Characterizing stable regions in the residual stream of LLMs

Jett Janiak
LASR Labs
jettjaniak@
gmail.com

Jacek Karwowski[*]
University of Oxford
jacek.karwowski@
cs.ox.ac.uk

Chatrik Singh Mangat[*]
LASR Labs
chatrikmangat@
outlook.com

Giorgi Giglemiani
LASR Labs
giglemiani@
gmail.com

Nora Petrova
LASR Labs
nora.axion@
gmail.com

Stefan Heimersheim
Apollo Research
stefan@
apolloresearch.ai

Abstract

We identify *stable regions* in the residual stream of Transformers, where the model's output remains insensitive to small activation changes, but exhibits high sensitivity at region boundaries. These regions emerge during training and become more defined as training progresses or model size increases. The regions appear to be much larger than previously studied polytopes. Our analysis suggests that these stable regions align with semantic distinctions, where similar prompts cluster within regions, and activations from the same region lead to similar next token predictions. This work provides a promising research direction for understanding the complexity of neural networks, shedding light on training dynamics, and advancing interpretability.

1 Introduction

We study the effects of perturbing Transformer activations, building upon previous work [6, 9, 12]. Specifically, we interpolate between different residual stream activations after the first layer, and measure the change in the model output. Our initial results suggest that:

1. The residual stream of a trained Transformer can be divided into *stable regions*. Within these regions, small changes in model activations lead to minimal changes in output. However, at region boundaries, small changes can lead to significant output differences.
2. These regions emerge during training and evolve with model scale. Randomly initialized models do not exhibit these stable regions, but as training progresses or model size increases, the boundaries between regions become sharper.
3. These stable regions appear to correspond to *semantic distinctions*. Dissimilar prompts occupy different regions, and activations from different regions produce different next token predictions.
4. These stable regions appear to be much larger than previously studied polytopes [2, 7, 13]. Our analysis of gate activations shows that while interpolating between two prompts typically crosses only one stable region boundary, hundreds or thousands of gates change their activation sign (Appendix F).
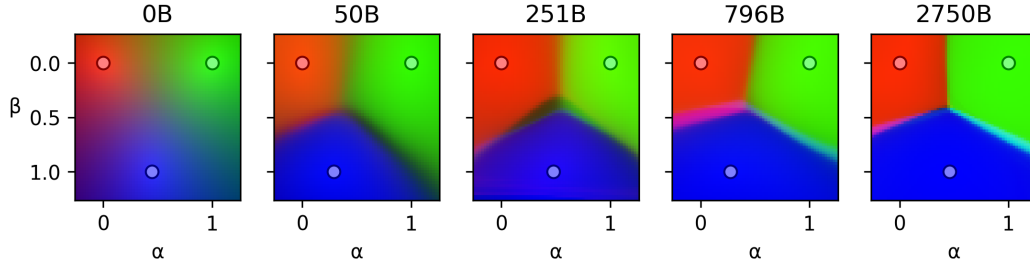
---

[*]equal contribution

Figure 1: Visualization of stable regions in `OLMo-7B` during training. Colors represent the similarity of model outputs to those produced by three model-generated activations (red, green, blue circles). Each subplot shows a 2D slice of the residual stream after the first layer at different stages of training, with the number of processed tokens indicated in the titles. As training progresses from left to right, distinct regions of solid color emerge and the boundaries between them sharpen. Refer to the end of Section 4 for details.

2 Related work

Previous work on characterizing the complexity of neural networks includes studying linear regions (aka polytopes) in piecewise linear networks [2, 7, 13] and bounding the VC dimension of network architectures [1, 8]. While these approaches provide theoretical insights, the extent to which they describe the practical expressivity of networks remains unclear. In this work, we study stable regions in the residual stream of Transformers that can be mapped to classes of very similar outputs. Our initial results suggest that stable regions are much larger than previously studies polytopes.

The linear representation hypothesis [3, 15, 16] suggests that features in neural networks correspond to directions in activation space. Since residual stream activations can be though of as sums of features, we can similarly think about model-generated activations as vectors. Due to normalization layers in Transformers, these vectors lie closely to a hypersphere in the residual stream. Because of high dimensionality of this space, most linear interpolations between two activations will stay close to the hypersphere.

Heimersheim & Mendel [9], motivated by theoretical results by Hänni et al. [10], studied robustness to noise in GPT-2 by interpolating between two residual stream activations and they observed nonlinear changes in the model output. Follow up studies by Giglemiani et al. [4], Lee & Heimersheim [11] connect this to SAEs work by Gurnee [6], Lindsey [12].

3 Methods

For a given prompt p, we define a Transformer forward pass with activation patching as a function $F_p : \mathbb{R}^D \to \mathbb{R}^D$ that returns residual stream activations after last layer[1], at last sequence position. Specifically, $F_p(X)$ replaces the residual stream activations after the first[2] layer, at the last sequence position, with X, before applying the rest of the model layers.

If small changes in model activations lead to minimal changes in output, then the residual stream after the first layer can be divided into stable regions. In this case, F_p's rate of change should be small away from region boundaries and large when approaching and crossing region boundaries.

To study this efficiently, we interpolate between prompts p_A and p_B, which may or may not belong to the same region. We introduce an interpolation coefficient $\alpha \in [0, 1]$, where $\alpha = 0$ corresponds to prompt p_A and $\alpha = 1$ corresponds to prompt p_B.

We measure the L_2 distance $d(\alpha) : [0, 1] \to \mathbb{R}$ between two outputs: a clean run on prompt p_A and a patched run where activations are modified depending on α. In the patched run, we replace the activations with a linear interpolation $A + \alpha(B - A)$, where A and B represent the model-generated activations after the first layer, at the last sequence position, corresponding to prompts p_A and p_B

---

[1]Before the unembedding and the softmax.

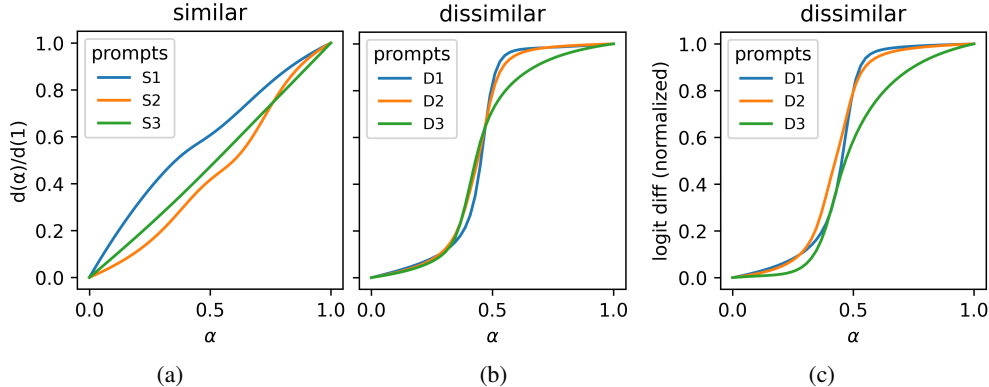[2]Preliminary results in Appendix E suggest that some of the reported effects hold for later layers.

Figure 2: (a,b) Relative output distance as a function of α for (a) similar and (b) dissimilar pairs of prompts in `Qwen2-0.5B`. (c) Normalized logit difference between top prediction for p_B and p_A.

respectively. Formally, this distance is expressed as:

$$d(\alpha) = \|F_{p_A}(A) - F_{p_A}(A + \alpha(B - A))\|_2.$$

In our experiments, we use models from `Qwen2` [17] and `OLMo` [5] families with a number of parameters ranging from 0.5B to 7B, see Appendix A for details. We compute the distance d for 50 uniformly spaced values of α between 0 and 1.

4 Experiments

Illustrative examples We suspect that semantically similar prompts are more likely to belong to the same stable region than semantically different prompts. To test this, we construct pairs of prompts dissimilar to each other numbered D1-D3, and pairs of similar prompts numbered S1-S3. We present prompts D1 and S1 below, and the remaining prompts in Appendix B. For simplicity, the prompts differ only in the last token.
Dissimilar prompts D1:
p_A = " The house at the end of the street was very", top prediction = " quiet"
p_B = " The house at the end of the street was in", top prediction = " a"
Similar prompts S1:
p_A = " She opened the dusty book and a cloud of mist"
p_B = " She opened the dusty book and a cloud of dust"

In Figure 2 we plot the relative distance $\frac{d(\alpha)}{d(1)}$ in model outputs as we interpolate between two prompts in `Qwen2-0.5B`[3]. In Figure 2a, we show the results for interpolation between similar pairs S1-S3, and in Figure 2b for dissimilar pairs D1-D3. We observe that the shapes for similar prompts S1-S3 are close to linear, or equivalently that the rate of change is similar throughout the interpolation. This is in contrast to results for dissimilar pairs D1-D3, where we observe flat regions at the beginning and at the end of the interpolations, separated by a sharp jump, which is in line with what Heimersheim & Mendel [9] observed in `GPT-2`. This relationship between semantic similarity and stable regions is further supported by a large-scale analysis of prompt pairs filtered by their maximum sensitivity (Appendix G). Our interpretation of these results is:

1. Activations for dissimilar prompts appear to occupy distinct stable regions. This is evidenced by the non-linear shape of $d(\alpha)$, which shows flat areas near the model-generated activations and a sharp jump between them.
2. Activations for similar prompts likely belong to the same stable region, as indicated by the linear change in $d(\alpha)$..

---

[3]For bigger models from the Qwen2 family, we observe smaller difference between semantically similar and different prompts. We hypothesize that more capable models are able to make more precise predictions, even for very similar prompts.
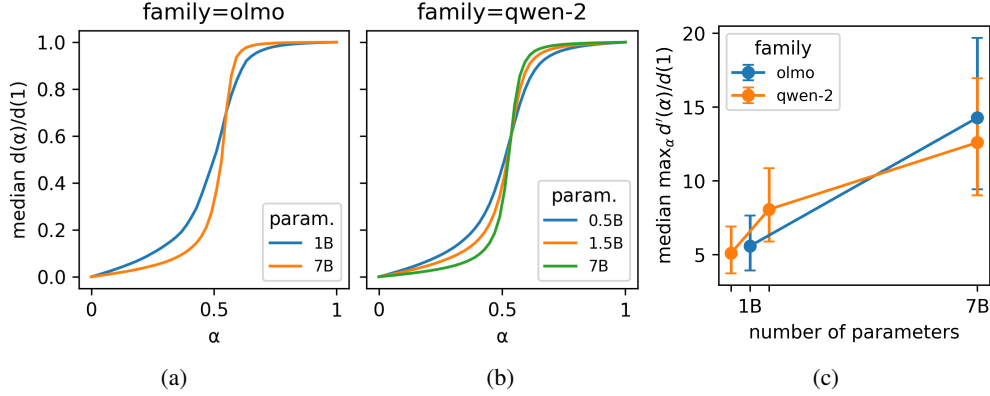
Figure 3: (a,b) Median relative output distance as a function of interpolation coefficient α for different models from the (a) `OLMo` and (b) `Qwen2` families. (c) Maximum slope as a function of the number of parameters for both model families. Dots represent median, and error bars represent 25th and 75th percentiles.

3. Stable regions are large compared to the polytopes described by Black et al. [2]. When interpolating between semantically different prompts, we appear to cross only a single stable region boundary, while hundreds or thousands of gates change their activation sign during the same interpolation (Appendix F).

Additionally, in Figure 2c, we plot the logit difference, normalized to 0-1 range, between top predictions for both prompts in D1-D3 pairs. The change in logit difference follows a similar dynamic to $d(\alpha)$, suggesting that change in $d(\alpha)$ reflects an interpretable change in model predictions.

Impact of model size We want to test if the results we report in the previous section generalize to other prompts and models, and whether the model size plays a role. We sample 1,000 pairs of 10-token-long, unrelated prompts (p_A, p_B) from the `sedthh/gutenberg_english` [14] dataset[4], and we use the methodology described for the previous experiment.

In Figures 3a and 3b we plot the median relative distance for different model sizes from `OLMo` and `Qwen2` families respectively. The curves appear to get sharper with increased model size. We quantify this sharpening effect in Figure 3c by plotting the median of maximum slope of the relative distance, which confirms the qualitative observation.

Impact of training progress We repeat the experiments from the previous section, this time varying the number of training tokens instead of model size. Results presented in Figure 4 reveal several insights. First, curves for randomly initialized models appear close to linear. However, as training progresses, we observe an increasing sharpness in the curves, similar to the effect seen with larger model sizes. This suggests the emergence and refinement of stable regions throughout the training process. Interestingly, the rate of this sharpening effect appears to plateau earlier for smaller `OLMo-1B` model compared to the larger `OLMo-7B`.

At least two different factors may contribute to the sharpening: (1) models develop additional stable regions; (2) the boundaries between existing regions become more defined.[5] To investigate the role of these factors, we examine a 2D slice of the residual stream after the first layer in Figure 1. This slice is spanned by three model-generated activations (A, B, C), corresponding to the red, green, and blue circles respectively. We create synthetic activations within this slice. For each synthetic activation, we compute how similar the model's output is to the outputs produced by A, B, and C. The

---

[4]We chose this popular dataset of English books, as we expect it was used in training of most Transformer LMs. Additionally, we wanted to have a dataset of a limited diversity, and easily understandable next token predictions, but this will only become relevant in future work.

[5]Factor 1 would result in more individual curves being sharp rather than linear, so the aggregated result would appear sharper. Another factor could be that the location of sharp jump is initially spread for different curves, but becomes more concentrated. That would affect the qualitative plots, (a,b) but not the quantitative plot (c).
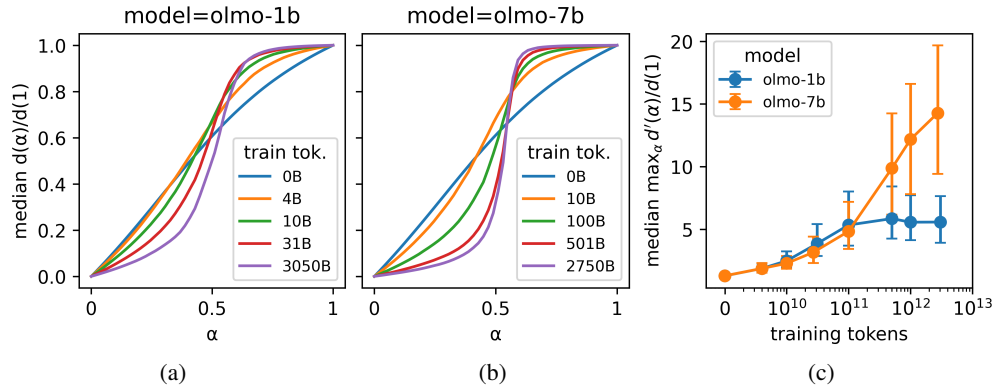
Figure 4: (a,b) Median relative output distance as a function of α for (a) `OLMo-1B` and (b) `OLMo-7B` models. (c) Maximum slope as a function of the number of training tokens for both models. Dots represent median, and error bars represent 25th and 75th percentiles. Note that in subfigures (a) and (b) we only show a few selected checkpoints for readability.

RGB color of each point in the plot represents these similarities: red, green, and blue components correspond to similarities with A, B, and C respectively. See Appendix C for details.

Solid colors indicate stable regions, while color transitions represent boundaries between them. Initially, the plot shows a smooth gradient of colors. As training progresses, distinct regions of solid color emerge, and the boundaries between these regions become increasingly sharp. In Appendix D, we present more slices for different triples of model-generated activations. Some of them hint at existing stable regions splitting during training, in particular Figure 11.

5 Discussion

We only look at a handful of relatively small Transformer language models. We have seen similar results for GPT, Pythia, Phi, and Llama, with a notable exception of Gemma.

We only present indirect evidence for the relatively large size of stable regions, compared to polytopes. Estimating their size or number directly, and how it changes during training and with model size, could provide a more solid ground for this claim.

6 Acknowledgements

References

[1] Bartlett, P. L. and Maass, W. Vapnik-Chervonenkis dimension of neural nets. In Arbib, M. A. (ed.), *The Handbook of Brain Theory and Neural Networks*, pp. 1188–1192. MIT Press, 2003.

[2] Black, S., Sharkey, L., Grinsztajn, L., Winsor, E., Braun, D., Merizian, J., Parker, K., Guevara, C. R., Millidge, B., Alfour, G., and Leahy, C. Interpreting Neural Networks through the Polytope Lens, November 2022. URL http://arxiv.org/abs/2211.12312. arXiv:2211.12312 [cs].

[3] Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy Models of Superposition, September 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs].

[4] Giglemiani, G., Petrova, N., Mangat, C. S., Janiak, J., and Heimersheim, S. Evaluating Synthetic Activations composed of SAE Latents in GPT-2, September 2024. URL `http://arxiv.org/abs/2409.15019`. arXiv:2409.15019.

[5] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. OLMo: Accelerating the Science of Language Models, June 2024. URL `http://arxiv.org/abs/2402.00838`. arXiv:2402.00838 [cs].

[6] Gurnee, W. SAE reconstruction errors are (empirically) pathological, March 2024. URL `https://www.alignmentforum.org/posts/rZPiuFxESMxCDHe4B/sae-reconstruction-errors-are-empirically-pathological`.

[7] Hanin, B. and Rolnick, D. Complexity of Linear Regions in Deep Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2596–2604. PMLR, May 2019. URL `https://proceedings.mlr.press/v97/hanin19a.html`. ISSN: 2640-3498.

[8] Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1064–1068. PMLR, June 2017. URL `https://proceedings.mlr.press/v65/harvey17a.html`. ISSN: 2640-3498.

[9] Heimersheim, S. and Mendel, J. Activation plateaus & sensitive directions in GPT2, July 2024. URL `https://www.lesswrong.com/posts/LajDyGyiyX8DNNsuF/interim-research-report-activation-plateaus-and-sensitive-1`.

[10] Hänni, K., Mendel, J., Vaintrob, D., and Chan, L. Mathematical Models of Computation in Superposition, August 2024. URL `http://arxiv.org/abs/2408.05451`. arXiv:2408.05451 [cs] version: 1.

[11] Lee, D. and Heimersheim, S. Investigating Sensitive Directions in GPT-2: An Improved Baseline and Comparative Analysis of SAEs, September 2024. URL `https://www.lesswrong.com/posts/dS5dSgwaDQRoWdTuu/investigating-sensitive-directions-in-gpt-2-an-improved`.

[12] Lindsey, J. How Strongly do Dictionary Learning Features Influence Model Behavior?, April 2024. URL `https://transformer-circuits.pub/2024/april-update/index.html#ablation-exps`.

[13] Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. On the Number of Linear Regions of Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/hash/109d2dd3608f669ca17920c511c2a41e-Abstract.html`.

[14] Nagyfi, R. sedthh/gutenberg_english, September 2024. URL `https://huggingface.co/datasets/sedthh/gutenberg_english`.

[15] Olah, C. What is a Linear Representation? What is a Multidimensional Feature?, July 2024. URL `https://transformer-circuits.pub/2024/july-update/#linear-representations`.

[16] Smith, L. The 'strong' feature hypothesis could be wrong. August 2024. URL `https://www.alignmentforum.org/posts/tojtPCCRpKLSHBdpn/the-strong-feature-hypothesis-could-be-wrong`.

[17] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu,

T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. Qwen2 Technical Report, September 2024. URL `http://arxiv.org/abs/2407.10671`. arXiv:2407.10671 [cs].

A Comparison of OLMo and Qwen2 model families

Feature	OLMo-1B	OLMo-7B	Qwen2-0.5B	Qwen2-1.5B	Qwen2-7B
Hidden Size	2048	4096	896	1536	3584
# Layers	16	32	24	28	28
# Attention Heads	16	32	14	12	28
# Key/Value Heads	16	32	2	2	4
Intermediate Size	8192	11008	4864	8960	18944
Normalization	Non-parametric layer norm		RMSNorm		
Vocabulary Size	50,304		151,936		
Weight Tying	Yes	No	Yes	Yes	No

Table 1: Comparison of OLMo and Qwen2 model configurations

B Similar and dissimilar prompts

D1:
p_A = " The house at the end of the street was very", top prediction = " quiet"
p_B = " The house at the end of the street was in", top prediction = " a"
D2:
p_A = " He suddenly looked at his watch and realized he was", top prediction = " late"
p_B = " He suddenly looked at his watch and realized he had", top prediction = " been"
D3:
p_A = " And then she picked up the phone to call her", top prediction = " mom"
p_B = " And then she picked up the phone to call him", top prediction = "."
S1:
p_A = " She opened the dusty book and a cloud of mist"
p_B = " She opened the dusty book and a cloud of dust"
S2:
p_A = " In the quiet library, students flipped through pages of"
p_B = " In the quiet library, students flipped through pages in"
S3:
p_A = " The hiker reached the peak and admired the breathtaking"
p_B = " The hiker reached the peak and admired the spectacular"

C Details of the 2D slice visualization

We create synthetic activations X in the 2D slice using the formula:

$$X = A + \alpha(B - A) + \beta(C - P),$$

where P is the orthogonal projection of C onto the line containing A and B, and α and β range from -0.25 to 1.25. This allows us to examine a broad area around and between the three reference activations.

To compute the colors for each point:

For each point X in the 2D slice, we compute three distances:

$$d_A = \|F_{p_A}(A) - F_{p_A}(X)\|_2$$
$$d_B = \|F_{p_B}(B) - F_{p_B}(X)\|_2$$
$$d_C = \|F_{p_C}(C) - F_{p_C}(X)\|_2$$

where F_p is defined as in Section 3. We normalize these distances by dividing by their respective maximum values over all points in the slice:

$$\hat{d}_A = \frac{d_A}{\max_X d_A}, \; \hat{d}_B = \frac{d_B}{\max_X d_B}, \; \hat{d}_C = \frac{d_C}{\max_X d_C}$$

The RGB color components for point X are then computed as: `Red` $= 1 - \hat{d}_A$, `Green` $= 1 - \hat{d}_B$, `Blue` $= 1 - \hat{d}_C$.
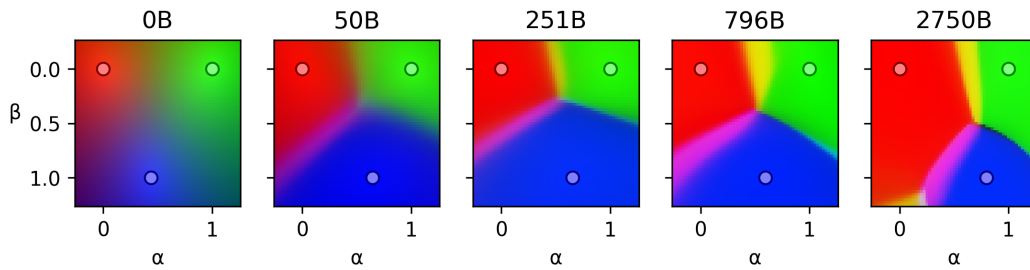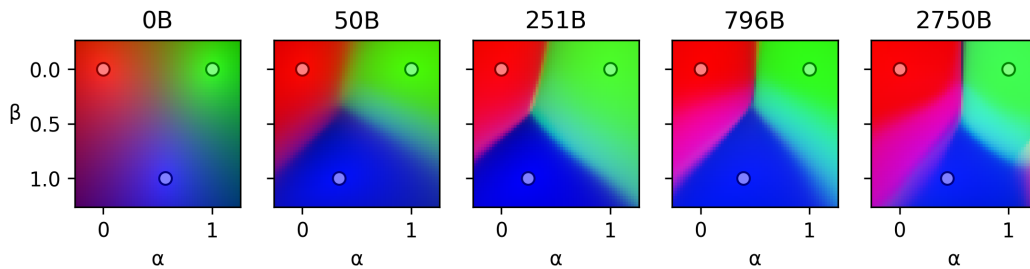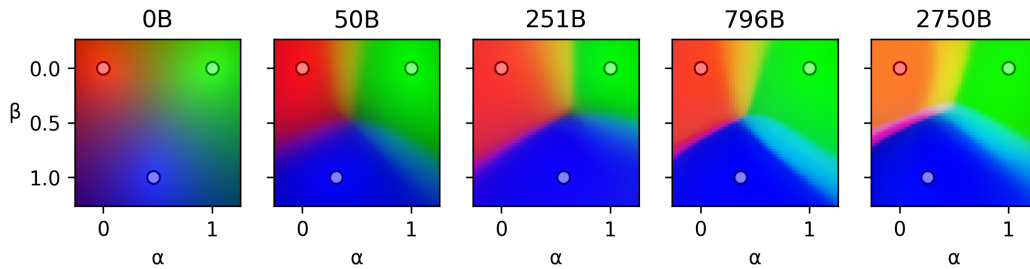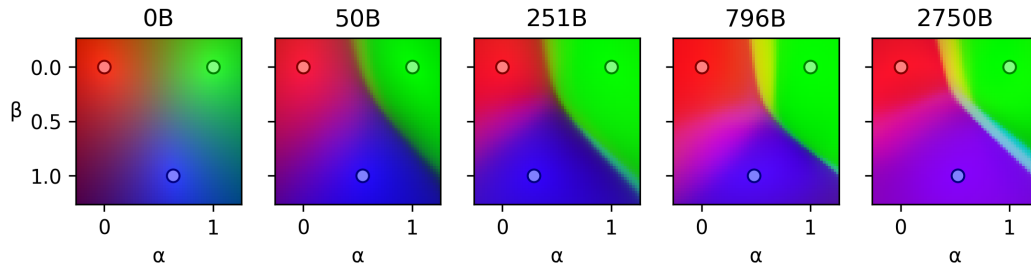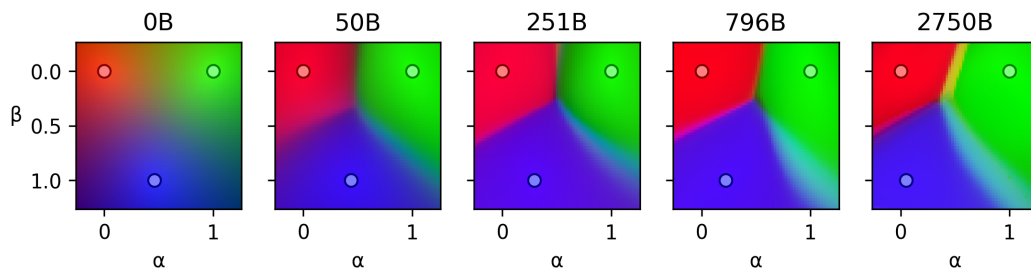
D   More 2D slice plots

Figure 5

Figure 6

Figure 7

8
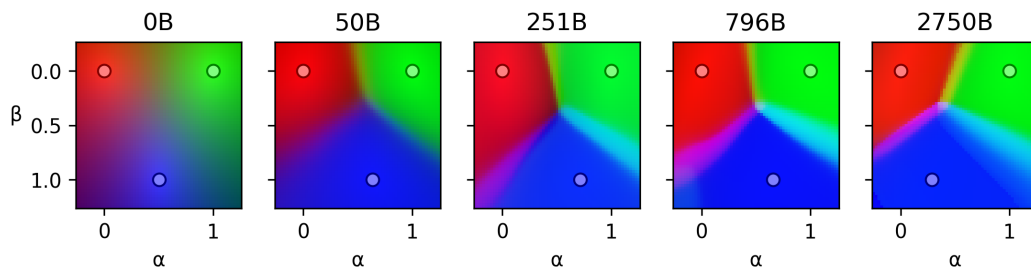
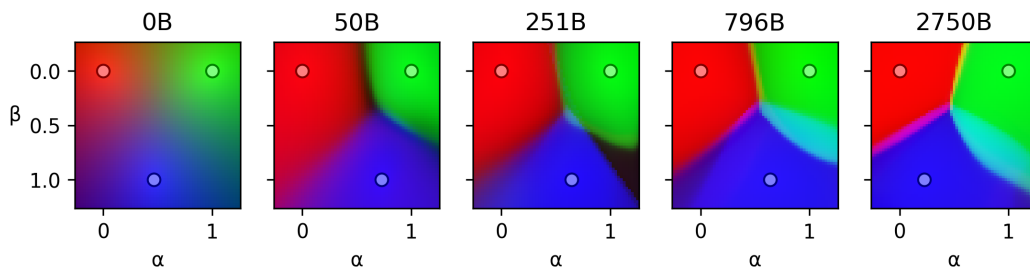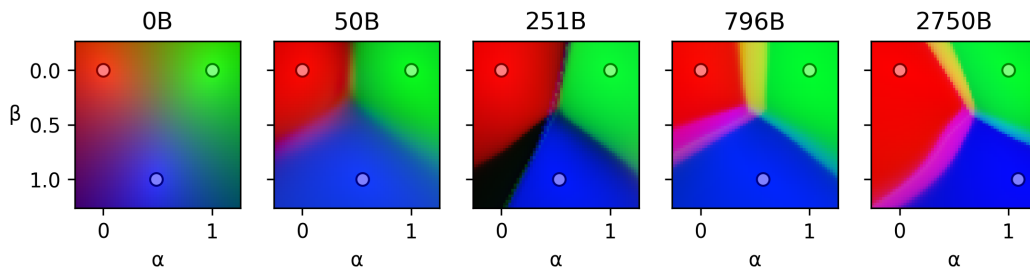Figure 8

Figure 9

Figure 10
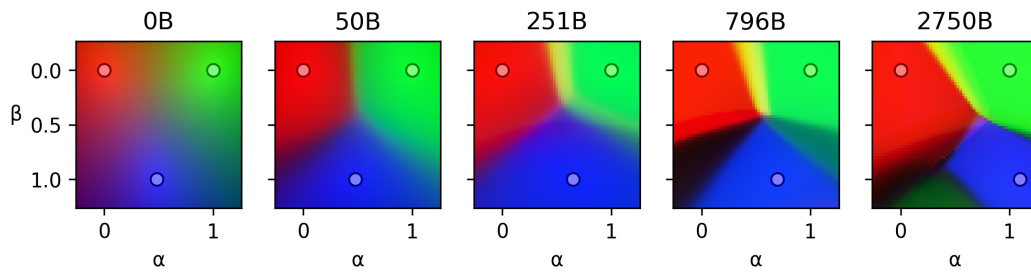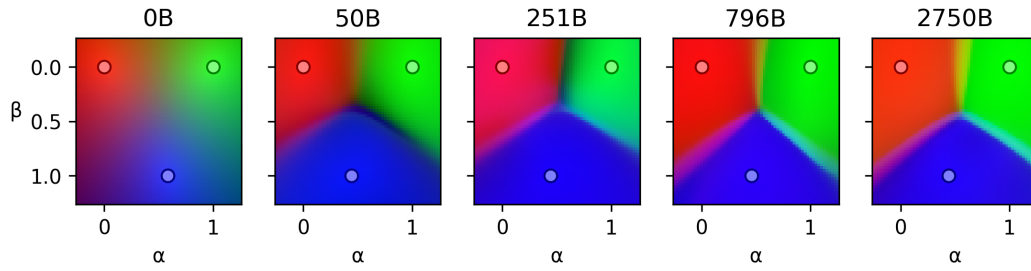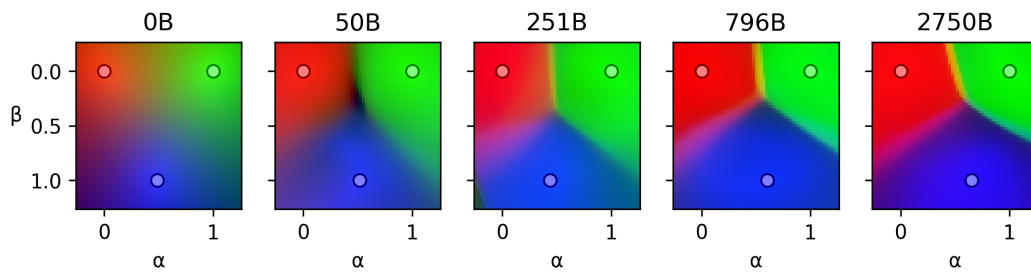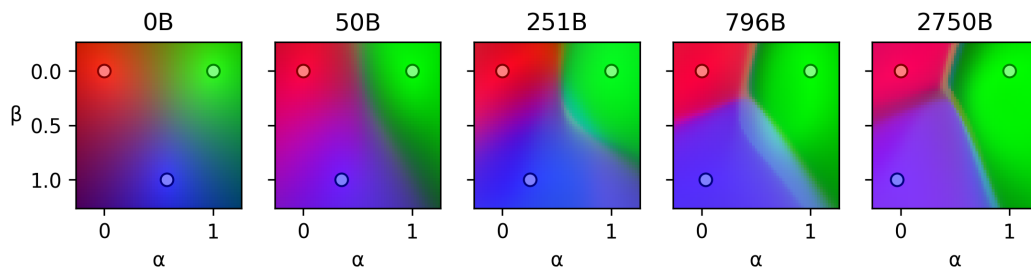

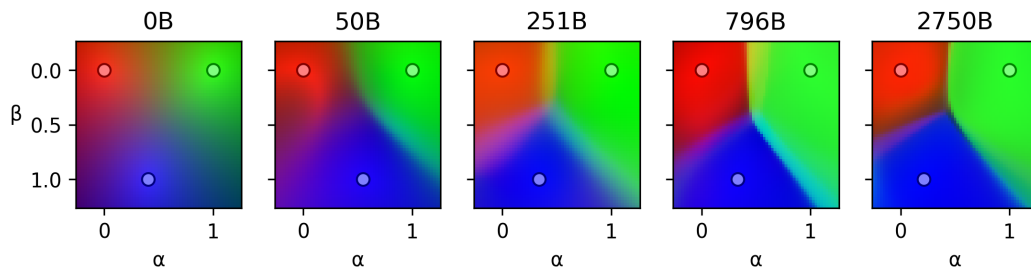
Figure 11
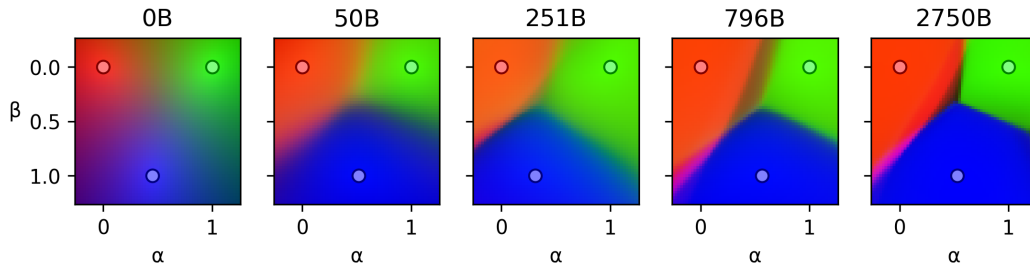
Figure 12
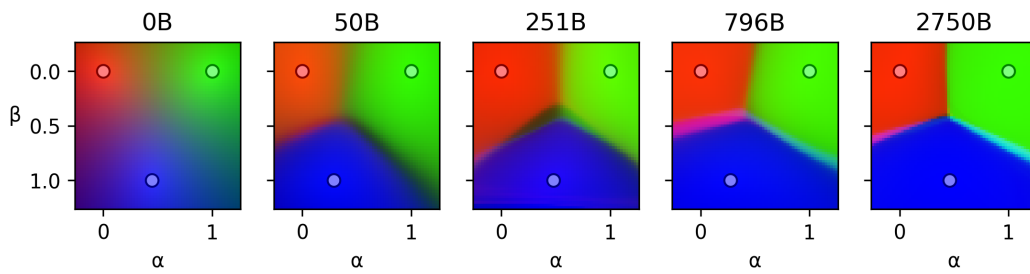
Figure 13

Figure 14

Figure 15

Figure 16

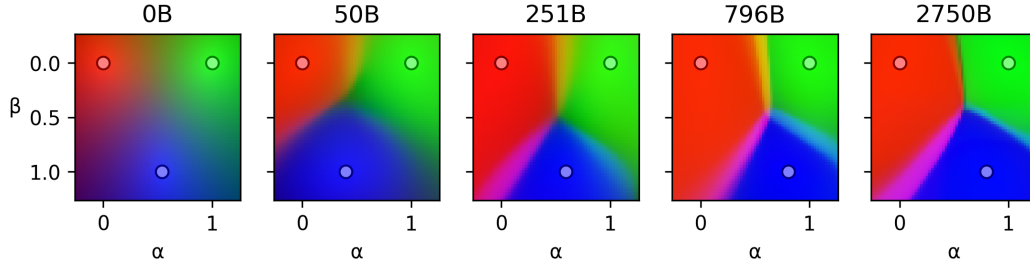Figure 17

Figure 18

Figure 19

11
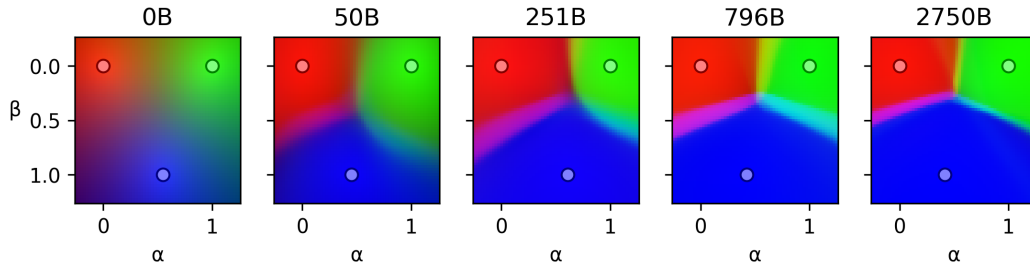
Figure 20

Figure 21
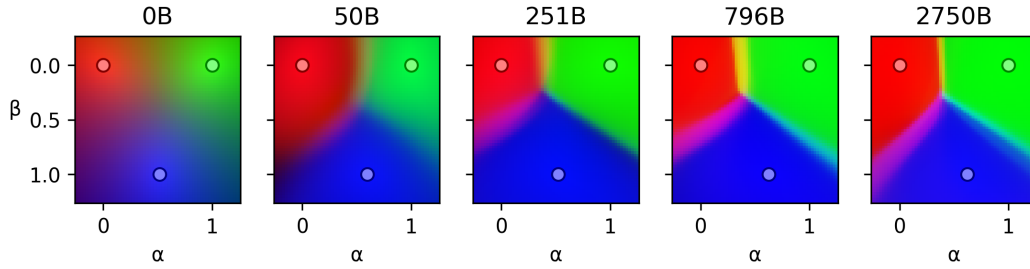
Figure 22

E   Shapes when patching after 1st vs after 7th layer

We present qualitative comparison of sharpness of the shapes of $d(\alpha)$, obtained using the same methods as in Section 4, when we interpolate and patch after 1st (Figure 23) and after 7th (Figure 24) layer in `Qwen2-1.5B`. The y-axis is $d(\alpha)/d(1)$ in both figures. While smoother, results for layer 7 still show a significant jump around the middle of the interpolation.

12
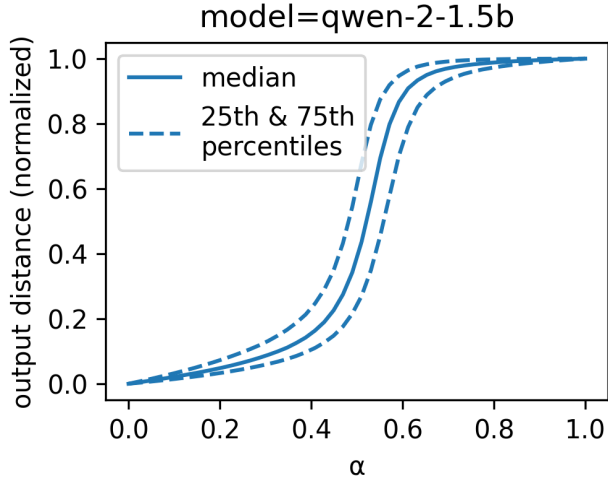
Figure 23: Interpolating and patching after 1st layer
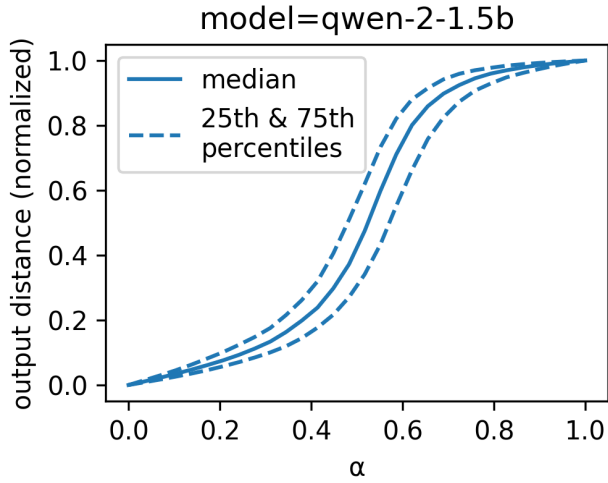
Figure 24: Interpolating and patching after 7th layer

F Direct comparison with polytopes

To directly compare the size of stable regions with polytopes, we analyze gate activations during interpolation between prompts in `Qwen2-0.5B`. For each pair of prompts (p_A, p_B), we collect gate activations at the last layer and count how many gates switch their sign (from positive to negative or vice versa) the beginning ($\alpha = 0$) and the end ($\alpha = 1$) of the interpolation. This provides a weak lower bound on the number of polytope boundaries crossed during interpolation.

13

Figure 25: Histogram of the number of gate activations that switch signs during the interpolation in `Qwen2-0.5B`. Despite observing typically only one stable region boundary during interpolation, we see hundreds or thousands of gate activations changing signs, suggesting stable regions encompass many polytopes.

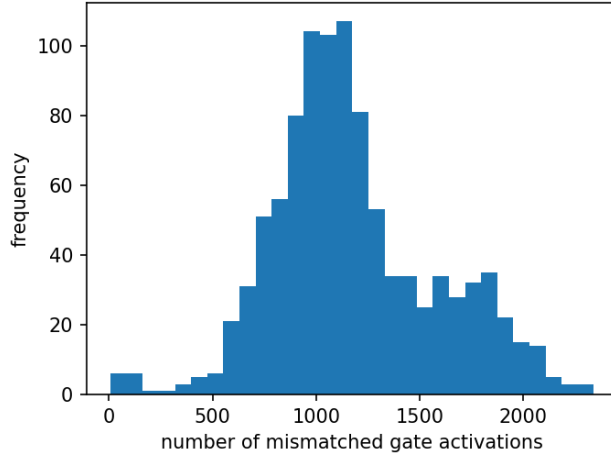As shown in Figure 25, we typically observe hundreds or thousands of gates switching signs between two prompts, while our interpolation experiments suggest crossing only one stable region boundary. This provides direct evidence that stable regions are indeed much larger than individual polytopes.

G Relationship between output sensitivity and semantic similarity
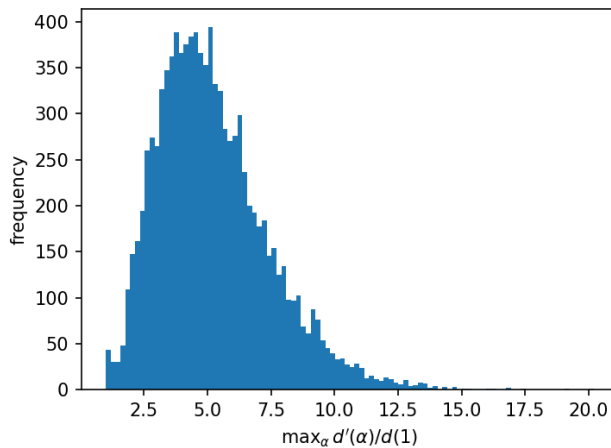
Figure 26: Distribution of maximum derivatives of the relative distance function in `Qwen2-0.5B`. The long right tail corresponds to prompt pairs that cross stable region boundaries.

To further investigate the relationship between semantic similarity and stable regions, we analyzed prompt pairs based on the maximum derivative of the relative distance ($\max_\alpha d'(\alpha)/d(1)$) in `Qwen2-0.5B`. We found that prompt pairs with low maximum derivative (< 1.2) consistently share the same last token, while pairs with high maximum derivative (> 10) have different last tokens. This analysis provides quantitative evidence for the relationship between semantic similarity and stable regions, supporting our earlier qualitative observations with manually selected prompts.

Here are representative examples:

G.1 Low sensitivity pairs (same last token)

p_A = "\r\n\r\nfor the little embellishments you speak of,"
p_B = "ckles' arrangements with the\r\n\r\n Angel,"

p_A = "\r\n\r\n\r\nOak then struck up "Jockey to the"
p_B = " nothing. The sale of big furniture barely paid the"

p_A = " being out of the question, we did the\r\n\r\n"
p_B = ";\r\n\r\nAnd from about him fierce effusion rolled\r\n\r\n"

p_A = "\r\n\r\n it had scratched the varnish from the"
p_B = "ically; but shall be content\r\n\r\nto produce the"

p_A = "003:006 And when the"
p_B = " a year since, how much I am in the"

p_A = "ilion\r\n\r\nand are beckoning to us."\r\n\r\n\r\n"
p_B = " answer; but the accent was decisive enough.\r\n\r\n\r\n"

p_A = " I scattered them among the heathen, and"
p_B = " show himself unto\r\n\r\nyou, my children, and"

p_A = "Israel, which he commanded our fathers, that they"
p_B = " the LORD.\r\n33:31 And they"

p_A = " quietly into the tree\r\n\r\nbeside him.\r\n\r\n\r\n"
p_B = " choir, Thea saw that she must have\r\n\r\n"

p_A = " be Jeddak of Helium. Say\r\n\r\n\r\n"
p_B = "ne to hit um wid a rock? How\r\n\r\n"

p_A = " hide the city of New York under it, and"
p_B = ": and their glory, and their multitude, and"

p_A = " subterranean\r\n\r\nchambers for ages,"
p_B = ", to Tarzan's\r\n\r\nconsternation,"

p_A = " to be\r\n\r\n entitled to his full name."
p_B = " He took up his candle \r\n\r\n\t\t\tto start."

p_A = " and there was the old man down the path a"
p_B = "as it used to be to keep from reading a"

p_A = " more natural time, and I was once more in"
p_B = "."\r\n\r\n\r\n"Well, there's a good deal in"

p_A = ", as represented by Sir Percy Blakeney,"
p_B = " a hundred paces,\r\n\r\nalmost without drawing breath,"

p_A = " thee.\r\nPALAMON.\r\n\r\nNo more; the"
p_B = " Neither boy spoke. If one moved, the"

p_A = "GONERIL.\r\n\r\nThis admiration, sir,"
p_B = ": Let us go forth to meet my people,"

p_A = " have been here a number\r\n\r\n of times,"
p_B = " had laboured to do: and, behold,"

p_A = " my things are all over Uncle Doc's house,"
p_B = " Paulvitch hastened back to his quarters,"

G.2 High sensitivity pairs (different last tokens)

p_A = "grandfather's opinions for the opinions of his father"
p_B = " local guides:—it falls into a fatal\r\n\r\n"

p_A = " life, but in short time\r\n\r\nAll offices of"
p_B = " did not know what else\r\n\r\nto do.\r\n\r\n\r\n"

p_A = " conformity; as a matter of course, of course"
p_B = " this the emigrant Evrémonde?"\r\n\r\n"

p_A = " kingdom\r\n\r\nalso; for he is mine elder brother"
p_B = " the host. And Joab and the captains\r\n\r\n"

p_A = " view with jealousy, view with a jealous eye.\r\n\r\n"
p_B = "ists of\r\n\r\n Octavius looked upon him"

p_A = ":\r\n\r\nWell may it sort that this portentous"
p_B = " in time of the cholera, some people\r\n\r\n"

p_A = ""Who are you?" asked the Scarecrow"
p_B = " mess of pottage, and that birthright\r\n\r\n"

p_A = "ema. Nature, poetic, silent, balmy"
p_B = " way, and walk therein, and ye shall\r\n\r\n"

p_A = ", saying, On this manner spake David.\r\n"
p_B = " medium of Clifford's\r\n\r\nhappiness, it would"

p_A = " before Thark, had I seen two men fight"
p_B = "tered. Elsewhere in the field,\r\n\r\nHere,"

p_A = " relatively few spoken words exchanged\r\n\r\neven in long conversations"
p_B = " whole congregation together was forty and two thousand three\r\n\r\n"

p_A = "ITUS LARTIUS; between them,\r\n\r\n"
p_B = " one of the retired\r\n\r\nstreets of a not"

p_A = "erringly, was also manifest and indisputable"
p_B = " revolution\r\n\r\ncovered with its cloud this summit where r"

p_A = " only look at her companion. Eleanor's countenance"
p_B = " the tribe of Judah, the mount Zion which"

p_A = "to go to Camden Place herself, she should not"
p_B = " the girl well knew, since he had been\r\n\r\n"

p_A = " lost\r\n\r\nin thought. He retraced his steps"
p_B = " seemed weak and harmless.\r\n\r\n \r\n\r\n\r\n What Black"

p_A = " \r\n\r\n\r\n "Keep it, then," said"
p_B = " the champions of the Cross?"\r\n\r\n\r\n"To the Knights"

p_A = " tears,\r\n\r\ntheir anguish, their terrors, their"
p_B = " his name be George, I'll call him Peter"

p_A = "\xa0MERCHANT. O, 'tis a"
p_B = " to walk to and fro through the earth.\r\n\r\n\r\n"

p_A = " made us for anything but this: to idolize"
p_B = " ruin.\r\n\r\n\xa0\xa0\xa0\xa0Is this your Christian counsel?"