

Settings (high→low)		en	de	fr	zh	es	ja	ko	Avg.
Zero-shot	text-davinci-003								
	<b>Basic Prompt</b>	53.3	52.9	50.8	53.6	53.0	50.3	50.3	52.0
	<b>XLT</b>	64.6	56.0	55.3	57.4	56.0	54.6	55.9	<b>57.1</b>
	gpt-3.5-turbo								
	<b>Basic Prompt</b>	73.6	68.3	68.4	63.4	69.6	59.7	55.7	<b>65.5</b>
	<b>XLT</b>	65.3	66.1	64.8	65.5	63.3	62.4	57.6	63.6
Few-shot	text-davinci-003								
	(Ahuja et al., 2023) <sup>†</sup>	72.5	69.8	71.3	65.2	70.1	65.4	65.8	68.6
	<b>Basic Prompt</b>	77.8	70.6	72.5	65.0	71.7	62.5	60.5	68.7
	<b>XLT</b>	76.5	78.8	77.4	61.1	78.0	65.0	68.7	<b>72.2</b>
	gpt-3.5-turbo								
	<b>Basic Prompt</b>	65.9	70.3	66.6	64.1	68.2	65.6	64.0	66.4
	<b>XLT</b>	73.4	69.8	68.5	70.9	67.8	66.4	66.7	<b>69.1</b>

Table 10: Accuracy scores on the PAWS-X benchmark. Ahuja et al. (2023)<sup>†</sup> utilize 8-shot learning.

Settings (high→low)		en	de	ru	fr	zh	es	ja	vi	tr	th	Avg.
Zero-shot	text-davinci-003											
	<b>Basic Prompt</b>	48.1	33.8	15.9	34.8	18.2	34.1	27.7	23.6	24.0	29.6	29.0
	<b>XLT</b>	51.1	42.3	27.3	43.0	36.7	43.3	46.8	35.8	37.8	38.1	<b>40.2</b>
	gpt-3.5-turbo											
	<b>Basic Prompt</b>	51.1	40.6	28.4	40.1	16.5	39.3	25.9	23.3	26.9	23.7	31.6
	<b>XLT</b>	56.7	46.0	33.9	47.6	33.0	47.9	47.5	36.5	39.1	38.6	<b>42.7</b>
Few-shot	text-davinci-003											
	<b>Basic Prompt</b>	52.6	42.3	21.8	42.9	33.1	42.8	45.5	35.5	37.5	36.6	39.1
	<b>XLT</b>	57.6	49.4	42.7	50.9	51.0	50.0	60.0	46.9	46.9	40.5	<b>49.6</b>
	gpt-3.5-turbo											
	<b>Basic Prompt</b>	53.2	48.6	31.0	46.1	40.9	47.9	51.4	38.5	40.0	39.3	43.7
	<b>XLT</b>	59.6	52.5	43.8	53.9	51.9	54.0	63.2	49.4	52.1	44.7	<b>52.5</b>

Table 11: F1 scores on the MKQA benchmark. The average score is the macro average F1 score.

Settings (high→low)		en	fr	zh	es	vi	tr	Avg.
Zero-shot	text-davinci-003							
	<b>Basic Prompt</b>	22.2	26.2	30.8	25.1	22.0	15.9	23.7
	<b>XLT</b>	24.4	28.2	32.2	26.0	22.3	17.9	<b>25.2</b>
	gpt-3.5-turbo							
	(Lai et al., 2023)	19.7	20.8	21.1	17.8	–	14.5	–
	<b>Basic Prompt</b>	25.3	26.2	30.2	26.3	21.1	19.2	24.7
	<b>XLT</b>	26.8	28.1	33.3	26.4	21.3	20.5	<b>26.1</b>
Few-shot	text-davinci-003							
	<b>Basic Prompt</b>	29.2	29.6	33.2	28.3	22.5	18.1	26.8
	<b>XLT</b>	28.2	30.3	34.4	29.4	22.7	18.6	<b>27.3</b>
	gpt-3.5-turbo							
	<b>Basic Prompt</b>	25.7	27.2	30.8	27.8	21.5	19.7	25.5
	<b>XLT</b>	28.5	29.2	35.0	28.6	23.7	22.3	<b>27.9</b>

Table 12: ROUGE-1 scores on the XL-Sum\* benchmark.

Settings		High-High				High-Low				Low-Low			
		zh-ru		de-vi		zh-th		zh-jv		th-gl		jv-th	
		→	←	→	←	→	←	→	←	→	←	→	←
Zero-shot	text-davinci-003												
	Basic Prompt	19.8	24.2	26.5	24.5	10.2	11.8	8.1	14.0	17.9	12.0	10.0	6.2
	XLT	<b>21.6</b>	<b>24.8</b>	<b>27.4</b>	<b>24.8</b>	<b>12.6</b>	<b>16.4</b>	<b>11.1</b>	<b>18.2</b>	<b>20.7</b>	<b>14.2</b>	<b>11.7</b>	<b>9.0</b>
	gpt-3.5-turbo												
	Basic Prompt	23.3	25.4	34.1	29.6	16.6	18.6	9.1	16.2	18.1	18.5	13.0	7.2
	XLT	<b>25.3</b>	<b>25.6</b>	<b>33.3</b>	<b>31.3</b>	<b>19.9</b>	<b>19.3</b>	<b>10.5</b>	<b>19.0</b>	<b>22.1</b>	<b>21.9</b>	<b>15.9</b>	<b>10.6</b>

Table 13: BLEU scores on the FLORES\* benchmark.

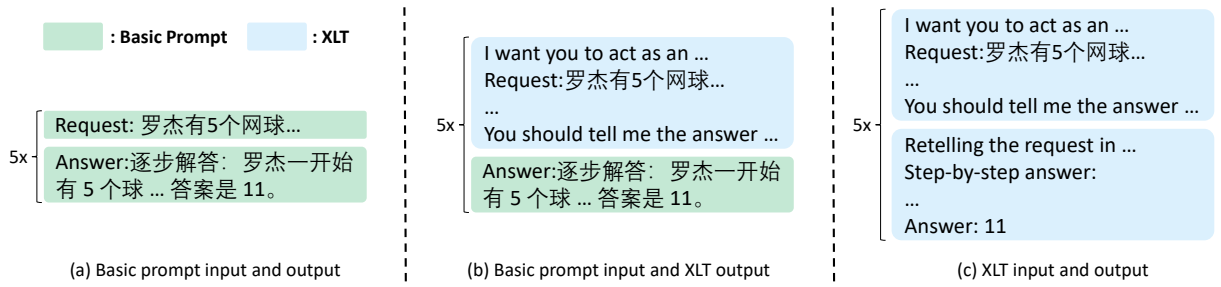


Figure 5: Illustrations of different demonstration input-output pairs in the few-shot learning.

Input:

Request: 制作一件袍子需要 2 匹蓝色纤维布料和这个数量一半的白色纤维布料。它一共需要用掉多少匹布料?

text-davinci-003 output:

Answer: 3 匹布料

gpt-3.5-turbo output:

3 匹布料。

Answer:

3

Figure 6: A Chinese example of the **MGS** benchmark using **basic prompt** and the corresponding outputs under the **zero-shot** setting.

Input:

I want you to act as an arithmetic reasoning expert for Chinese .

Request: 制作一件袍子需要 2 匹蓝色纤维布料和这个数量一半的白色纤维布料。它一共需要用掉多少匹布料?

You should retell the request in English.

You should do step-by-step answer to obtain a number answer .

You should step-by-step answer the request.

You should tell me the answer in this format 'Answer: '.

text-davinci-003 output:

The request is: To make a robe, you need 2 pieces of blue fiber fabric and half of that amount of white fiber fabric. How many pieces of fabric are needed in total?  
Step-by-step answer:  
1. Start with the amount of blue fiber fabric: 2 pieces  
2. Half of that amount is 1 piece  
3. Add the 1 piece of white fiber fabric to the 2 pieces of blue fiber fabric  
4. Answer: 3 pieces of fabric are needed in total.

gpt-3.5-turbo output:

The request is asking how many pieces of fabric are needed to make a robe, given that it requires 2 pieces of blue fiber fabric and half that amount of white fiber fabric.  
To solve this problem, we need to first find out how many pieces of white fiber fabric are needed. We can do this by dividing the number of blue fiber fabric pieces by 2:  $2 \div 2 = 1$   
So we need 1 piece of white fiber fabric.  
To find the total number of fabric pieces needed, we can add the number of blue and white pieces together:  $2 + 1 = 3$   
Therefore, Answer: 3 pieces of fabric are needed to make the robe.

Answer:

3

Figure 7: A Chinese example of the **MGS** benchmark using **XLT** and the corresponding outputs under the **zero-shot** setting.