Table 1: Additional experimenst on GPT-J and GPT2-XL to illustrate our method's generalization ability. The mean and standard deviation are reported for 3 repetitions with different ICL examples.

	GPT-J					GPT2-XL				
Method	ICL	CC	DC	PC	UniBias	ICL	CC	DC	PC	UniBias
SST-2	$90.67_{1.06}$	$90.75_{1.88}$	$91.90_{1.12}$	$89.14_{1.68}$	$\mathbf{92.01}_{1.04}$	$56.69_{3.39}$	$81.35_{4.36}$	$87.00_{2.66}$	$\mathbf{88.69}_{1.03}$	$\mathbf{88.69}_{1.10}$
WiC	$50.21_{0.99}$	$51.83_{0.53}$	$50.89_{0.90}$	$51.67_{0.59}$	$\mathbf{52.77}_{0.52}$	$49.06_{0.25}$	$50.88_{0.92}$	$50.31_{1.33}$	$48.69_{1.79}$	$\mathbf{51.22}_{0.69}$
COPA	$47.33_{0.47}$	$47.00_{0.35}$	$48.66_{0.94}$	$52.99_{5.88}$	$\mathbf{55.33}_{2.62}$	$49.00_{1.63}$	$46.33_{1.70}$	$49.33_{1.25}$	$\mathbf{52.33}_{2.36}$	$52.00_{3.74}$
MR	$89.26_{1.65}$	$85.31_{2.36}$	$90.29_{0.45}$	$89.66_{0.27}$	$\mathbf{90.39}_{0.36}$	$51.11_{1.76}$	$72.00_{4.44}$	$82.63_{1.80}$	$83.07_{1.56}$	$\mathbf{83.53}_{1.02}$
RTE	$50.54_{3.19}$	$53.99_{0.68}$	$54.13_{1.19}$	$49.58_{4.85}$	$\mathbf{56.20}_{1.80}$	$52.37_{0.34}$	$52.91_{0.45}$	$53.15_{0.29}$	$53.30_{1.28}$	$\mathbf{56.16}_{1.23}$
Avg.	65.60	65.78	67.17	66.61	**69.34**	51.65	60.69	64.48	65.32	**66.32**

Table 2: Additional experiments on streamlining grid search by adopting a fixed set of thresholds. The **best** and second-best results are marked by bold and undrline, respectively. Experiments are conducted using Llama2-7b.

	SST2	MMLU	COPA	RTE	MR	Trec	Avg.
ICL	$87.22_{6.03}$	$41.73_{2.25}$	$67.60_{2.30}$	$66.21_{7.30}$	$89.37_{1.83}$	$72.92_{12.42}$	70.84
CC	$92.24_{3.39}$	$43.72_{0.97}$	$67.80_{2.17}$	$64.33_{3.68}$	$91.77_{1.42}$	$76.44_{3.21}$	72.72
DC	$94.15_{1.22}$	$43.57_{1.38}$	$60.40_{2.79}$	$65.49_{2.09}$	$\mathbf{92.35}_{0.23}$	$77.16_{3.94}$	72.19
PC	$93.90_{1.54}$	$34.12_{3.41}$	$67.80_{3.70}$	$62.59_{4.71}$	$91.39_{1.65}$	$74.92_{5.78}$	70.79
UniBias	$\mathbf{94.54}_{0.62}$	$\mathbf{44.83}_{0.24}$	$\mathbf{69.00}_{2.74}$	$\mathbf{67.65}_{6.44}$	$92.19_{0.37}$	$\mathbf{80.80}_{3.17}$	**74.84**
UniBias with fixed thresholds	$\underline{94.42}_{0.80}$	$\underline{44.47}_{0.93}$	$\underline{68.20}_{1.79}$	$\underline{67.51}_{4.97}$	$\mathbf{92.35}_{0.17}$	$\underline{80.32}_{4.40}$	<u>74.54</u>

Table 3: Shared biased attention heads and their frequency of occurrence across 12 datasets in the Llama2-7b model. Each entry represents a specific (layer index, head index) combination.

$(19, 10)$	$(19, 14)$	$(16, 29)$	$(19, 21)$	$(25, 21)$	$(16, 11)$	$(18, 31)$	$(18, 1)$
6	5	4	4	3	3	3	3

Table 4: Additional experiments on eliminating common biased components. Attention heads list in Table 3 are removed from the original Llama2-7b model, and the modified model is then evaluated across multiple tasks.

	SST2	MMLU	COPA	RTE	MR	Trec	Avg.
ICL	$87.22_{6.03}$	$41.73_{2.25}$	$67.60_{2.30}$	$66.21_{7.30}$	$89.37_{1.83}$	$72.92_{12.42}$	70.84
Unibias	$\mathbf{94.54}_{0.62}$	$\mathbf{44.83}_{0.24}$	$\mathbf{69.00}_{2.74}$	$\mathbf{67.65}_{6.44}$	$92.19_{0.37}$	$\mathbf{80.80}_{3.17}$	**74.84**
Eliminating Common Biased Components	$94.32_{0.60}$	$44.20_{1.14}$	$68.00_{2.87}$	$67.37_{4.60}$	$\mathbf{92.43}_{0.09}$	$77.60_{4.75}$	73.98

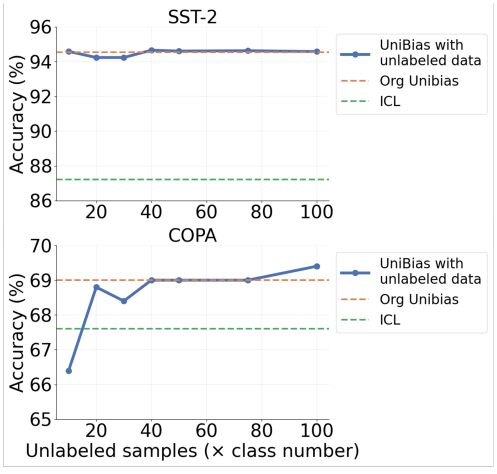

Figure 1: Performance of Unibias using unlabeled samples as support set. It is compared against standard ICL and the original Unibias method.

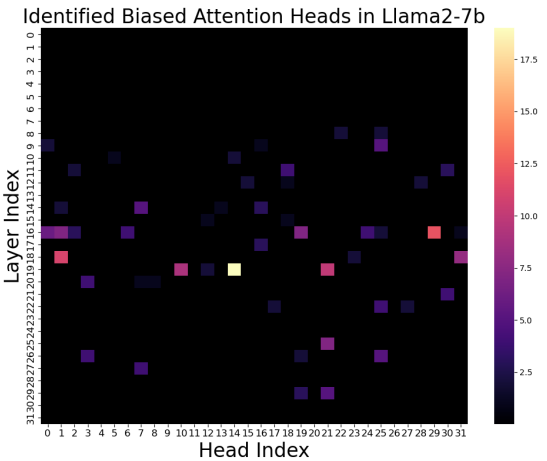

Figure 2: Identified biased attention heads across 12 datasets with 5 repetitions for each dataset.