Table 1: Refusal rates on GSM8K of models "pseudo-aligned" to consider math questions as harmful, using one of our three alignment techniques. Updated values are shown in bold.

	Alignment method		
Model	Prompting	SFT	`EvilMath`
LLaMA 3.1 8B	**69.5**	**95.1**	-
LLaMA 3.1 405B	78.3	-	58.2
Claude 3.5 Haiku	-	-	95.0