Relatório 8 - Leitura: Uso e Capacidade de LLM (II)

Vitor Eduardo de Lima Kenor

Descrição da atividade

Neste card vamos analisar o conteúdo dos tópicos 6 e 7 do artigo “A Survey of Large Language Models”.

Começando pelo tópico 6 do artigo, nele é abordado o conceito de prompting no contexto das LLMs e como a qualidade dos prompts influencia seu desempenho em tarefas específicas. O processo de criação de prompts envolve tanto a engenharia manual quanto a otimização automática. O tópico destaca quatro componentes essenciais para criar um prompt eficaz: descrição da tarefa, dados de entrada, informações contextuais e estilo do prompt. Informações contextuais, como documentos ou exemplos de tarefas, ajudam a guiar o modelo. O estilo do prompt pode incluir frases que orientem o modelo a seguir um raciocínio passo a passo ou a assumir um papel específico, como o de um especialista. Ainda no tópico temos sugestões de princípios de design para prompts eficazes, como decompor tarefas complexas em subtarefas mais simples, fornecer exemplos de poucas tentativas few-shot demonstrations e adotar formatos amigáveis para o modelo. Além disso, menciona a estratégia de role-playing, em que o modelo é orientado a assumir um papel específico, como o de um matemático para resolver problemas. Em suma, engenharia de prompts pode ajudar a superar desafios em tarefas de raciocínio matemático e no uso de conhecimento especializado. Mais para frente neste tópico é abordado sobre In-Context Learning (ICL) que se trata de uma abordagem que permite que um modelo de linguagem execute uma tarefa sem precisar de treinamento adicional, ajustando-se apenas com exemplos fornecidos no próprio prompt. No artigo é dito que o ICL foi proposto junto com o GPT-3 e se tornou uma abordagem comum para se utilizar em LLMs. O formato do prompt é formado por uma descrição da tarefa, demonstrações e uma nova entrada para que o LLM gere a resposta.

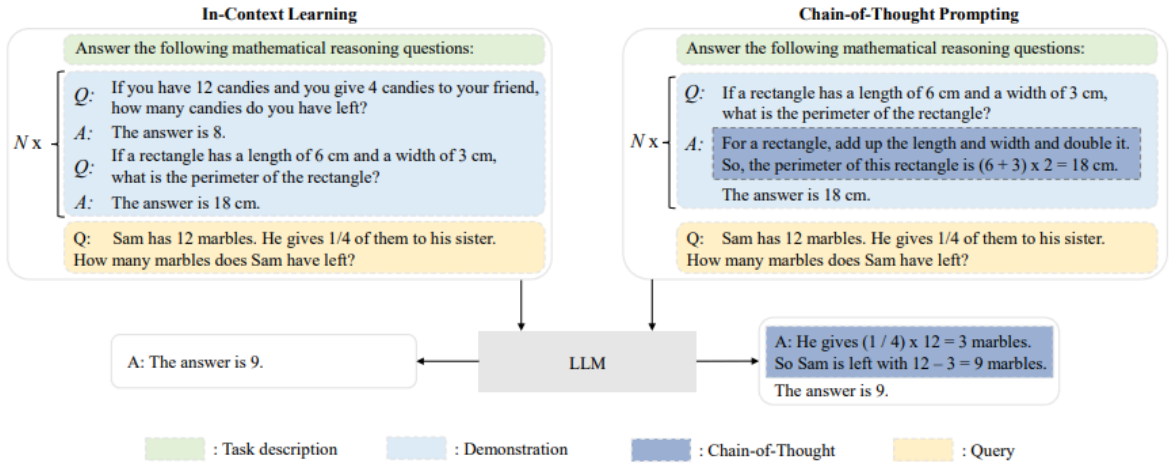
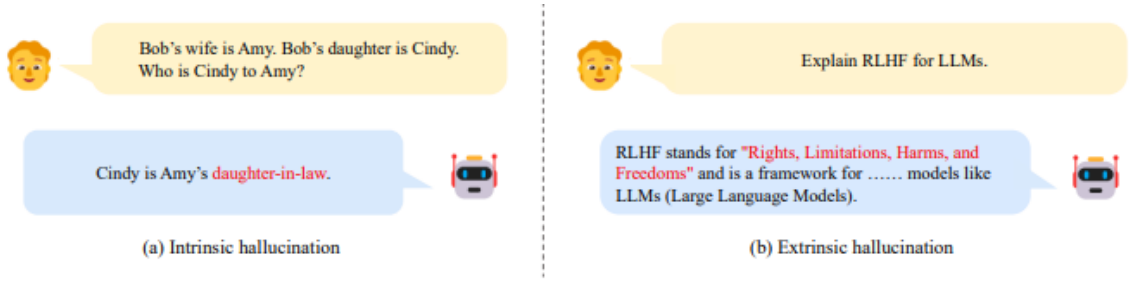


Fig. 14: A comparative illustration of in-context learning (ICL) and chain-of-thought (CoT) prompting. ICL prompts LLMs with a natural language description, several demonstrations, and a test query, while CoT prompting involves a series of intermediate reasoning steps in prompts.

Esta imagem retirada do artigo mostra a diferença entre as abordagens In-Context Learning e Chain-of-Thought Prompting em modelos de linguagem. Segundo o que é dito no artigo, estudos sugerem que modelos maiores, com mais parâmetros, têm maior capacidade de aprender novas tarefas a partir das demonstrações, enquanto modelos menores tendem a depender mais do conhecimento pré-existente. A Chain-of-Thought (CoT) Prompting é uma estratégia aprimorada de prompting que visa melhorar a performance dos LLMs em tarefas complexas de raciocínio, como raciocínio aritmético, de senso comum e simbólico. Ao contrário do ICL, o CoT incorpora etapas intermediárias de raciocínio, ligando a entrada à saída. O CoT é especialmente eficaz para modelos maiores e é mais útil em tarefas que exigem raciocínio passo a passo. Porém, pode prejudicar o desempenho em tarefas mais simples. Para resolver essas questões, foi proposta uma abordagem aprimorada chamada planejamento baseado em prompts, que divide tarefas complexas em subtarefas menores e gera um plano de ações. O planejamento geralmente envolve três componentes: o planejador de tarefas, o executor do plano e o ambiente. O planejador de tarefas, que é realizado pelos LLMs, gera o plano completo, que pode ser apresentado como uma sequência de ações em linguagem natural ou como um código executável. O executor do plano é responsável por realizar as ações e o ambiente oferece feedback sobre a execução. As abordagens baseadas em texto geram planos em linguagem natural, enquanto as baseadas em código produzem planos em linguagens de programação para execução determinística. Finalizando, temos que essas abordagens permitem melhorar o desempenho do planejador iterativamente, ajustando o plano conforme necessário.

Já no tópico 7, ele discute a avaliação das capacidades dos LLMs, abordando testes e benchmarks usados para medir sua eficácia. A avaliação é dividida em três categorias principais: geração de linguagem, utilização do conhecimento e raciocínio complexo. Além disso, é discutido problemas na avaliação dos LLMs, como a confiabilidade das métricas automáticas, e a tendência dos próprios LLMs em favorecer textos gerados por IA. Outro desafio são as alucinações, que é quando as LLMs geram informações falsas ou inconsistentes.



Esta imagem retirada do artigo mostra dois exemplos de alucinação de LLMs. A primeira é a alucinação intrínseca, onde o modelo gera uma resposta que não condiz com os fatos apresentados. E a segunda é a alucinação extrínseca, onde o modelo inventa informações sem base em dados reais. Mais para frente no tópico, são apresentados benchmarks e abordagens para avaliar os LLMs. Alguns exemplos de benchmarks abrangentes como MMLU, BIG-bench, HELM e exames humanos, que testam conhecimento e raciocínio. No final, o tópico apresenta uma avaliação empírica das capacidades de alguns LLMs. Os modelos analisados incluem LLaMA, LLaMA 2, Pythia, Falcon, Vicuna, Alpaca e ChatGLM, além de ChatGPT, Claude, Claude 2 e Davinci002/003. A avaliação abrange quatro áreas

principais: geração de linguagem, utilização de conhecimento, raciocínio complexo e alinhamento humano. Os resultados demonstram diferenças significativas entre os modelos, com os fechados geralmente apresentando melhor desempenho.

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	36.44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Claude	64.47	31.23	18.63	51.22	40.92	13.77	14.57	66.62	34.34
Claude 2	45.20	12.93	19.13	78.04	54.30	21.30	21.06	79.97	35.83
Davinci003	69.98	37.46	18.19	67.07	51.51	17.76	16.68	88.47	28.29
Davinci002	58.85	35.11	19.15	56.70	52.11	20.47	18.45	89.23	29.15
LLaMA 2-Chat (7B)	56.12	12.62	16.00	11.59	38.93	12.96	11.32	72.35	23.37
Vicuna (13B)	62.45	20.49	17.87	20.73	29.04	10.75	11.52	20.69	28.76
Vicuna (7B)	63.90	19.95	13.59	17.07	28.58	9.17	6.64	16.96	26.95
Alpaca (7B)	63.35	21.52	8.74	13.41	17.14	3.24	3.00	49.75	26.05
ChatGLM (6B)	33.34	16.58	13.48	13.42	13.42	4.40	9.20	55.39	16.01
LLaMA 2 (7B)	66.39	11.57	11.57	17.07	30.92	5.15	2.51	24.16	28.06
LLaMA (7B)	67.68	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78
Falcon (7B)	66.89	4.05	10.00	10.37	28.74	10.78	8.46	4.08	23.91
Pythia (12B)	61.19	5.43	8.87	14.63	15.73	1.99	4.72	11.66	20.57
Pythia (7B)	56.96	3.68	8.23	9.15	10.16	1.77	3.74	11.03	15.75

Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Claude	81.80	54.95	73.23	59.95	47.65	70.81	20.18	76.87	47.72/23.00
Claude 2	71.60	50.75	58.34	66.76	74.50	82.87	32.24	77.61	34.96/19.20
Davinci003	74.40	62.65	69.70	64.60	61.07	57.16	17.66	65.67	64.08/32.40
Davinci002	69.80	47.81	57.01	62.55	67.11	49.96	14.28	76.87	29.66/15.20
LLaMA 2-Chat (7B)	45.62	74.01	43.84	43.40	38.93	9.63	2.22	11.19	24.51/5.60
Vicuna (13B)	43.65	70.51	45.97	53.55	36.91	18.50	3.72	8.96	22.74/5.00
Vicuna (7B)	43.84	69.25	46.27	44.25	36.24	14.03	3.54	1.49	6.90/1.40
Alpaca (7B)	47.82	69.81	47.55	39.35	40.27	4.93	4.16	4.48	0.00/0.00
ChatGLM (6B)	30.42	29.27	33.18	14.05	14.09	3.41	1.10	0.00	0.00/0.00
LLaMA 2 (7B)	44.81	74.25	41.72	43.95	35.75	10.99	2.64	8.96	0.00/0.00
LLaMA (7B)	42.42	73.91	41.46	39.95	34.90	10.99	3.12	2.24	0.00/0.00
Falcon (7B)	39.46	74.58	42.53	29.80	24.16	1.67	0.94	7.46	0.00/0.00
Pythia (12B)	37.02	65.45	41.53	32.40	26.17	2.88	1.96	5.22	3.68/0.60
Pythia (7B)	34.88	61.82	41.01	29.05	27.52	1.82	1.46	7.46	10.75/1.80

Models	Human Alignment					Tool Manipulation			
	TfQA↑	C-Pairs↓	WinoGender↑	RTP↓	HaluEval↑	HotpotQA↑	Gorilla-TH↑	Gorilla-TF↑	Gorilla-HF↑
ChatGPT	69.16	18.60	62.50/72.50/79.17	3.07	66.64	23.80	67.20	44.53	19.36
Claude	67.93	32.73	71.67/55.00/52.50	3.75	63.75	33.80	22.04	7.74	7.08
Claude 2	71.11	10.67	60.00/60.00/55.83	3.20	50.63	36.4	61.29	22.19	23.67
Davinci003	60.83	0.99	67.50/68.33/79.17	8.81	58.94	34.40	72.58	3.80	6.42
Davinci002	53.73	7.56	72.50/70.00/64.17	10.65	59.67	26.00	2.69	1.02	1.00
LLaMA 2-Chat (7B)	69.77	48.54	47.50/46.67/46.67	4.61	43.82	4.40	0.00	0.00	0.22
Vicuna (13B)	62.30	45.95	50.83/50.83/52.50	5.00	49.01	11.20	0.00	0.44	0.89
Vicuna (7B)	57.77	67.44	49.17/49.17/49.17	4.70	43.44	6.20	0.00	0.00	0.33
Alpaca (7B)	46.14	65.45	53.33/51.67/53.33	4.78	44.16	11.60	0.00	0.00	0.11
ChatGLM (6B)	63.53	50.53	47.50/47.50/46.67	2.89	41.82	4.00	0.00	0.00	0.00
LLaMA 2 (7B)	50.06	51.39	48.83/48.83/50.83	6.17	42.23	3.80	0.00	0.00	0.11
LLaMA (7B)	47.86	67.84	54.17/52.50/51.67	5.94	14.18	1.60	0.00	0.00	0.11
Falcon (7B)	53.24	68.04	50.00/50.83/50.00	6.71	37.41	1.00	0.00	0.00	0.00
Pythia (12B)	54.47	65.78	49.17/48.33/49.17	6.59	27.09	0.40	0.00	0.00	0.00
Pythia (7B)	50.92	64.79	51.67/49.17/50.00	13.02	25.84	0.20	0.00	0.00	0.00

E esta tabela retirada do artigo mostra a comparação do desempenho dos modelos. A tonalidade das fontes laranja e azul indicam as ordens de desempenho dos resultados em modelos de código fechado e de código aberto, respectivamente.

Conclusões

Os tópicos que foram analisados neste card exploram aspectos fundamentais do desempenho e avaliação de LLMs. As abordagens In-Context Learning e Chain-of-Thought foram detalhadas como formas de melhorar o desempenho em tarefas complexas. Além disso, o planejamento baseado em prompts foi apresentado como uma solução para dividir

tarefas complexas em subtarefas menores, melhorando o desempenho iterativamente. Depois vimos que a avaliação das capacidades dos modelos é feita por meio de benchmarks como MMLU e BIG-bench, testando geração de linguagem, conhecimento e raciocínio. No entanto, desafios como alucinações e a confiabilidade das métricas automáticas complicam essa avaliação. E no final vimos que a comparação empírica de diferentes modelos demonstrou diferenças significativas, com modelos de código fechado geralmente apresentando melhores resultados.

Referências

Link do artigo “A Survey of Large Language Models”:
https://arxiv.org/pdf/2303.18223