Model	$K = 1$				$K = 4$			
	t1	t2	t3	t4	t1	t2	t3	t4
GPT-3.5	100	99	85	91	95 (−5)	90 (−9)	84 (−1)	84 (−7)
Llama-3 70B	100	99	97	99	100 (0)	99 (0)	96 (−1)	92 (−7)
Llama-2 70B	100	96	69	77	88 (−12)	96 (0)	63 (−6)	46 (−31)
Qwen-1.5 72B	100	94	52	70	66 (−34)	91 (−3)	28 (−24)	34 (−36)

(a) Setting 1: Addition in original numerical form and in different languages where t1 = add, t2 = add_in_en, t3 = add_in_fr, t4 = add_in_es.

Model	$K = 1$			$K = 3$		
	t1	t2	t3	t1	t2	t3
GPT-3.5	100	100	100	93 (−7)	62 (−38)	56 (−44)
Llama-3 70B	100	100	100	82 (−18)	90 (−10)	83 (−17)
Llama-2 70B	97	94	90	75 (−22)	75 (−19)	50 (−40)
Qwen-1.5 72B	100	99	91	65 (−35)	87 (−12)	48 (−43)

(b) Setting 2: Naming the capital, continent and capitalize the country name where t1 = capital, t2 = continent, t3 = capitalization.

Model	$K = 1$			$K = 3$		
	t1	t2	t3	t1	t2	t3
GPT-3.5	100	100	100	100 (0)	97 (−3)	97 (−3)
Llama-3 70B	100	100	100	99 (−1)	100 (0)	99 (−1)
Llama-2 70B	100	100	95	95 (−5)	99 (−1)	84 (−11)
Qwen-1.5 72B	100	100	99	100 (0)	98 (−2)	98 (−1)

(c) Setting 3: t1 = copy(op1), t2 = copy(op2) and t3 = op1+op2.

Model	$K = 1$				$K = 4$			
	t1	t2	t3	t4	t1	t2	t3	t4
GPT-3.5	100	87	100	54	94 (−6)	56 (−31)	97 (−3)	12 (−42)
Llama-3 70B	100	63	100	40	99 (−1)	29 (−34)	99 (−1)	13 (−27)
Llama-2 70B	100	55	100	38	99 (−1)	35 (−20)	97 (−3)	7 (−31)
Qwen-1.5 72B	100	62	100	34	89 (−11)	30 (−32)	100 (0)	15 (−19)

(d) Setting 4: First or last letter in upper or lower cases where t1 = first_letter, t2 = last_letter, t3 = first_letter_cap, t4 = last_letter_cap.

Table 2: Accuracy for each task in percentage, with the delta change given in parenthesis. For each setting we calculate the accuracy with prompts consisting of task examples of only one task ($K = 1$ case) and with prompts consisting of examples from multiple tasks ($K > 1$ case).