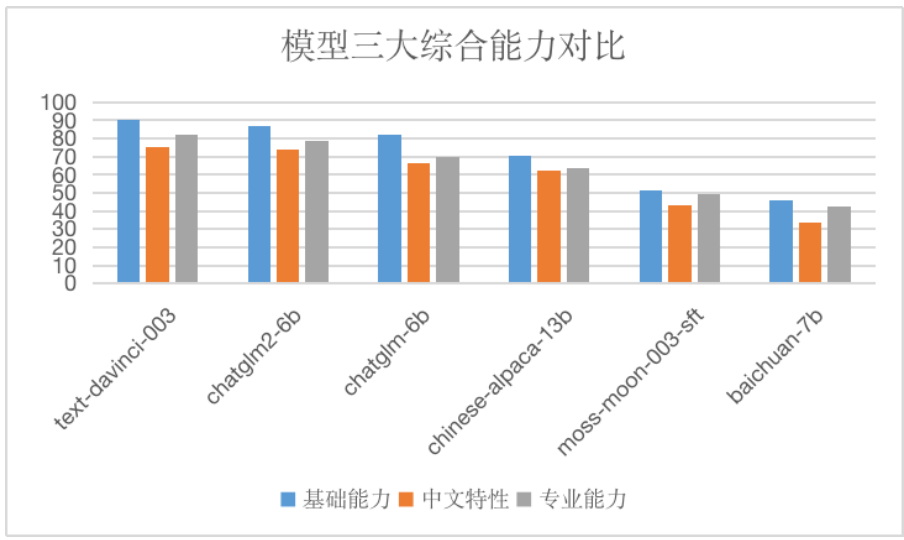
特赛发LLM第一轮评测结果

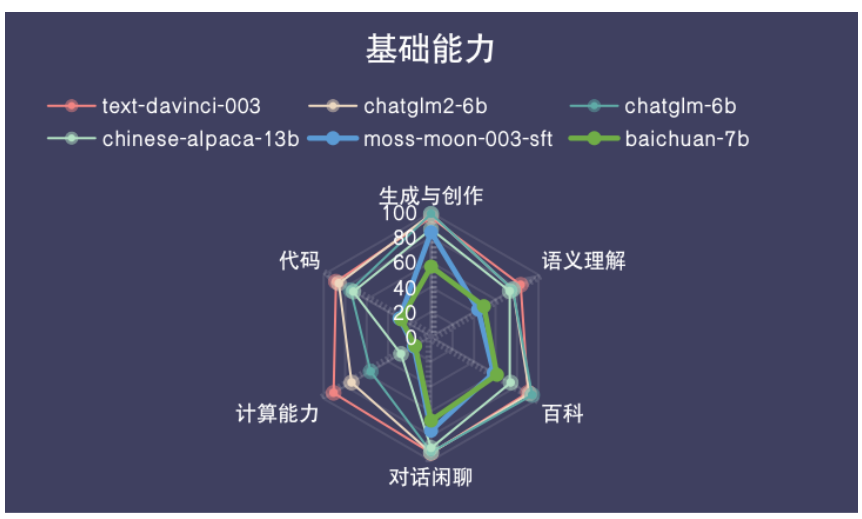
一、模型三大综合能力对比

1.1 三大综合能力对比



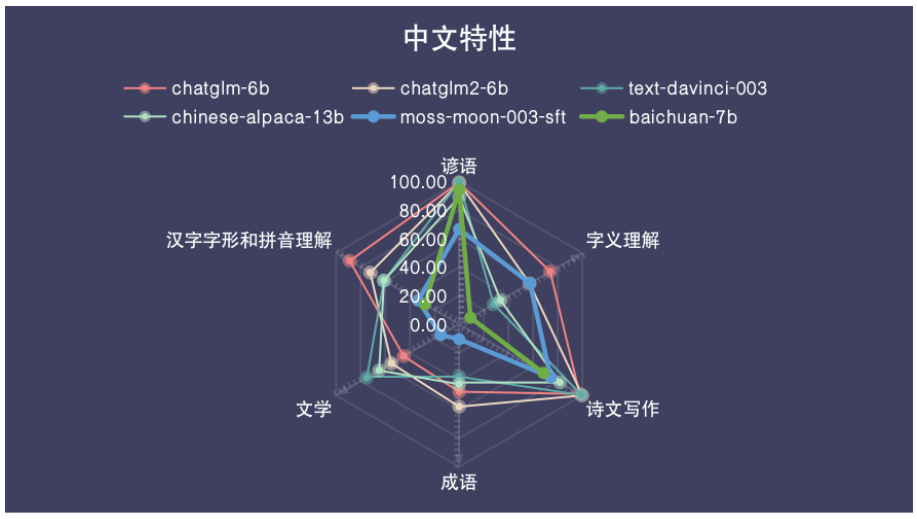
	A	B	C	D	E	F
1	排名	模型名称	平均分	基础能力	中文特性	专业能力
2	1	text-davinci-003	82.56	90.47	75.4	81.81
3	2	chatglm2-6b	79.56	86.69	73.57	78.42
4	3	chatglm-6b	72.81	82.23	66.68	69.52
5	4	chinese-alpaca-13b	65.50	70.68	62.17	63.64
6	5	moss-moon-003-sft	47.79	51.26	42.88	49.23
7	6	baichuan-7b	40.73	46.21	33.33	42.64

1.2 基础能力



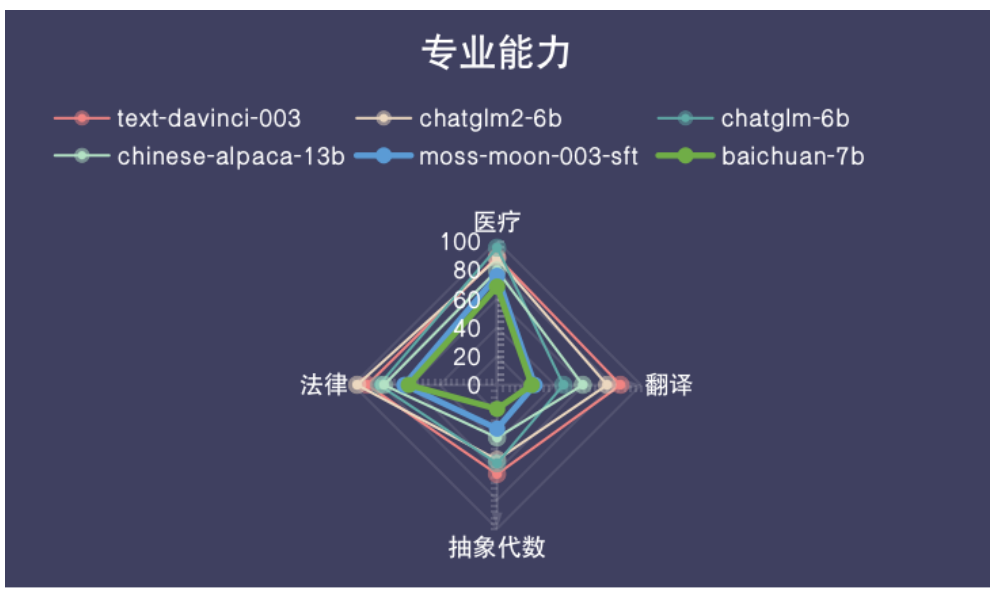
	A	B	C	D	E	F	G	H	I
1	排名	model_name	基础能力						
2			平均分	生成与创作	语义理解	百科	对话闲聊	计算能力	代码
3	1	text-davinci-003	90.47	96.19	83.42	90.3	93.64	90.75	88.55
4	2	chatglm2-6b	86.69	98.73	76.15	91.73	93.64	74	85.9
5	3	chatglm-6b	82.23	99.05	76.07	94.12	93.03	56.25	74.87
6	4	hinese-alpaca-13	70.68	86.67	73.25	74.09	89.7	28	72.39
7	5	oss-moon-003-s	51.26	84.29	43.85	58.27	75.76	16	29.4
8	6	baichuan-7b	46.21	56.51	49.15	61.33	67.58	14.83	27.86

1.3 中文特性



	A	B	C	D	E	F	G	H	I
1	排名	model_name	中文特性						
2			平均分	谚语	字义理解	诗文写作	成语	文学	汉字字形和拼音理解
3	1	chatglm-6b	75.4	99.39	73.89	97.88	47.37	45	88.89
4	2	chatglm2-6b	73.57	99.09	57.22	100	57.89	55	72.22
5	3	ext-davinci-00	66.68	99.7	28.06	99.39	36.84	75	61.11
6	4	nese-alpaca-1	62.17	89.39	33.61	81.82	42.11	65	61.11
7	5	ss-moon-003-	42.88	66.67	57.5	74.24	10.53	15	33.33
8	6	baichuan-7b	33.33	94.55	9.17	68.48			27.78

1.4 专业能力



	A	B	C	D	E	F	G
1	排名	model_name	专业能力				
2			平均分	医疗	翻译	抽象代数	法律
3	1	text-davinci-003	81.81	89.12	85.74	62.08	90.28
4	2	chatglm2-6b	78.42	87.89	76.48	52.08	97.22
5	3	chatglm-6b	69.52	95.44	46.11	54.58	81.94
6	4	chinese-alpaca-13b	63.64	79.65	59.63	36.67	78.61
7	5	moss-moon-003-sft	49.23	75.96	26.11	30.42	64.44
8	6	baichuan-7b	42.64	68.42	24.07	16.67	61.39

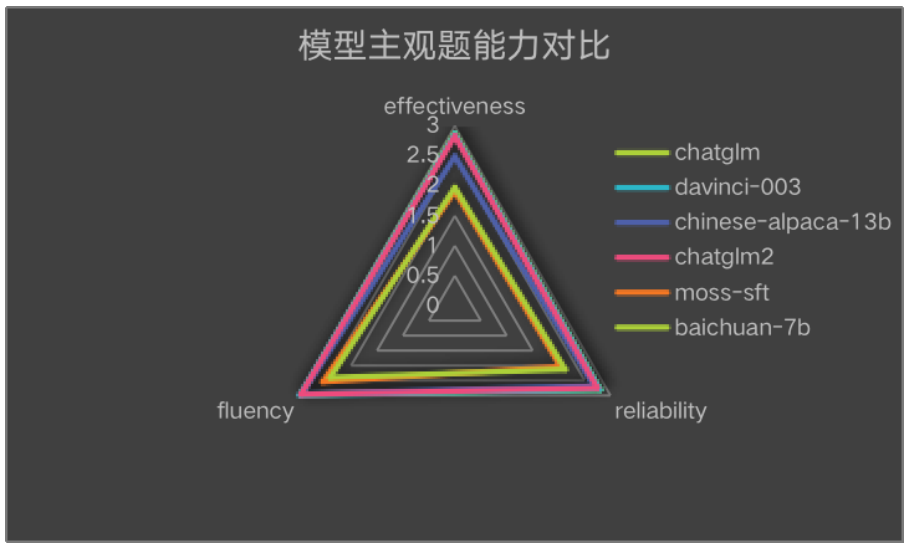
二、模型16个小垂类能力对比

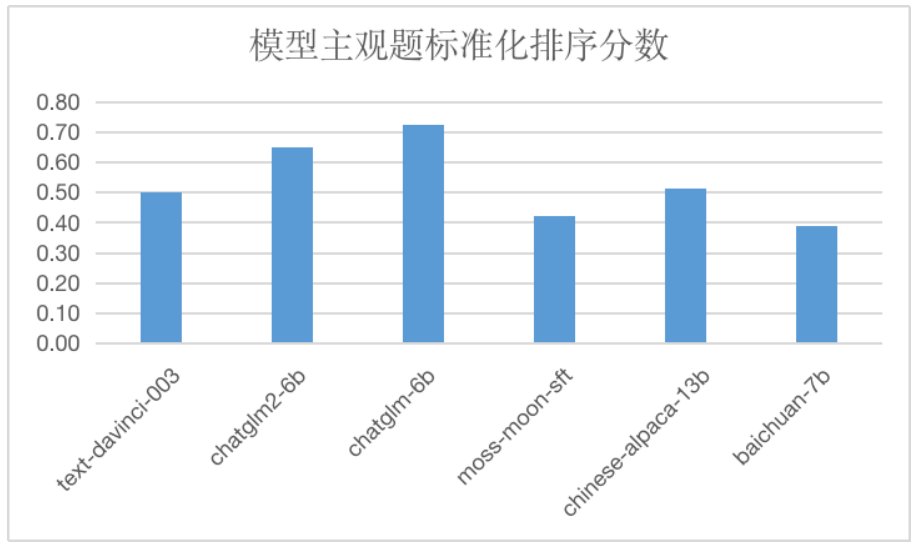
2.1 数据图表展示

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	排名	模型名称	平均分	生成与创作	语义理解	百科	对话闲聊	计算能力	代码	谚语	字义理解	诗文写作	医疗
2	1	text-davinci-003	77.1	96.19	83.42	90.3	93.64	90.75	88.55	99.39	73.89	97.88	89.12
3	2	chatglm2-6b	75.19	98.73	76.15	91.73	93.64	74	85.9	99.09	57.22	100	87.89
4	3	chatglm-6b	69.4	99.05	76.07	94.12	93.03	56.25	74.87	99.7	28.06	99.39	95.44
5	4	chinese-alpaca-13b	61.91	86.67	73.25	74.09	89.7	28	72.39	89.39	33.61	81.82	79.65
6	5	moss-moon-003-sft	45.53	84.29	43.85	58.27	75.76	16	29.4	66.67	57.5	74.24	75.96
7	6	baichuan-7b	38.75	56.51	49.15	61.33	67.58	14.83	27.86	94.55	9.17	68.48	68.42

三、模型主观题能力对比

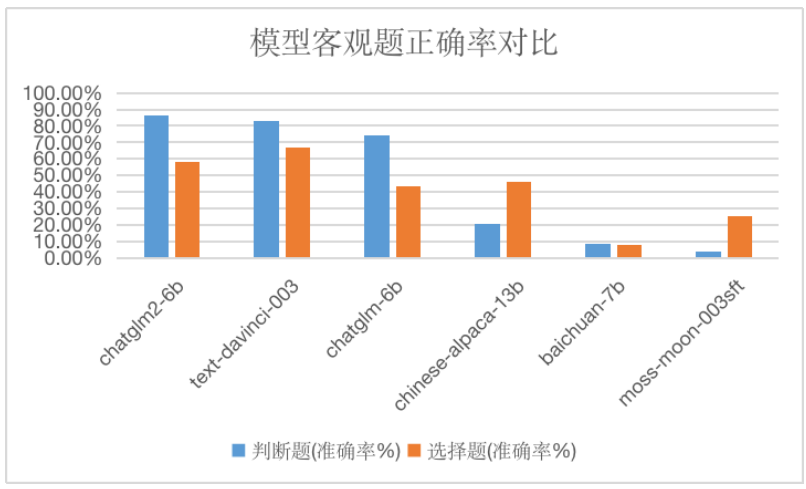
3.1 数据图表展示





	A	B	C	D	E	F	G
1	排名	模型名称	平均分	有效性	可靠性	流畅性	满分
2	1	text-davinci-003	2.89	2.868	2.8	2.988	3
3	2	chatglm-6b	2.88	2.884	2.8	2.96	3
4	3	chatglm2-6b	2.85	2.832	2.748	2.968	3
5	4	chinese-alpaca-13b	2.71	2.492	2.684	2.948	3
6	5	moss-moon-003-sft	2.17	1.912	2.04	2.544	3
7	6	baichuan-7b	2.17	1.976	2.128	2.404	3

四、模型客观题准确率对比



4.1 判断题

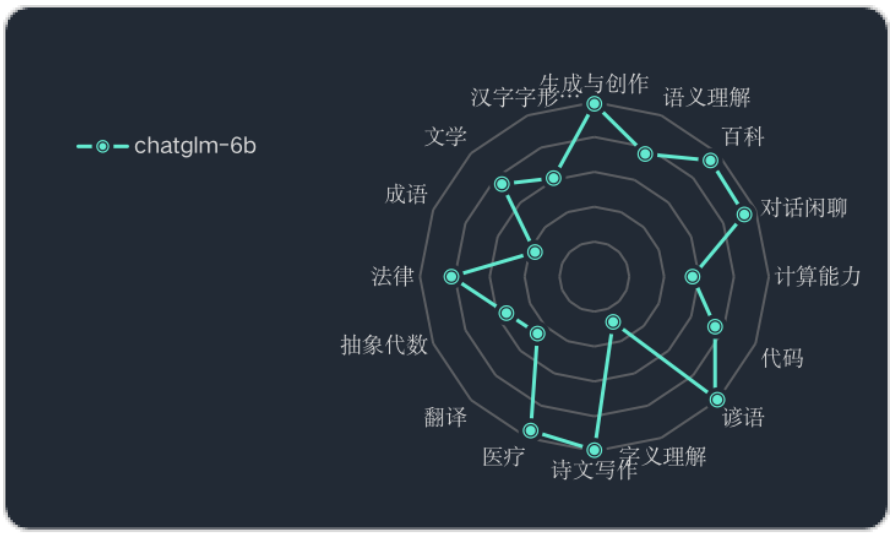
	A	B	C
1	排名	模型名称	判断题(准确率%)
2	1	chatglm2-6b	86.44%
3	2	text-davinci-003	83.05%
4	3	chatglm-6b	74.58%
5	4	chinese-alpaca-13b	20.34%
6	5	baichuan-7b	8.47%
7	6	moss-moon-003sft	3.39%

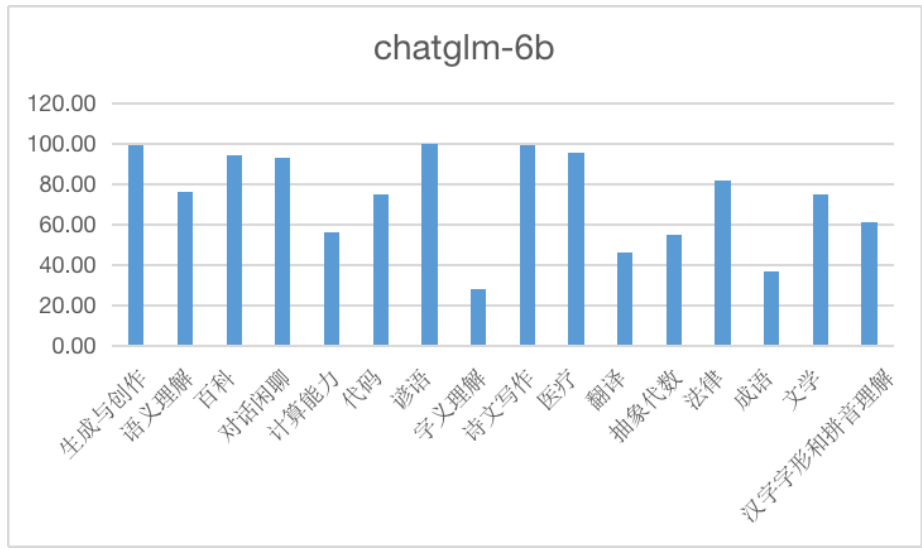
4.2 选择题

	A	B	C
1	排名	模型名称	选择题(准确率%)
2	1	text-davinci-003	67.14%
3	2	chatglm2-6b	57.86%
4	3	chinese-alpaca-13b	45.71%
5	4	chatglm-6b	43.57%
6	5	moss-moon-003sft	25.00%
7	6	baichuan-7b	7.86%

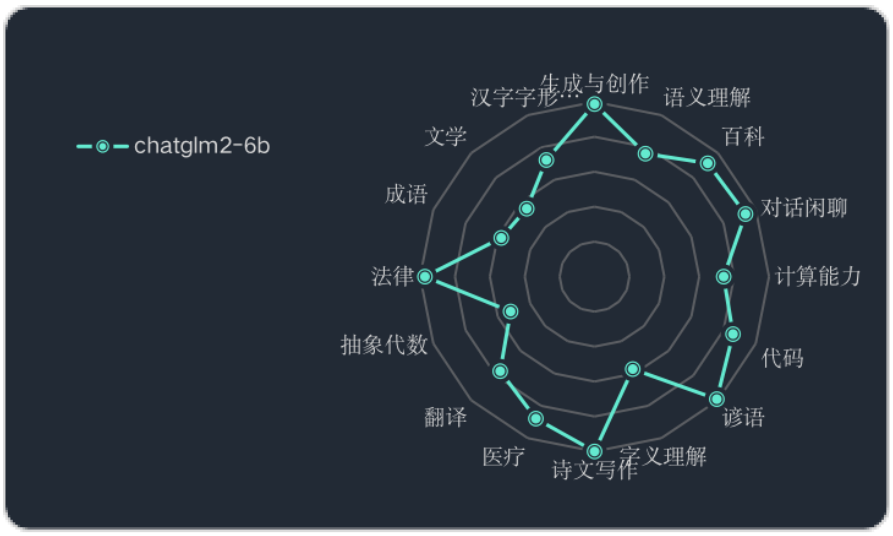
五、单个模型综合能力

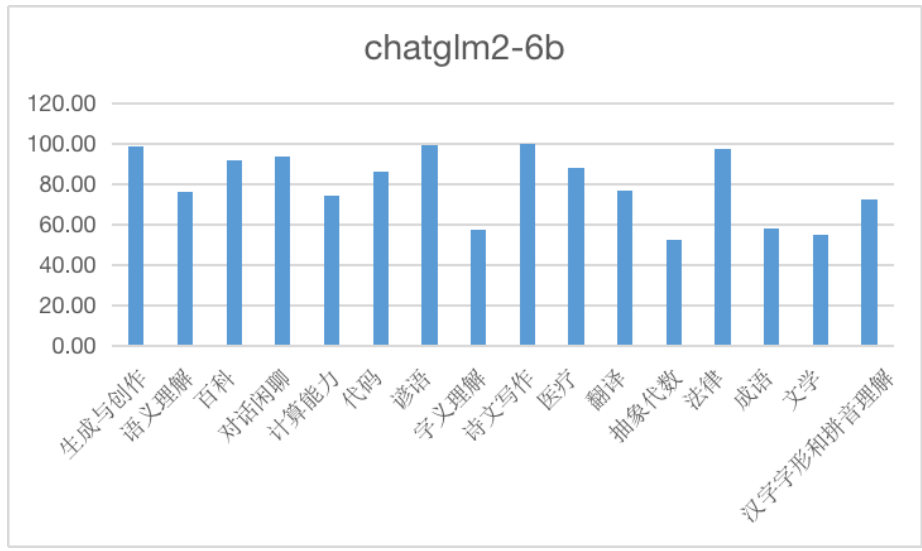
5.1 chatglm-6b



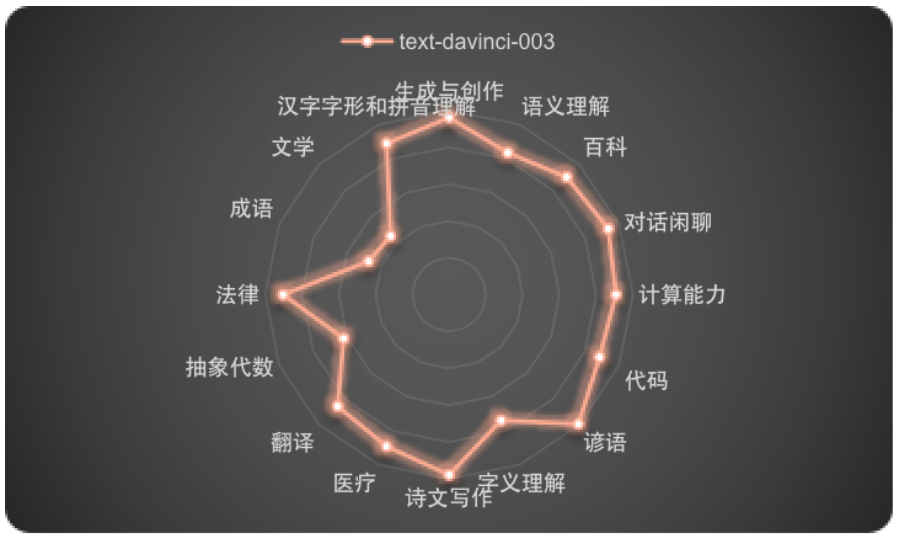


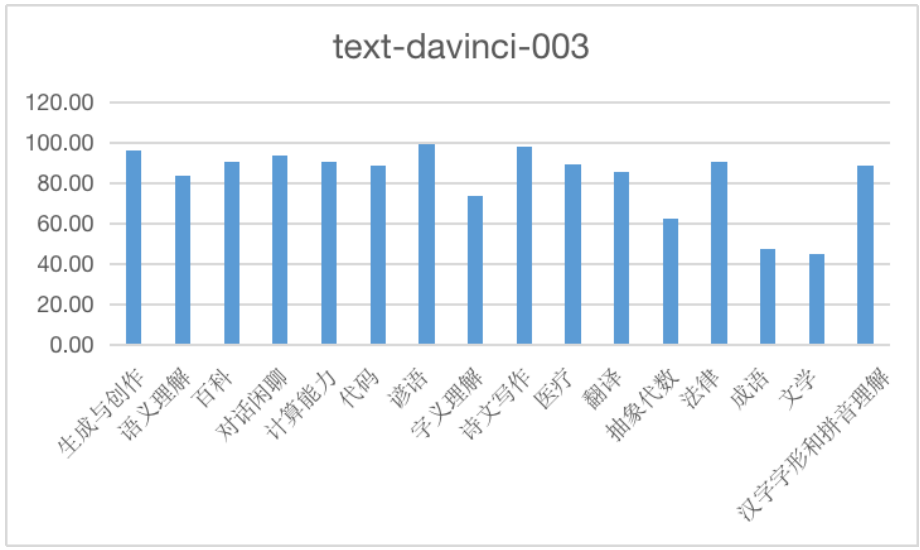
5.2 chatglm2-6b



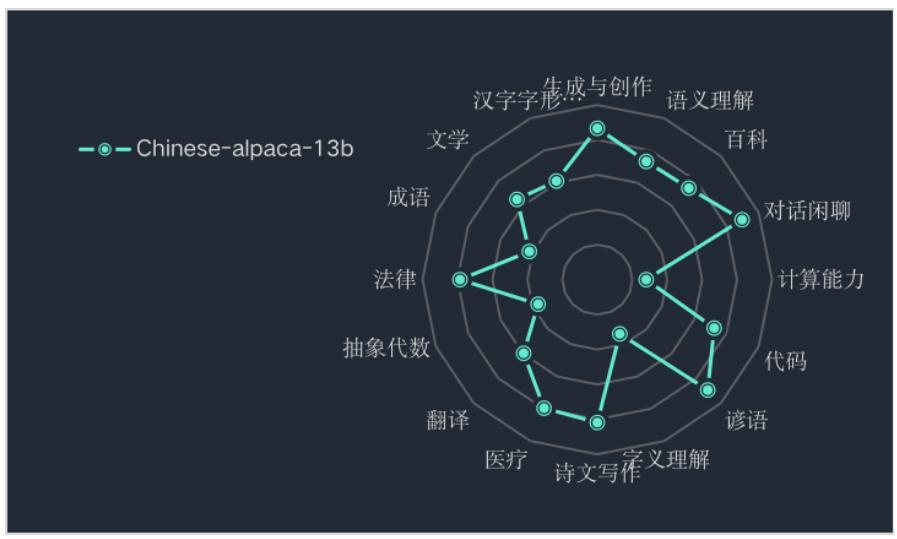


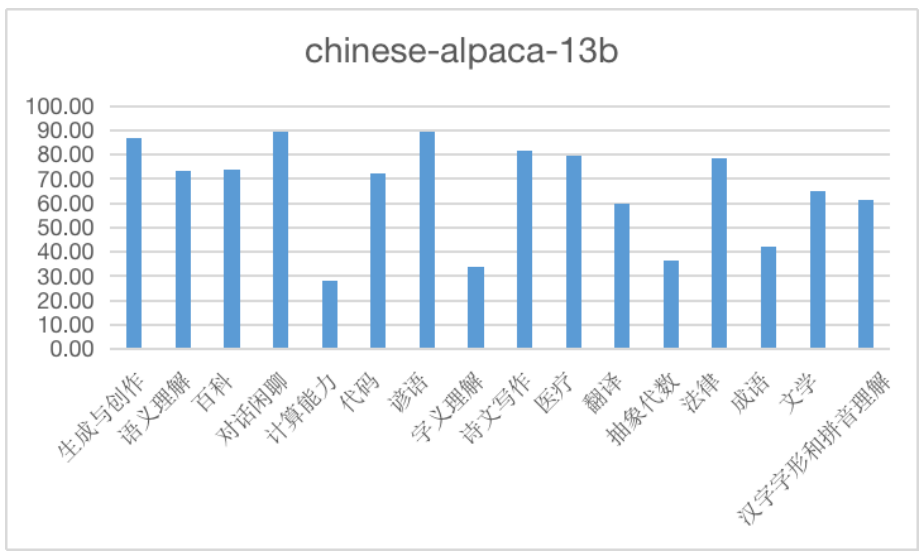
5.3 text-davinci-003



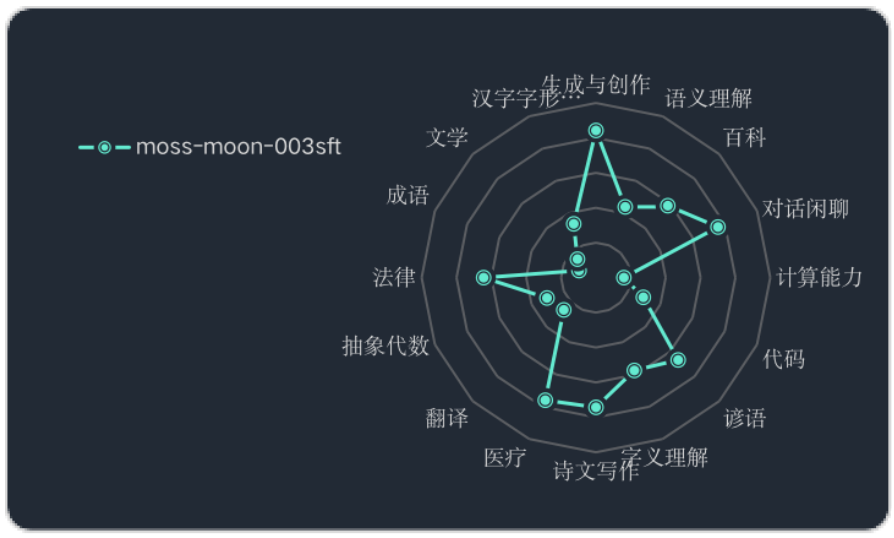


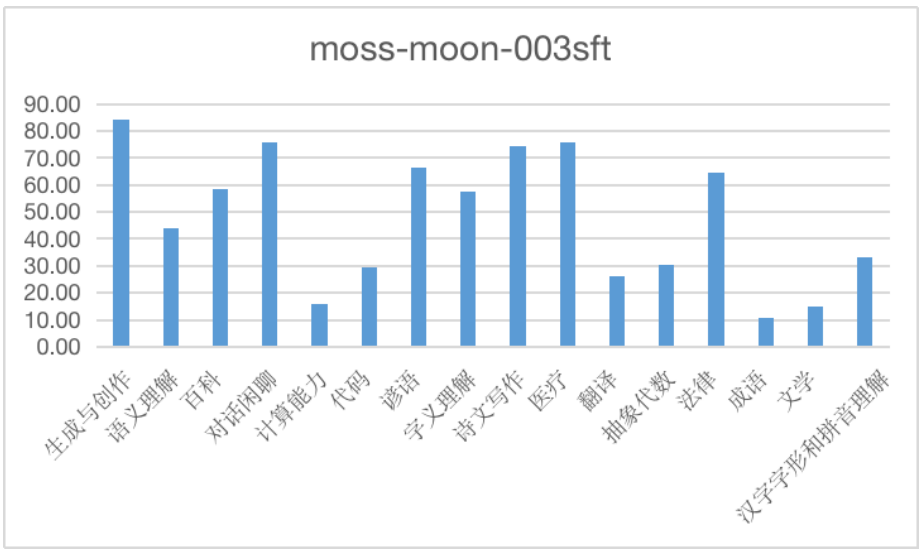
5.4 Chinese-alpaca-13b



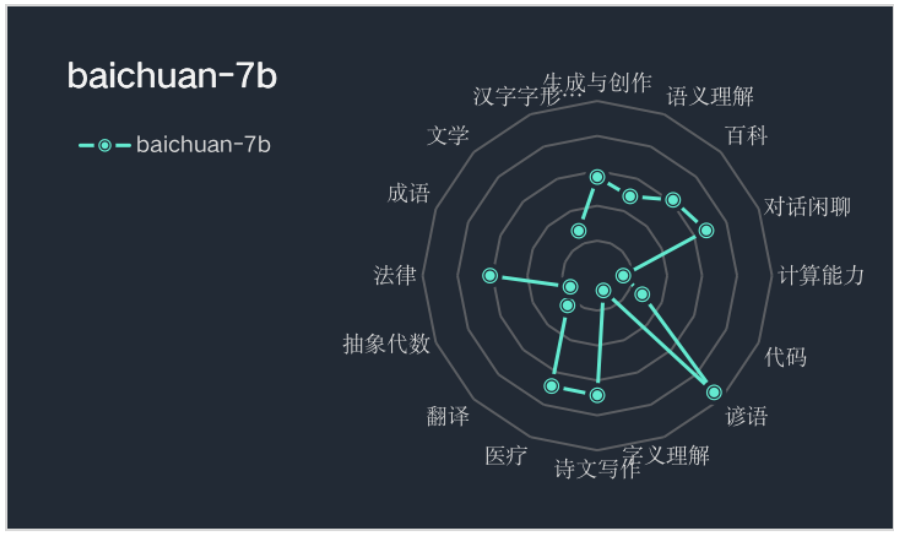


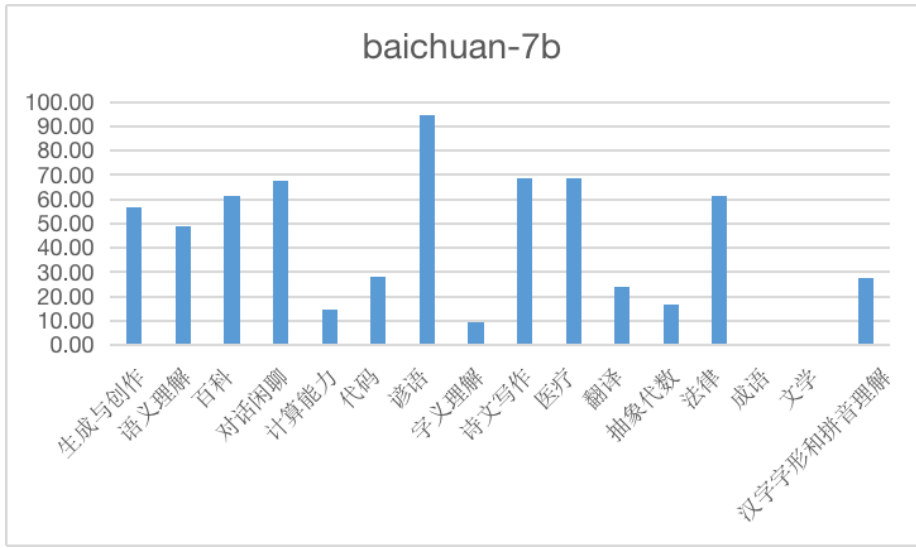
5.5 moss-moon-003sft





5.6 baichuan-7b


baichuan-7b
baichuan-7b
生成与创作
语义理解
百科
对话闲聊
计算能力
代码
谚语
字义理解
诗文写作
医疗
翻译
抽象代数
法律
成语
文学
汉字字形...


baichuan-7b
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
生成与创作
语义理解
百科
对话闲聊
计算能力
代码
谚语
字义理解
诗文写作
医疗
翻译
抽象代数
法律
成语
文学
汉字字形和拼音理解