Nome: Fernando Buligon Antunes **Data:**30/04/2025

Large Language Models (Tópicos 6 e 7)

O primeiro tópico retrata sobre a utilização das LLMs após o pré treinamento e ajustes. A principal técnica usada é a modelação de prompts estratégicos para a solução de diversas tarefas.

A qualidade dos prompts vai alterar significativamente a qualidade das saídas em tarefas específicas ou gerais, por isso houve uma série de estudos que tentaram buscar qual seria um modelo ideal de prompt para tarefas específicas.

Criação de prompts, o processo de criação de prompt também é chamado de engenharia de prompt. Um prompt bem organizado é de extrema utilidade para extrair o máximo das habilidades de uma LLM em tarefas, para isso existem alguns componentes que são interessantes de se ter em mente. Ingredientes chaves, a estrutura básica de um bom prompt é separada em quatro partes principais. Descrição da tarefa, se trata de uma instrução contendo o comportamento esperado da LLM, é importante manter o essencial, de forma clara e detalhada. Dados de entrada, de maneira geral podem ser descritos em linguagem natural, mas caso se tratem de dados estruturados, como tabelas e grafos, são transformados em sequências ou códigos para serem entendidos. Informação contextual, em adição as outras partes, é essencial informar o contexto da pergunta para que o modelo consiga responder de maneira mais coerente com o que você quer. Estilo de prompt, dependendo da LLM, o estilo do prompt pode variar, então devem ser feitas as alterações que melhor se enquadrem com o modelo atual. Princípios de design, baseados nos ingredientes chaves, há alguns princípios a serem seguidos também, como expressar o objetivo de forma clara, separar em sub tarefas detalhadas e fáceis, prover algumas demonstrações e usar um formato agradável ao modelo.

Otimização de prompts, apesar da criação manual de prompts ser mais intuitiva, acaba consumindo bastante tempo e também tem a questão de que os modelos são altamente sensitivos aos prompts, então prompts impróprios levam a uma performance ineficaz do modelo. Com base nesse problema, estudos propuseram abordagens de otimização automática para prompts discretos e contínuos para conseguir alcançar a performance esperada do modelo.

Otimização de prompt discreto, um prompt discreto é composto basicamente por uma sequência de tokens da linguagem natural, de forma simples e flexível. Existem algumas abordagens para esse tipo de prompt, como por exemplo:

- Baseada em gradiente: Tentam melhorar o prompt ajustando palavras com base no que aumenta a chance de uma boa resposta do modelo com base em gradientes;
- Baseada em RL: Tentam melhorar o prompt com base em recompensas que refletem o desempenho do modelo naquele prompt, para isso existem alguns métodos como o RLPrompt e o TEMPERA;
- Baseada em edição: Os dois métodos acima possuem uma demanda computacional relativamente alta, então outra linha de trabalho busca editar prompts já existentes;
- Baseada em LLM: Usam as próprias LLMs para gerar os prompts iniciais, depois os que possuírem a melhor acurácia são selecionados.

Otimização de prompt contínuo, consiste em uma série de embeddings contínuos, que podem ser otimizados com base na atualização de gradiente baseada na loss das tarefas anteriores. Possui dois tipo de abordagens:

- Aprendizado de prompt com dados suficientes: Nesse abordagem, a maioria dos métodos tratam prompts contínuos como modelos treináveis de parâmetros e depois usam o aprendizado supervisionado para otimização reduzindo a cross-entropy loss baseada em dados suficientes de tarefas anteriores;
- Transferência de prompt com dados escassos: abordagens com aprendizado supervisionado demandam dados de treino suficientes, então não funcionam muito bem em domínios com poucos dados, e essa técnica foi proposta para solucionar esse problema. Consiste em primeiramente aprender um único prompt contínuo para várias tarefas, depois esse prompt é usado para inicializar um prompt para uma tarefa específica.

Outra abordagem especial de formato de prompts para trabalhar com LLMs é a ICL (In-Context Learning, em português, aprendizado em contexto). O prompt é composto basicamente por linguagem natural formatada, contendo a descrição da tarefa e alguns exemplos. Apesar da ideia geral ser simples, existem duas diferentes maneiras de se trabalhar com esse método, como a heurística que é simples, com o adicional de baixo custo, e também a abordagem baseada em LLM, que basicamente usam LLMs durante o processo para avaliar como está indo.

CoT (Chain-of-Thought, em português, cadeia de pensamento) prompting, é uma estratégia de prompt com o objetivo de aprimorar a resposta das LLMs em tarefas complexas. Como o próprio nome já sugere, essa técnica envolve montar uma cadeia de pensamentos que demonstram um caminho lógico que o modelo deve seguir.

Apesar de tudo, essas duas técnicas ainda sofrem com tarefas mais complexas, como questões de raciocínio matemático ou perguntas que exigem múltiplos passos. Para resolver esse problema, foi proposto o prompt baseado em planejamento, que consiste em quebrar um problema complexo em tarefas menores e gerar um plano de ações para concluir a tarefa.

Então basicamente consiste em uma abordagem que divide as tarefas complexas em três componentes principais: um planejador de tarefas (geralmente a própria LLM), um executor do plano e o ambiente onde as ações são realizadas. O planejador gera uma sequência de ações em linguagem natural ou código com base no objetivo da tarefa. O executor realiza essas ações, e o ambiente fornece o feedback, que pode vir do próprio modelo ou de fontes externas como interpretadores de código ou mundos simulados. Esse processo ocorre de forma iterativa, com o planejador refinando o plano com base no feedback recebido do anterior.

A geração de planos pode seguir dois formatos principais: baseada em texto, que é mais simples, e baseada em código, oferecendo mais precisão na execução. Além disso, técnicas como raciocínio, retrocesso (backtracking) e uso de memória de longo prazo são utilizadas para melhorar continuamente os planos gerados. Essa abordagem amplia significativamente a capacidade das LLMs em tarefas que exigem múltiplas etapas, raciocínio profundo ou persistência de memória.

O Segundo tópico é sobre como as LLMs são avaliadas, examinando a efetividade de cada uma sobre tarefas específicas, com base no desempenho é possível medir a

superioridade entre uma e outra. Existem três tipos de habilidades básicas que são usadas para fazer essa medição:

- Geração de linguagem: pode ser separado em três classificações diferentes. Modelação de linguagem, consiste em uma das tarefas mais fundamentais de qualquer LLM, em que o objetivo é conseguir prever o próximo token se baseando nos anteriores, fazendo uso de capacidades básicas de entendimento e geração. Geração de texto condicional, é a geração de texto com base no que foi pedido, por exemplo, uma tradução, um resumo, reescrita de maneira formal, e etc. Síntese de código, as LLMs demonstraram grande capacidade em geração de linguagem formal, especialmente códigos, especialmente porque códigos, ao contrário de linguagem natural, podem ser compilados e checados para testar se estão corretos. A geração de linguagem infelizmente possui dois problemas principais. Avaliação inconfiável de geração, apesar das LLMs serem capazes de gerar textos semelhantes ao humano no quesito de qualidade, essa qualidade pode ser subestimada por métricas automáticas. Geração especializada com performance abaixo da esperada, as LLMs ainda tem dificuldade em tarefas que exigem domínios muito específicos não inseridos em sua base de dados.
- Utilização de conhecimento: é uma habilidade importante que consiste no modelo ser capaz de usar conhecimento factual para resolver tarefas complexas, como perguntas baseadas em senso comum e complementação de fatos. Isso requer que a LLM use adequadamente o conhecimento do corpus usando para o pré treinamento ou buscando dados externos quando necessário. Essa habilidade pode ser dividida em três tarefas principais. Livro fechado PR (Perguntas e Respostas), a LLM deve responder apenas com o conhecimento obtido do corpus no pré treinamento, se baseando no contexto da tarefa, sem buscar por fontes externas. Livro aberto PR, o modelo pode responder fazendo buscas externas ao corpus do pré treinamento. Completude de conhecimento, nessas tarefas a LLM deve ser considerada como uma base de conhecimentos, em que seja capaz de prever ou completar as partes faltantes de unidades de conhecimento, essa parte pode ser separada em duas partes, a completude de conhecimentos gráficos e completude de fatos. Apesar de que o progresso atual seja bom, as LLMs sofrem principalmente com dois grandes problemas, primeiro, alucinação, em que a LLM não consegue encontrar os dados então ela decide apenas criar alguma resposta para responder algo, e o segundo, dados atualizados, as LLMs encontram grandes dificuldades quando pedem algo muito recente, aquele dado requisitado ainda não foi incluído na base de dados dela, então não há como ela saber com certeza.
- Raciocínio complexo: se refere a habilidade de fazer uso do compreendimento e da lógica para responder, assim como as outras habilidades, essa também é separada em três classes. Raciocínio de conhecimento, busca responder as perguntas com base na relação dos fatos. Raciocínio simbólico, manipula os símbolos com regras formais para atingir os objetivos. Raciocínio matemático, são responsáveis pelos problemas que precisam de lógica e conhecimentos da matemática e da computação para resolução de problemas ou geração de provas. Essa habilidade possui dois problemas principais, o primeiro, raciocínio inconsistente, em que a LLM talvez acerte a resposta seguindo um caminho de raciocínio incorreto, ou, consegue

encontrar o caminho certo mas chega na resposta incorreta, o segundo, envolve computação numérica, as LLMs possuem dificuldades em fazer contas, especialmente quando envolve símbolos que não estavam presente durante seu pré treinamento.

Em adição a essas três habilidades básicas, também existem algumas habilidades superiores que são mais simples de definir:

- Alinhamento humano: as LLMs devem responder de acordo com as necessidades e valores humanos. Essa é uma habilidade chave para que a LLM possa ser usada publicamente. O alinhamento é feito geralmente com feedback humano.
- Interação com o ambiente externo: as LLMs possuem a habilidade de receber feedbacks de ambientes externos, e trabalhar com base neles. Para testar essa habilidade diversos ambientes de inteligência artificial foram usados.
- Manipulação de ferramentas: caso necessário quando estão resolvendo problemas muito complexos, as LLMs podem acessar ferramentas externas, isso ajuda elas a resolverem problemas em áreas que não possuem domínio. Em adição, elas também conseguem criar suas próprias ferramentas para resolver tarefas específicas de forma autônoma.

Anteriormente foram discutidas as habilidades básicas e avançadas das LLMs, agora são vistos alguns benchmarks e abordagens para avaliação. Foram mostradas quais são os benchmarks mais usados:

- MMLU: é um benchmark versátil para avaliação em larga escala da compreensão de tarefas múltiplas, cobrindo áreas humanas e exatas. A dificuldade dessas tarefas variam de fáceis até difíceis.
- BIG-bench: é um benchmark que é composto por 204 tarefas que cobrem diversos assuntos das diversas áreas. Aumentando a escala, os modelos foram capazes de superar o desempenho médio humano em 65%.
- HELM: é um benchmark compreensivo que implementa 16 cenários e 7 categorias de métricas para avaliar as LLMs.
- Benchmarks a nível humano: tentam avaliar o entendimento das LLMs em questões designadas para testar humanos, esses testes também cobrem uma grande variedade de cenários, variando a dificuldade, área, linguagem, e etc.

Além desses benchmarks gerais, existem outros que não cobrem uma variedade tão grande de casos, esses outros geralmente focam em domínios específicos, o que faz bastante sentido dependendo de qual for o objetivo da LLM. Depois de mostrar os benchmarks, são mostradas algumas abordagens de avaliação, que variam de acordo com o tipo da LLM.

Avaliação de LLMs base, LLMs bases são os checkpoints do modelo obtido após o pré treinamento. Então o foco são as habilidades básicas, como geração de linguagem, utilização do conhecimento e raciocínio complexo. Para avaliar usam benchmarks comuns, avaliações por humanos e por modelos.

Avaliação de LLMs especializadas, LLMs especializadas são os checkpoints adaptados para algum domínio ou aplicação específica como saúde ou finanças, então não faria sentido avaliar essas LLMs somente em habilidades gerais, a avaliação tem que ser adaptada para o que o modelo faz, então é feita uma combinação dos benchmarks para que ela seja avaliada nesses dois âmbitos, gerais e especializados. Assim como na avaliação de

LLMs básicas, existem diferentes métodos de avaliação, os que usam benchmarks, humanos e modelos.