

LLM Performance in IPV Detection

F1 scores across different models with error bars (± 1 SD)

Model

mlx-community/gpt-oss-120b

qwen3-next-80b-a3b-thinking-mlx

qwen/qwen3-next-80b

0%

20%

40%

60%

80%

F1 Score

