Token Sequence	n	ct	ψ
lower case	3	2	0.736012
storm	2	4	0.716379
excursion	4	2	0.713134
====... *(72 'equals' signs)*	8	2	0.712982
Mom	3	2	0.706778
acre	3	2	0.629213
Subject	3	2	0.607172
ninth	3	2	0.606669
processing elements	3	2	0.599549
CVC	3	2	0.596735
VPN	3	3	0.596052
Regul	3	2	0.591968
bore	2	2	0.590212
$\dot{G	5	2	0.589714
Rates	3	2	0.589637
INSURANCE	5	2	0.584323
Commercial	4	2	0.581543
Barney	3	3	0.574872
PTA	3	2	0.571932
penetrated	4	2	0.570164
MG	3	2	0.569830
Leigh	3	2	0.567894
jail	3	3	0.567225
TNS	3	2	0.567003
peptides	4	2	0.565775
John Arena	3	2	0.565648
Disease	4	2	0.564662
welfare	4	4	0.564364
wild type	3	2	0.560699
uws	3	3	0.557799
ongrel	4	3	0.554208
liquid cry	3	3	0.553408
princess	3	2	0.551672
Denmark	3	2	0.548702
birthday	3	2	0.548504
atedmes	4	2	0.548171
"ENOENT	5	2	0.547169
third-party	4	2	0.546949
aliens	3	2	0.546507
Durban	3	4	0.545848
Bouncy	4	3	0.545826
CHO	3	2	0.542762
unjust	3	2	0.538813
these motivational	4	3	0.537485
DLS	3	4	0.535933
\n&	3	2	0.534510
uneven	3	2	0.533137
watt	3	2	0.532243
'She	3	2	0.531300
HP	3	3	0.529555

Table 6: **Llama-2-7b** Pile results (1658 sequences total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment. ψ is averaged over all occurrences.

Token Sequence	n	ct	ψ
</td>\n<td>	9	2	0.627583
{d}x	5	3	0.599395
*\n	4	3	0.587016
_{n=1}{\in	7	4	0.585434
</td>\n<td	8	2	0.573310
-2-2007-061	12	3	0.551581
reticulum	4	3	0.549337
INSURANCE	5	2	0.548263
32;\n internal static	8	2	0.547893
;\n internal static	6	9	0.540374
: At	4	2	0.538609
(2,9,'	6	4	0.537495
Respondent	4	2	0.534509
\t\t}\n\n\t	7	3	0.530669
(3,0,'	6	4	0.529493
_{n-1}\ar	7	2	0.527303
thank you for your understanding	6	2	0.513979
hydroxyl	4	2	0.510059
>\n*/\private $	9	2	0.510054
in mukaan	5	2	0.506333
{w}{B}_{	6	2	0.505970
/2\Z	5	2	0.501998
'); \nINSERT INTO	6	10	0.501055
7-f131	7	2	0.496881
0, 1L>	8	2	0.495809
/0 S	5	2	0.492042
5 Audi	4	2	0.491043
all that apply	4	3	0.490469
": true,\n	6	2	0.486807
4,\n	5	2	0.485315
to as DSP	5	2	0.484967
B]{}\	6	2	0.483484
;\ninternal	5	3	0.479777
100% used	6	2	0.475673
", "x":	5	3	0.474701
2.7	4	2	0.473720
</td>\n	6	2	0.473578
" code="	4	4	0.473514
e2d-d	6	2	0.473418
is under conversion	4	5	0.473355
{ int|sys	5	3	0.471213
();\n}\n\nprivate boolean isAny	12	2	0.470941
(2,8,'	6	4	0.470214
trachea	4	2	0.469154
use in an automobile	6	2	0.467788
at org.apache.c	7	5	0.467637
world around us	4	2	0.464469
2\left(1+x	8	2	0.463555
or Commodore	5	3	0.463106
11-117	7	2	0.459824

Table 7: **Llama-3-8b** Pile results (819 sequences total). n is the number of tokens in the sequence, and 'ct' represents occurrences of this segment. ψ is averaged over all occurrences.