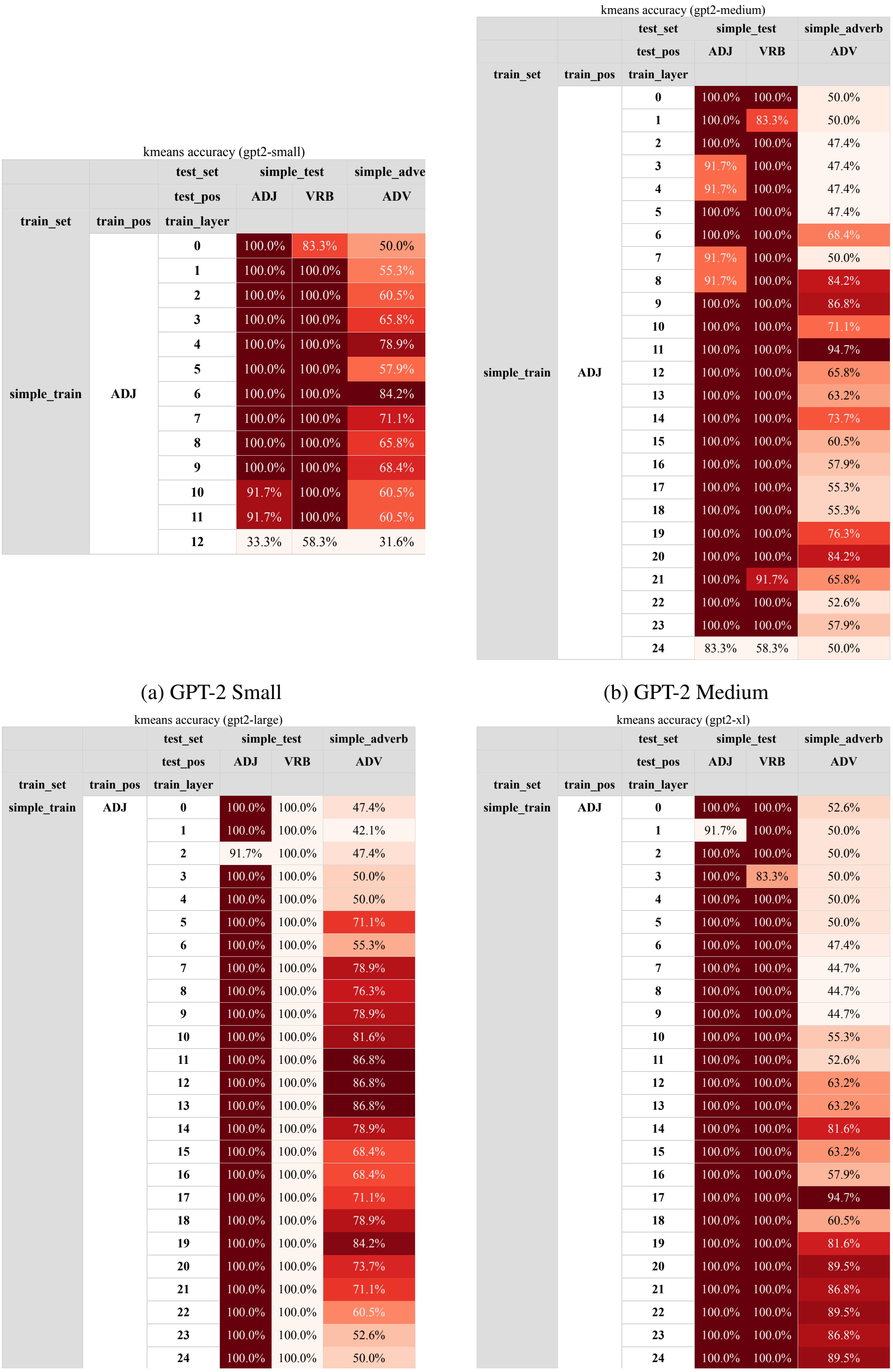


kmeans accuracy (gpt2-small)

| | | test_set | simple_test | | simple_adve |
|---|---|---|---|---|---|
| | | test_pos | ADJ | VRB | ADV |
| train_set | train_pos | train_layer | | | |
| simple_train | ADJ | 0 | 100.0% | 83.3% | 50.0% |
| | | 1 | 100.0% | 100.0% | 55.3% |
| | | 2 | 100.0% | 100.0% | 60.5% |
| | | 3 | 100.0% | 100.0% | 65.8% |
| | | 4 | 100.0% | 100.0% | 78.9% |
| | | 5 | 100.0% | 100.0% | 57.9% |
| | | 6 | 100.0% | 100.0% | 84.2% |
| | | 7 | 100.0% | 100.0% | 71.1% |
| | | 8 | 100.0% | 100.0% | 65.8% |
| | | 9 | 100.0% | 100.0% | 68.4% |
| | | 10 | 91.7% | 100.0% | 60.5% |
| | | 11 | 91.7% | 100.0% | 60.5% |
| | | 12 | 33.3% | 58.3% | 31.6% |

(a) GPT-2 Small

kmeans accuracy (gpt2-medium)

| | | test_set | simple_test | | simple_adverb |
|---|---|---|---|---|---|
| | | test_pos | ADJ | VRB | ADV |
| train_set | train_pos | train_layer | | | |
| simple_train | ADJ | 0 | 100.0% | 100.0% | 50.0% |
| | | 1 | 100.0% | 83.3% | 50.0% |
| | | 2 | 100.0% | 100.0% | 47.4% |
| | | 3 | 91.7% | 100.0% | 47.4% |
| | | 4 | 91.7% | 100.0% | 47.4% |
| | | 5 | 100.0% | 100.0% | 47.4% |
| | | 6 | 100.0% | 100.0% | 68.4% |
| | | 7 | 91.7% | 100.0% | 50.0% |
| | | 8 | 91.7% | 100.0% | 84.2% |
| | | 9 | 100.0% | 100.0% | 86.8% |
| | | 10 | 100.0% | 100.0% | 71.1% |
| | | 11 | 100.0% | 100.0% | 94.7% |
| | | 12 | 100.0% | 100.0% | 65.8% |
| | | 13 | 100.0% | 100.0% | 63.2% |
| | | 14 | 100.0% | 100.0% | 73.7% |
| | | 15 | 100.0% | 100.0% | 60.5% |
| | | 16 | 100.0% | 100.0% | 57.9% |
| | | 17 | 100.0% | 100.0% | 55.3% |
| | | 18 | 100.0% | 100.0% | 55.3% |
| | | 19 | 100.0% | 100.0% | 76.3% |
| | | 20 | 100.0% | 100.0% | 84.2% |
| | | 21 | 100.0% | 91.7% | 65.8% |
| | | 22 | 100.0% | 100.0% | 52.6% |
| | | 23 | 100.0% | 100.0% | 57.9% |
| | | 24 | 83.3% | 58.3% | 50.0% |

(b) GPT-2 Medium

kmeans accuracy (gpt2-large)

| | | test_set | simple_test | | simple_adverb |
|---|---|---|---|---|---|
| | | test_pos | ADJ | VRB | ADV |
| train_set | train_pos | train_layer | | | |
| simple_train | ADJ | 0 | 100.0% | 100.0% | 47.4% |
| | | 1 | 100.0% | 100.0% | 42.1% |
| | | 2 | 91.7% | 100.0% | 47.4% |
| | | 3 | 100.0% | 100.0% | 50.0% |
| | | 4 | 100.0% | 100.0% | 50.0% |
| | | 5 | 100.0% | 100.0% | 71.1% |
| | | 6 | 100.0% | 100.0% | 55.3% |
| | | 7 | 100.0% | 100.0% | 78.9% |
| | | 8 | 100.0% | 100.0% | 76.3% |
| | | 9 | 100.0% | 100.0% | 78.9% |
| | | 10 | 100.0% | 100.0% | 81.6% |
| | | 11 | 100.0% | 100.0% | 86.8% |
| | | 12 | 100.0% | 100.0% | 86.8% |
| | | 13 | 100.0% | 100.0% | 86.8% |
| | | 14 | 100.0% | 100.0% | 78.9% |
| | | 15 | 100.0% | 100.0% | 68.4% |
| | | 16 | 100.0% | 100.0% | 68.4% |
| | | 17 | 100.0% | 100.0% | 71.1% |
| | | 18 | 100.0% | 100.0% | 78.9% |
| | | 19 | 100.0% | 100.0% | 84.2% |
| | | 20 | 100.0% | 100.0% | 73.7% |
| | | 21 | 100.0% | 100.0% | 71.1% |
| | | 22 | 100.0% | 100.0% | 60.5% |
| | | 23 | 100.0% | 100.0% | 52.6% |
| | | 24 | 100.0% | 100.0% | 50.0% |

(c) GPT-2 Large

kmeans accuracy (gpt2-xl)

| | | test_set | simple_test | | simple_adverb |
|---|---|---|---|---|---|
| | | test_pos | ADJ | VRB | ADV |
| train_set | train_pos | train_layer | | | |
| simple_train | ADJ | 0 | 100.0% | 100.0% | 52.6% |
| | | 1 | 91.7% | 100.0% | 50.0% |
| | | 2 | 100.0% | 100.0% | 50.0% |
| | | 3 | 100.0% | 83.3% | 50.0% |
| | | 4 | 100.0% | 100.0% | 50.0% |
| | | 5 | 100.0% | 100.0% | 50.0% |
| | | 6 | 100.0% | 100.0% | 47.4% |
| | | 7 | 100.0% | 100.0% | 44.7% |
| | | 8 | 100.0% | 100.0% | 44.7% |
| | | 9 | 100.0% | 100.0% | 44.7% |
| | | 10 | 100.0% | 100.0% | 55.3% |
| | | 11 | 100.0% | 100.0% | 52.6% |
| | | 12 | 100.0% | 100.0% | 63.2% |
| | | 13 | 100.0% | 100.0% | 63.2% |
| | | 14 | 100.0% | 100.0% | 81.6% |
| | | 15 | 100.0% | 100.0% | 63.2% |
| | | 16 | 100.0% | 100.0% | 57.9% |
| | | 17 | 100.0% | 100.0% | 94.7% |
| | | 18 | 100.0% | 100.0% | 60.5% |
| | | 19 | 100.0% | 100.0% | 81.6% |
| | | 20 | 100.0% | 100.0% | 89.5% |
| | | 21 | 100.0% | 100.0% | 86.8% |
| | | 22 | 100.0% | 100.0% | 89.5% |
| | | 23 | 100.0% | 100.0% | 86.8% |
| | | 24 | 100.0% | 100.0% | 89.5% |

(d) GPT-2 XL

Figure A.2: 2-means classification accuracy for various GPT-2 sizes, split by layer (showing up to 24 layers)

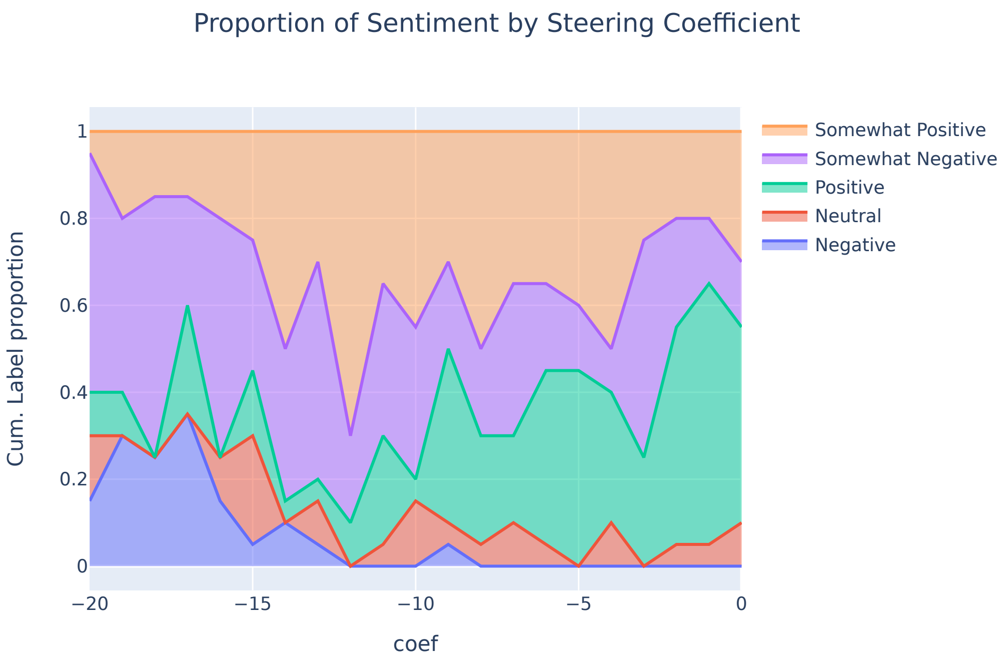
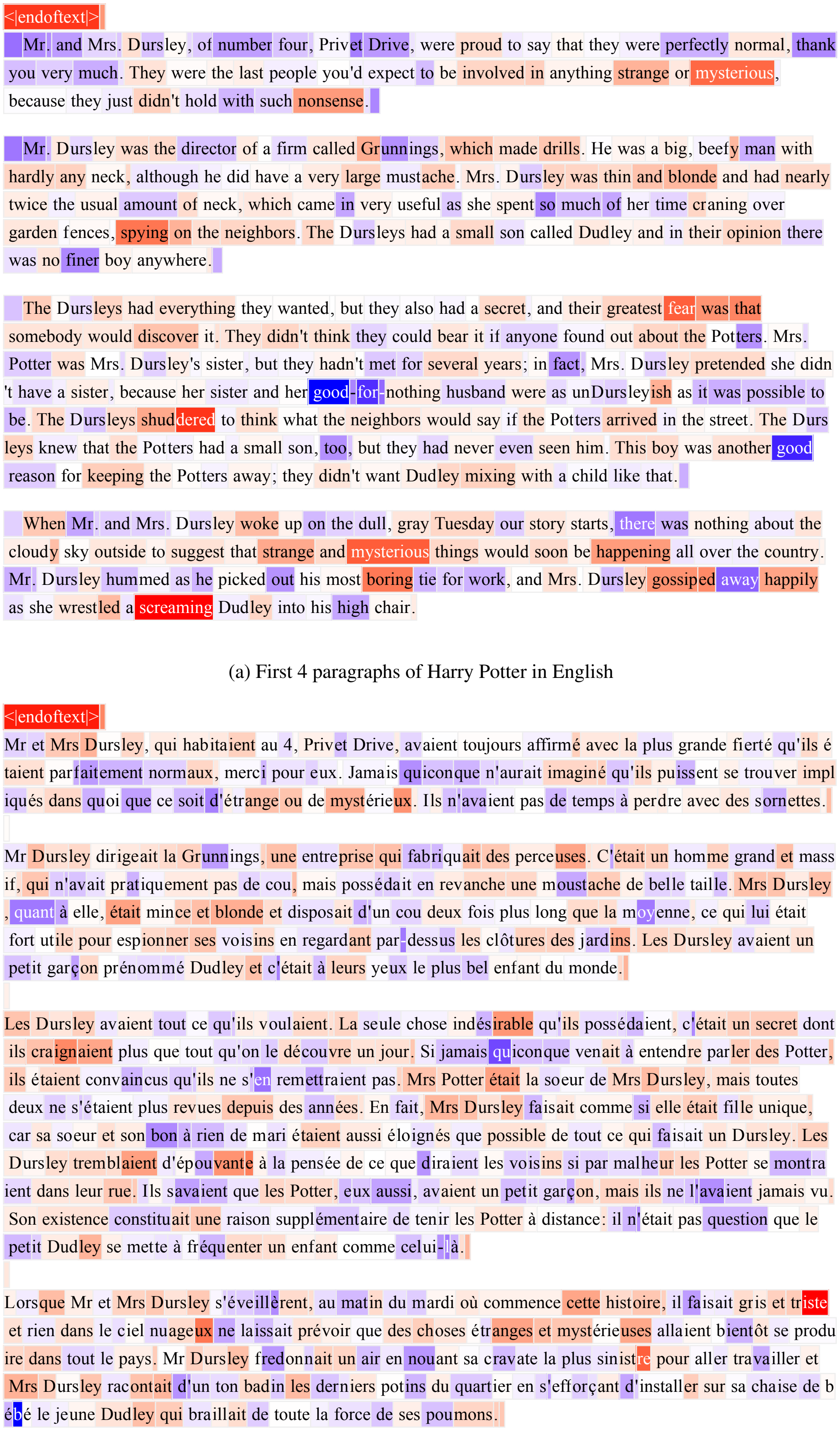


Figure A.3: Area plot of sentiment labels for generated outputs by activation steering coefficient, starting from a single positive movie review continuation prompt. Activation addition (Turner et al., 2023) was performed in GPT2-small's first residual stream layer. Classification was performed by GPT-4.



<|endoftext|>

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that.

When Mr. and Mrs. Dursley woke up on the dull, gray Tuesday our story starts, there was nothing about the cloudy sky outside to suggest that strange and mysterious things would soon be happening all over the country. Mr. Dursley hummed as he picked out his most boring tie for work, and Mrs. Dursley gossiped away happily as she wrestled a screaming Dudley into his high chair.

(a) First 4 paragraphs of Harry Potter in English

<|endoftext|>

Mr et Mrs Dursley, qui habitaient au 4, Privet Drive, avaient toujours affirmé avec la plus grande fierté qu'ils étaient parfaitement normaux, merci pour eux. Jamais quiconque n'aurait imaginé qu'ils puissent se trouver impliqués dans quoi que ce soit d'étrange ou de mystérieux. Ils n'avaient pas de temps à perdre avec des sornettes.

Mr Dursley dirigeait la Grunnings, une entreprise qui fabriquait des perceuses. C'était un homme grand et massif, qui n'avait pratiquement pas de cou, mais possédait en revanche une moustache de belle taille. Mrs Dursley, quant à elle, était mince et blonde et disposait d'un cou deux fois plus long que la moyenne, ce qui lui était fort utile pour espionner ses voisins en regardant par-dessus les clôtures des jardins. Les Dursley avaient un petit garçon prénommé Dudley et c'était à leurs yeux le plus bel enfant du monde.

Les Dursley avaient tout ce qu'ils voulaient. La seule chose indésirable qu'ils possédaient, c'était un secret dont ils craignaient plus que tout qu'on le découvre un jour. Si jamais quiconque venait à entendre parler des Potter, ils étaient convaincus qu'ils ne s'en remettraient pas. Mrs Potter était la soeur de Mrs Dursley, mais toutes deux ne s'étaient plus revues depuis des années. En fait, Mrs Dursley faisait comme si elle était fille unique, car sa soeur et son bon à rien de mari étaient aussi éloignés que possible de tout ce qui faisait un Dursley. Les Dursley tremblaient d'épouvante à la pensée de ce que diraient les voisins si par malheur les Potter se montraient dans leur rue. Ils savaient que les Potter, eux aussi, avaient un petit garçon, mais ils ne l'avaient jamais vu. Son existence constituait une raison supplémentaire de tenir les Potter à distance: il n'était pas question que le petit Dudley se mette à fréquenter un enfant comme celui-là.

Lorsque Mr et Mrs Dursley s'éveillèrent, au matin du mardi où commence cette histoire, il faisait gris et triste et rien dans le ciel nuageux ne laissait prévoir que des choses étranges et mystérieuses allaient bientôt se produire dans tout le pays. Mr Dursley fredonnait un air en nouant sa cravate la plus sinistre pour aller travailler et Mrs Dursley racontait d'un ton badin les derniers potins du quartier en s'efforçant d'installer sur sa chaise de bébé le jeune Dudley qui braillait de toute la force de ses poumons.

(b) First 3 paragraphs of Harry Potter in French

Figure A.4: First paragraphs of Harry Potter in different languages. Model: pythia-2.8b.