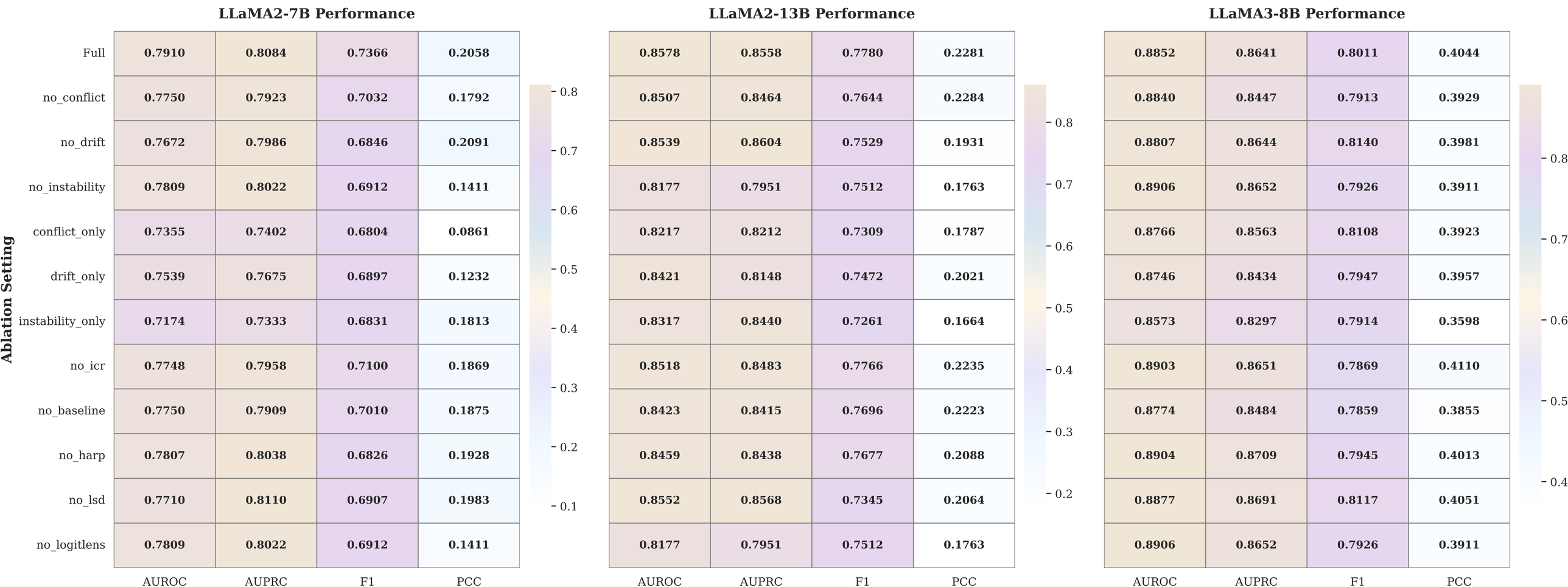
LLaMA2-7B Performance
LLaMA2-13B Performance
LLaMA3-8B Performance
Ablation Setting
Full
no_conflict
no_drift
no_instability
conflict_only
drift_only
instability_only
no_icr
no_baseline
no_harp
no_lsd
no_logitlens
AUROC
AUPRC
F1
PCC
0.7910 0.8084 0.7366 0.2058
0.7750 0.7923 0.7032 0.1792
0.7672 0.7986 0.6846 0.2091
0.7809 0.8022 0.6912 0.1411
0.7355 0.7402 0.6804 0.0861
0.7539 0.7675 0.6897 0.1232
0.7174 0.7333 0.6831 0.1813
0.7748 0.7958 0.7100 0.1869
0.7750 0.7909 0.7010 0.1875
0.7807 0.8038 0.6826 0.1928
0.7710 0.8110 0.6907 0.1983
0.7809 0.8022 0.6912 0.1411
0.8578 0.8558 0.7780 0.2281
0.8507 0.8464 0.7644 0.2284
0.8539 0.8604 0.7529 0.1931
0.8177 0.7951 0.7512 0.1763
0.8217 0.8212 0.7309 0.1787
0.8421 0.8148 0.7472 0.2021
0.8317 0.8440 0.7261 0.1664
0.8518 0.8483 0.7766 0.2235
0.8423 0.8415 0.7696 0.2223
0.8459 0.8438 0.7677 0.2088
0.8552 0.8568 0.7345 0.2064
0.8177 0.7951 0.7512 0.1763
0.8852 0.8641 0.8011 0.4044
0.8840 0.8447 0.7913 0.3929
0.8807 0.8644 0.8140 0.3981
0.8906 0.8652 0.7926 0.3911
0.8766 0.8563 0.8108 0.3923
0.8746 0.8434 0.7947 0.3957
0.8573 0.8297 0.7914 0.3598
0.8903 0.8651 0.7869 0.4110
0.8774 0.8484 0.7859 0.3855
0.8904 0.8709 0.7945 0.4013
0.8877 0.8691 0.8117 0.4051
0.8906 0.8652 0.7926 0.3911
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1