Value	Top-10 Tokens
$\mathbf{v}^{14}_{1853}$	transparency, disclosure, clearer, parency, iquette, humility, modesty, disclosures, accountability, safer
$\mathbf{v}^{15}_{73}$	respectful, honorable, healthy, decent, fair, erning, neutral, peacefully, respected, reconc
$\mathbf{v}^{15}_{1395}$	safe, neither, safer, course, safety, safe, Safe, apologize, Compact, cart
$\mathbf{v}^{16}_{216}$	refere, Messages, promises, Relations, accept, acceptance, Accept, assertions, persistence, warn
$\mathbf{v}^{17}_{462}$	should, should, MUST, ought, wisely, Should, SHOULD, safely, shouldn, urgently
$\mathbf{v}^{17}_{3209}$	peaceful, stable, healthy, calm, trustworthy, impartial, stability, credibility, respected, peace
$\mathbf{v}^{17}_{4061}$	Proper, proper, moder, properly, wisely, decency, correct, corrected, restraint, professionalism
$\mathbf{v}^{18}_{2921}$	thank, THANK, thanks, thank, Thank, apologies, Thank, thanks, Thanks, apologise
$\mathbf{v}^{19}_{1891}$	thanks, thank, Thanks, thanks, THANK, Thanks, Thank, Thank, thank, congratulations
$\mathbf{v}^{23}_{3770}$	free, fit, legal, und, Free, leg, pless, sound, qualified, Free

Table 8: The 10 manually picked value vectors used for toxic language suppression and the top-10 tokens in their projection to the vocabulary. Repetitions in the projections are a result of special characters not being shown. These vectors were found by manually searching for non-toxic words such as *"safe"* and *"peace"* in the projections to the vocabulary.

Layer	% Examples	Layer	% Examples
1	6.70	9	2.96
2	5.25	10	3.78
3	13.74	11	4.74
4	3.13	12	7.45
5	1.02	13	10.79
6	1.07	14	9.88
7	1.86	15	9.81
8	2.60	16	15.22

Table 9: The percentage of saturation events per layer using WIKILM, for the WIKITEXT-103 validation set.

Layer	% Examples	Layer	% Examples
1	2.21	13	1.24
2	0.77	14	1.62
3	1.06	15	2.37
4	0.74	16	2.72
5	0.85	17	2.99
6	0.83	18	3.80
7	0.83	19	4.15
8	0.72	20	5.21
9	0.93	21	5.67
10	0.99	22	9.31
11	1.16	23	14.52
12	1.32	24	34.15

Table 10: The percentage of saturation events per layer using GPT2, for the WIKITEXT-103 validation set.