forall i \in \{1, \dots, K\}$. Thus, $\text{Tr}(G) = K$. Now, let $\{\lambda_k\}_{k=1}^K$ be the eigenvalues of G . Since $\text{Tr}(G) = \sum_{k=1}^K \lambda_k$, PR from Equation 4 can be written as:

$$\text{PR} = \frac{\left(\sum_{k=1}^K \lambda_k\right)^2}{\sum_{k=1}^K \lambda_k^2} = \frac{K^2}{\text{Tr}(G^2)}$$

Since G is a Gram matrix, symmetric by construction, we have $\text{Tr}(G^2) = \sum_{i,j} G_{ij}^2$, which gives:

$$\text{PR} = \frac{K^2}{\sum_{i,j} G_{ij}^2} \quad (7)$$

Now, the denominator can be split between diagonal and off-diagonal elements as:

$$\begin{aligned} \sum_{i,j} G_{ij}^2 &= \sum_{i=1}^K G_{ii}^2 + \sum_{i \neq j} G_{ij}^2 \\ &= K + \sum_{i \neq j} G_{ij}^2 \end{aligned}$$

By the Cauchy-Schwarz inequality, $\sum_{i \neq j} G_{ij}^2$ is bounded by:

$$\sum_{i \neq j} G_{ij}^2 \geq \frac{1}{K(K-1)} \left(\sum_{i \neq j} G_{ij} \right)^2$$

which can be expressed in terms of PS (using Equation 6):

$$\sum_{i \neq j} G_{ij} = K(K-1)\text{PS}$$

Thus,

$$\begin{aligned} \sum_{i \neq j} G_{ij}^2 &\geq \frac{1}{K(K-1)} (K(K-1)\text{PS})^2 \\ \Leftrightarrow \sum_{i \neq j} G_{ij}^2 &\geq K(K-1)\text{PS}^2 \end{aligned}$$

Finally, substituting back into PR definition, an upper bound of PR is given by:

$$\text{PR} \leq \frac{K}{1 + (K-1)\text{PS}^2}$$

Yet, as can be intuitively understood, PR can reach its theoretical minimum value of 1 (when $\sum_{i,j} G_{ij}^2 = K^2 \Rightarrow G_{ij} \in [-1, 1] \forall i, j$), for any value of PS. Hence, considering only PS-dependent boundaries, PR is bounded by:

$$1 \leq \text{PR} \leq \frac{K}{1 + (K - 1)\text{PS}^2} \quad (8)$$

Nonetheless, by taking into account not only the first mode of the distribution $\{G_{ij}\}_{i \neq j}$ (which is PS), but also its second mode (the variance σ_{PS}^2), PR can be fully described.

C.2.2 PR(PS, σ_{PS}^2)

PR can be completely expressed in terms of the mean and variance of the off-diagonal elements of G , respectively PS and σ_{PS}^2 , providing additional intuition as to how this quantity varies depending on the distribution of cross-condition contrast vectors' similarities $\{G_{ij}\}_{i \neq j}$.

Indeed, by definition the variance is given by:

$$\begin{aligned} \sigma_{\text{PS}}^2 &= \left(\frac{1}{K(K-1)} \sum_{i \neq j} G_{ij}^2 \right) - \text{PS}^2 \\ \Rightarrow \sum_{i \neq j} G_{ij}^2 &= K(K-1)(\sigma_{\text{PS}}^2 + \text{PS}^2) \end{aligned}$$

Plugging it back into PR formula from Equation 7, we have:

$$\text{PR} = \frac{K}{1 + (K-1)(\sigma_{\text{PS}}^2 + \text{PS}^2)} \quad (9)$$

with $\sigma_{\text{PS}}^2 \leq 1 - \text{PS}^2$ and $\text{PS} \in [-1, 1]$.

The formula of PR from Equation 9 highlights that PR decreases both when $|\text{PS}|$ or σ_{PS}^2 increase (while keeping the other value constant).

This relation shows, on one hand, that PR reaches its minimal value of 1 when $(\sigma_{\text{PS}}^2 + \text{PS}^2) = 1$. This is verified when the variance is maximized under the PS-constraint ($\sigma_{\text{PS}}^2 = 1 - \text{PS}^2$), meaning that contrast vectors are highly polarized (some perfectly parallel, some perfectly antiparallel).

On the other hand, it reaches its maximal value of $\text{PR} = K$ when $(\sigma_{\text{PS}}^2 + \text{PS}^2) = 0 \Rightarrow \text{PS} = 0$ and $\sigma_{\text{PS}}^2 = 0$, meaning that all contrast vectors are perfectly orthogonal from one another: $v_i \cdot v_j = 0 \forall i \neq j$.

For additional visual insights, the PR curve as a function of both PS and its variance is represented on Figure 6.

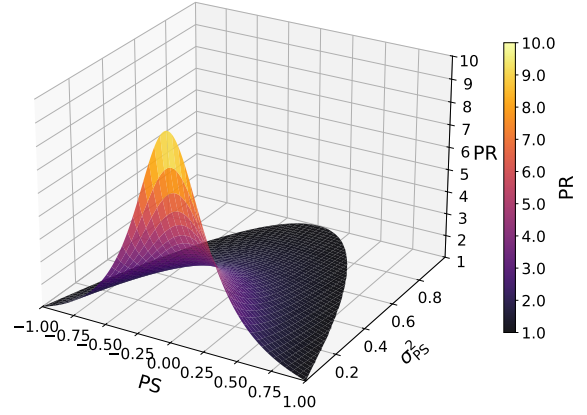


Figure 6: Participation ratio as a function of mean parallelism score and its variance.

D Detailed Results

This section provides detailed results for the different experiments and metrics used in the core of the paper.

The subsection D.1 presents domain-specific performance, related to the calculation of both WCP and CCGP.

The subsection D.2 showcases similarity scores between domains' contrast vectors, used in the computation of PR.

D.1 Cross-condition performance

The detailed cross-condition performance for the TF-IDF comparison baseline is shown in Figure 7.

The detailed matrices of cross-condition performance across layers for Llama-3.1-8B, Llama-3.1-8B-Instruct, and OLMo-2-1124-7B-Instruct are displayed respectively in Figure 8, Figure 9, Figure 10.

Note that the diagonal elements of the displayed matrices are within-domain accuracy. Contrary to off-diagonal elements that are cross-domain performance evaluated on the whole set of domain-constrained evaluation data, the performance within domain is the average of the accuracy obtained on the holdout samples of cross-validated folds. This methodological difference is highlighted on the plots through hatches on the diagonal within-domain scores.

D.2 Particiation ratio (PR)

The detailed Gram matrices of contrast vectors' similarities (used to compute PR) across layers for Llama-3.1-8B, Llama-3.1-8B-Instruct,

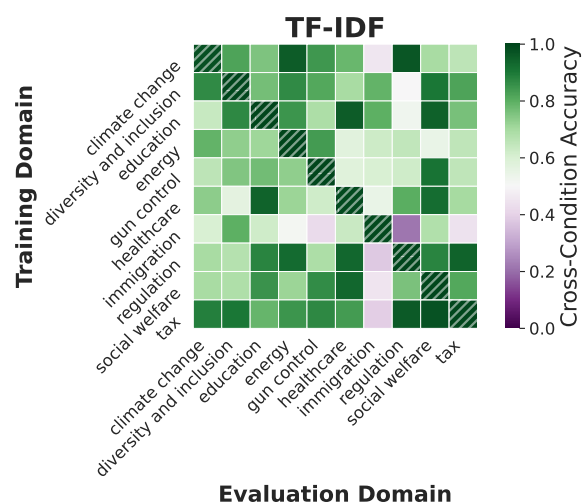


Figure 7: Detailed cross-domain accuracy for the TF-IDF model.

and OLMo-2-1124-7B-Instruct are displayed respectively in [Figure 11](#), [Figure 12](#), [Figure 13](#).

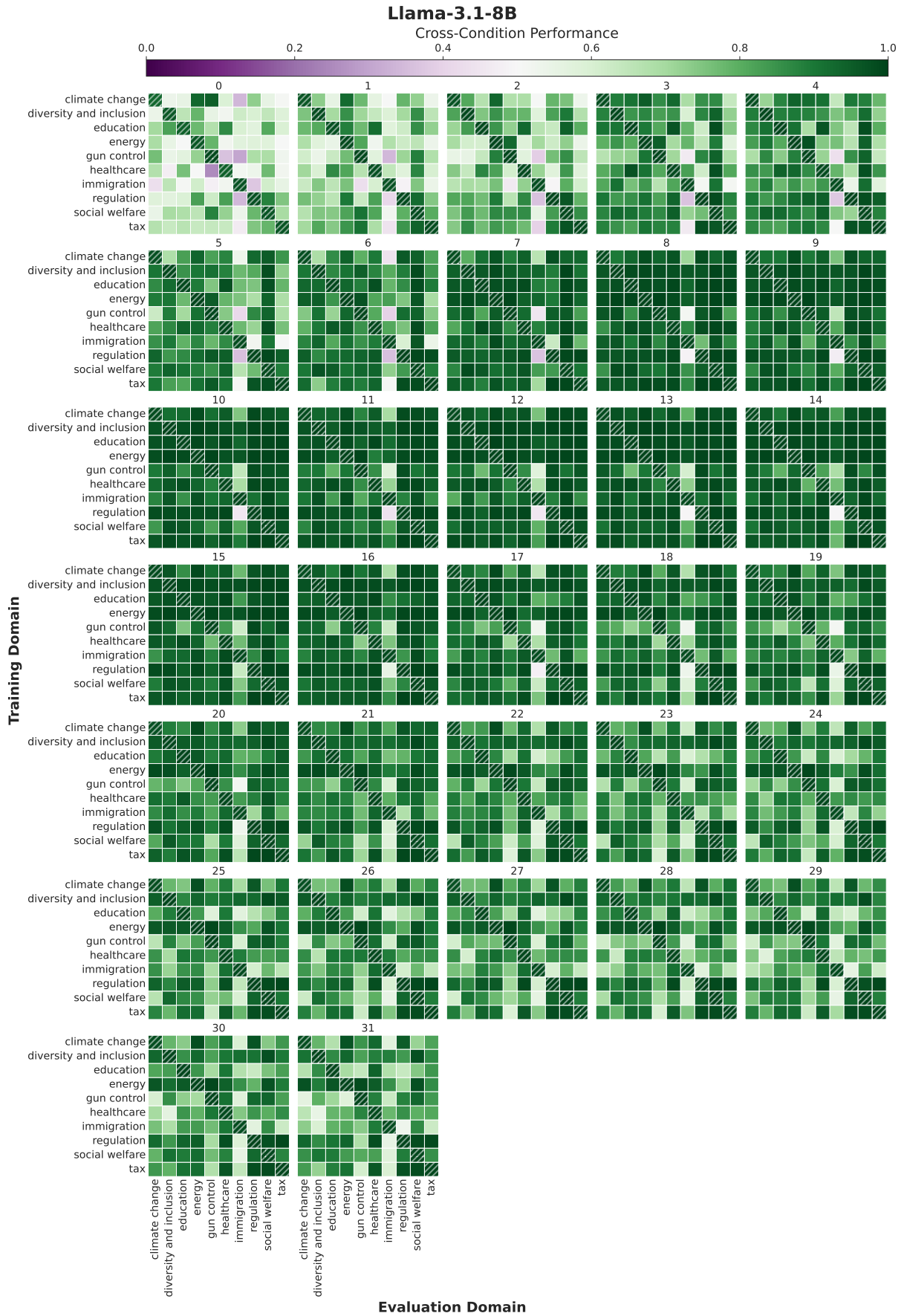


Figure 8: Detailed cross-domain accuracy across layers for Llama-3.1-8B. Off-diagonal, cross-domain performance are evaluated on the entirety of the test domain data, hatched diagonal elements disclose within domain performance evaluated on 20% uphold samples.

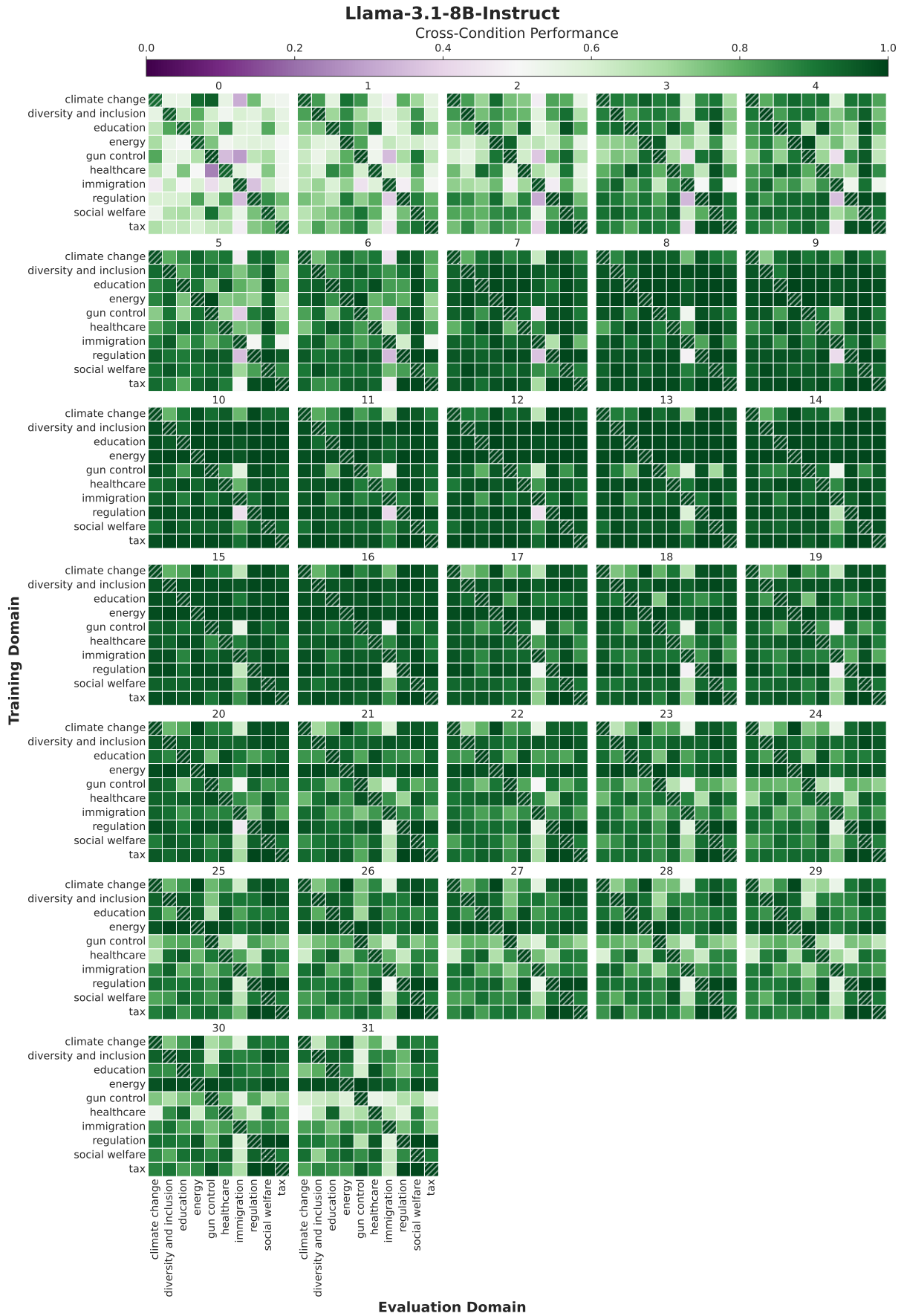


Figure 9: Detailed cross-domain accuracy across layers for Llama-3.1-8B-Instruct. Off-diagonal, cross-domain performance are evaluated on the entirety of the test domain data, hatched diagonal elements disclose within domain performance evaluated on 20% uphold samples.

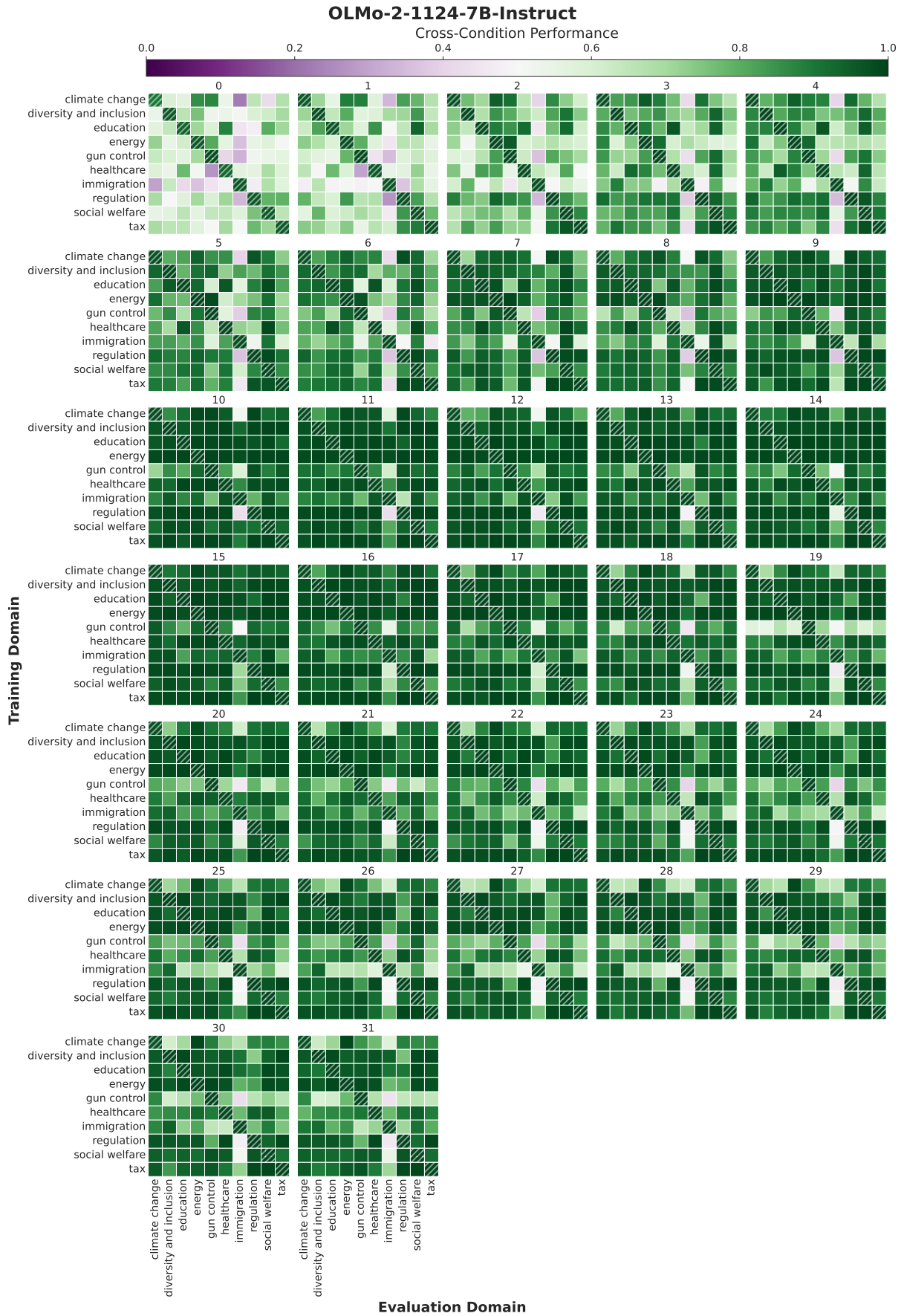


Figure 10: Detailed cross-domain accuracy across layers for OLMo-2-1124-7B-Instruct. Off-diagonal, cross-domain performance are evaluated on the entirety of the test domain data, hatched diagonal elements disclose within domain performance evaluated on 20% uphold samples.

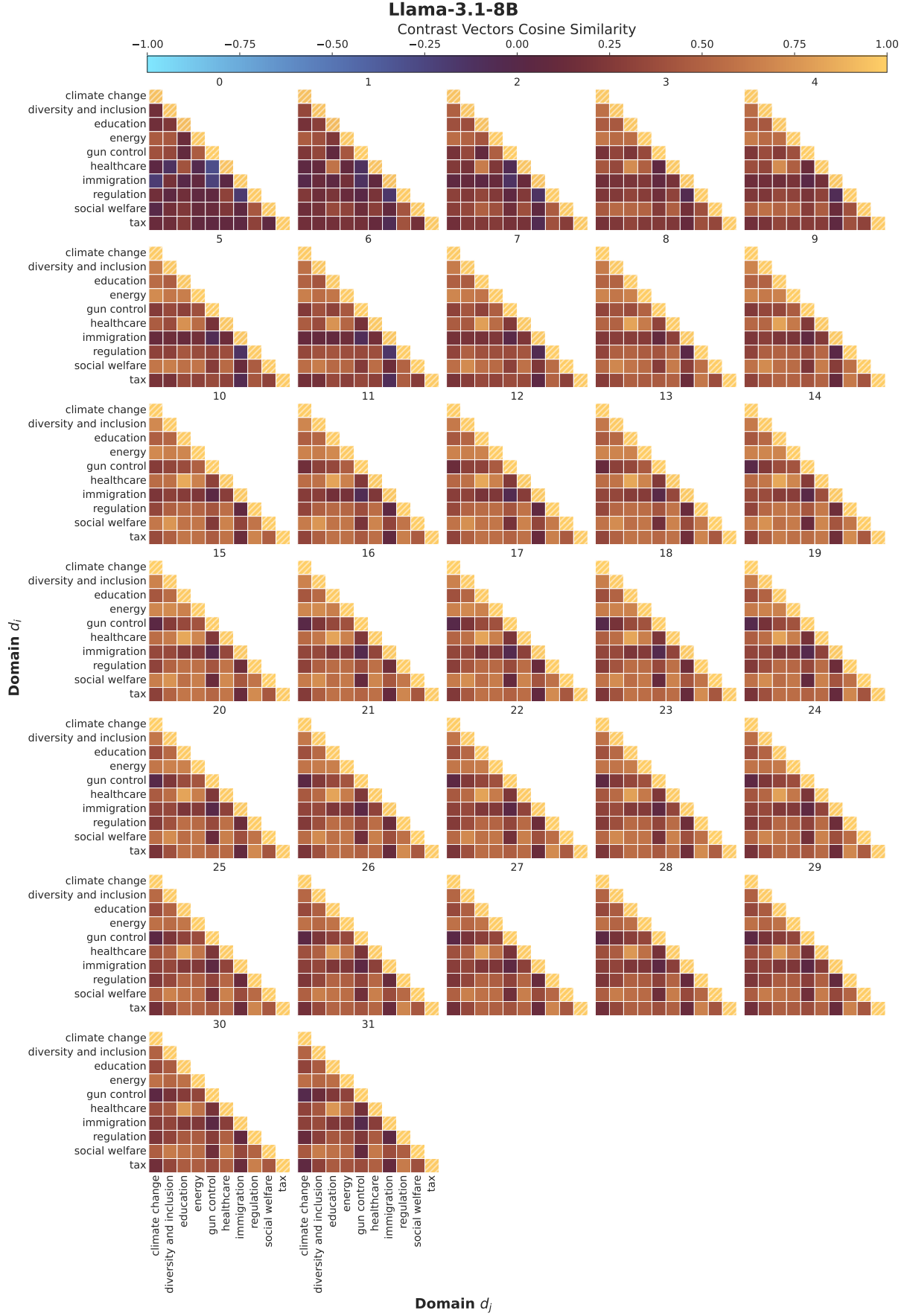


Figure 11: Detailed contrast vectors cosine similarities across layers for Llama-3.1-8B.

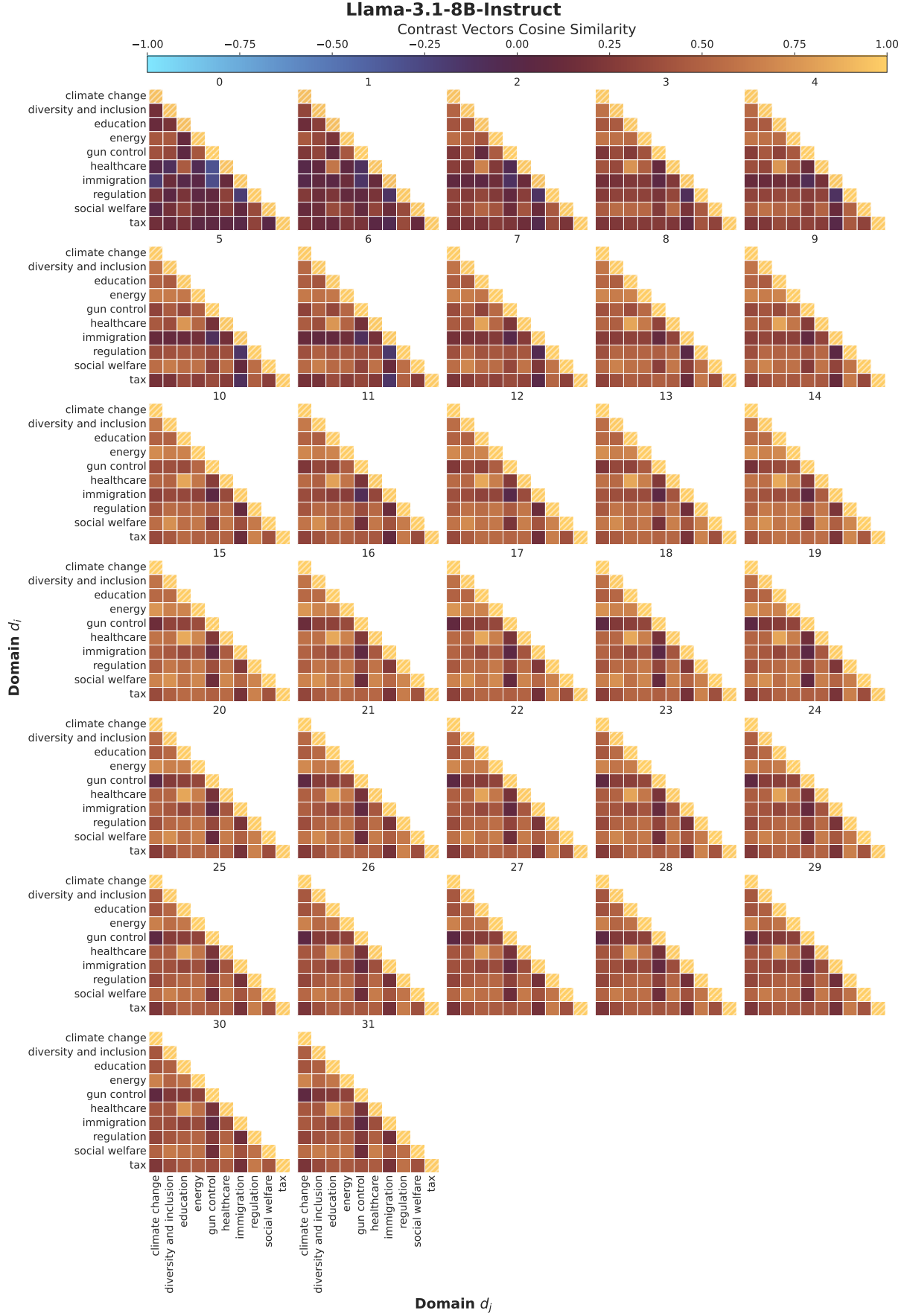


Figure 12: Detailed contrast vectors cosine similarities across layers for Llama-3.1-8B-Instruct.

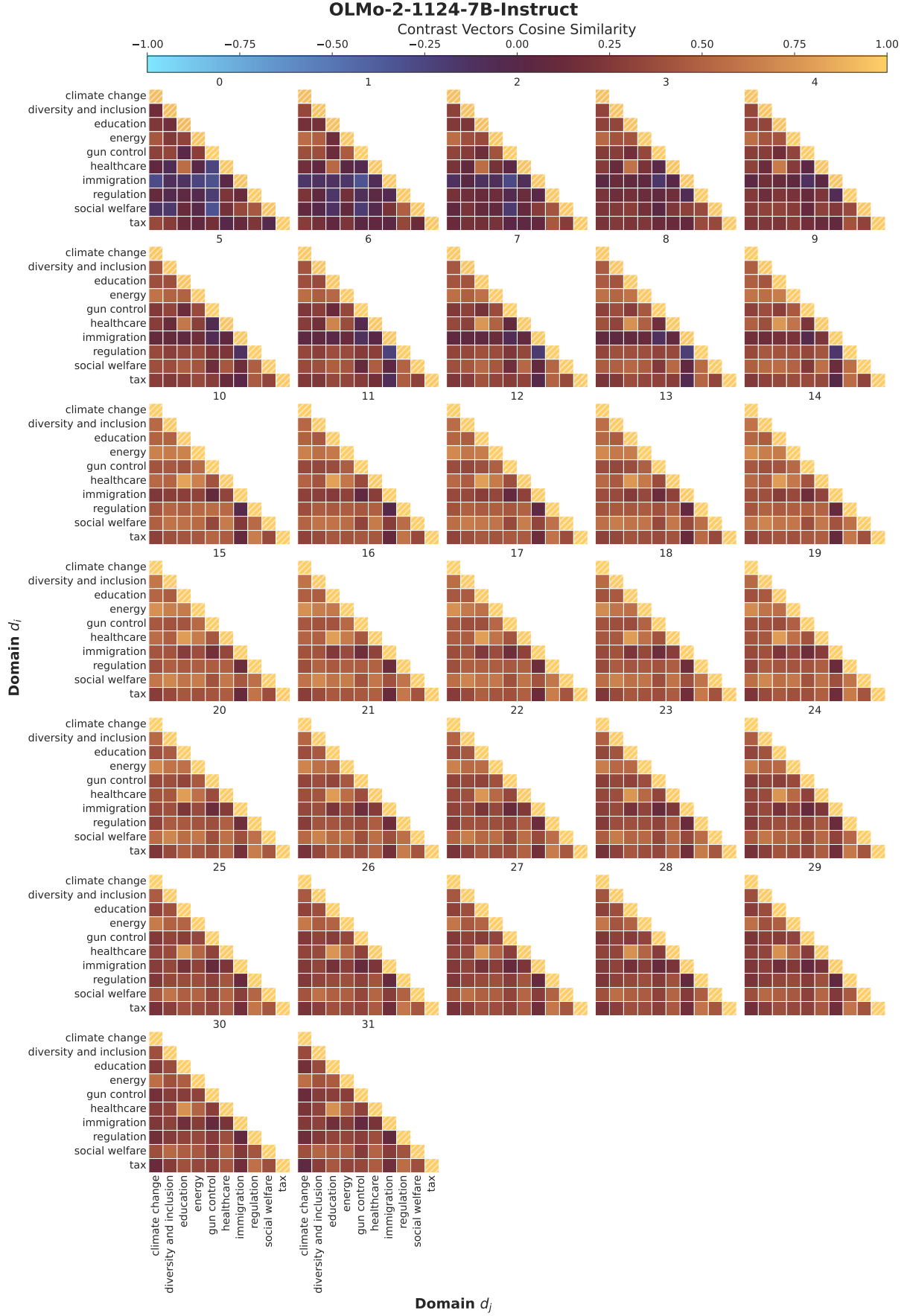


Figure 13: Detailed contrast vectors cosine similarities across layers for OLMo-2-1124-7B-Instruct.