Resumo Artigos

gustavo.gbaggio

April 2025

1 Resumo do Artigo: OPT-R (2023)

O estudo *OPT-R* (Meta AI, 2023) investiga o impacto da inclusão de explicações durante o *finetuning* e o *prompting* de modelos de linguagem de grande porte (LLMs) na resolução de tarefas de raciocínio. Utilizando a família de modelos OPT (com 1.3B, 6.7B e 13B parâmetros), os autores comparam três variantes: o modelo base (OPT), um modelo *finetuned* sem explicações (OPT-R) e um modelo *finetuned* com explicações (OPT-RE). Também avaliam três formas de *prompting*: *zero-shot*, *few-shot* e *few-shot* com explicações. A combinação dessas dimensões resultou em 27 configurações, avaliadas em 57 tarefas extraídas do benchmark SUPER-NATURALINSTRUCTIONS, totalizando 6.156 testes.

Corpo de Dados para Finetuning Foram utilizados conjuntos de dados com explicações associadas, incluindo: AQUA-RAT, CoQA, CoS-E, ECQA, ESNLI, GSM8K, ProofWriter e StrategyQA. Essas bases cobrem uma diversidade de tarefas que exigem raciocínio numérico, lógico, textual e de senso comum.

Resultados Principais Os modelos *finetuned* apresentaram melhorias significativas em diversas habilidades de raciocínio, destacando-se:

- **Numérico:** OPT (44.8), OPT-R (65.2), OPT-RE (64.7)
- **Analógico:** OPT (49.0), OPT-R (62.9), OPT-RE (60.8)
- **Contagem:** OPT (19.8), OPT-R (13.1), OPT-RE (31.3)
- **Físico:** OPT (38.2), OPT-R (37.8), OPT-RE (49.1)
- **Entailment:** OPT (42.6), OPT-R (47.2), OPT-RE (51.6)

Por outro lado, houve degradação de desempenho em habilidades como:

- **Argumentação:** OPT (57.9), OPT-R (46.1), OPT-RE (48.7)
- **Entailment Dedutivo:** OPT (36.0), OPT-R (29.0), OPT-RE (29.4)
- **Raciocínio de senso comum:** OPT (33.4), OPT-R (29.7), OPT-RE (28.8)

Impacto das Explicações Durante o *finetuning*, a inclusão de explicações impactou positivamente principalmente nas tarefas numéricas, físicas e de *entailment* textual. Já durante o *prompting*, as explicações só mostraram efeitos relevantes no modelo base (OPT), e não nos modelos já *finetuned*. A Tabela 1 resume os efeitos médios das explicações nos diferentes modelos.

Conclusões O estudo conclui que o *finetuning* com explicações oferece ganhos modestos, mas consistentes, em tarefas específicas de raciocínio. As habilidades que mais se beneficiam são: raciocínio numérico, analógico, físico, de contagem e textual. Já para o *prompting*, explicações são mais úteis quando o modelo ainda não foi *finetuned*. A análise destaca também quais tarefas sofrem ou permanecem inalteradas com a inclusão de explicações.

Limitações A avaliação foi limitada aos modelos OPT e a um conjunto específico de bases abertas com explicações. Resultados podem não generalizar para LLMs maiores ou treinados com dados fechados. Além disso, não foram consideradas estratégias alternativas de *finetuning*, como *parameter-efficient tuning*.

Utilidade Esse artigo pode ser útil para nosso trabalho, pois sugere diversas ideias de tarefas que podem ser utilizadas para avaliar o desempenho do modelo após o finetuning. Também fornece sugestões de bases de dados para essas tarefas. O estudo fornece um comparativo entre os modelos sem finetuning, com prompting e com três tipos de finetuning, o que constitui parcialmente o nosso objetivo, e pode servir como base de inspiração e referência para o nosso trabalho.

2 Resumo do Artigo: LLM-Adapters (2023)

O artigo *LLM-Adapters* (Hu et al., 2023) investiga métodos de *parameter-efficient fine-tuning* (PEFT) para modelos de linguagem de grande porte (LLMs), com foco em arquiteturas baseadas em adaptadores. O trabalho propõe a **LLM-Adapter**, uma estrutura unificada que permite a integração de diferentes tipos de adaptadores (Series, Paralelo, Reparametrização e Prompt-based) em LLMs abertos, como *LLaMA*, *BLOOMz* e *GPT-J*.

Contribuições Principais

- Desenvolvimento do framework **LLM-Adapter** para facilitar experimentos com métodos PEFT.
- Construção de dois conjuntos de dados de *fine-tuning*:
  - **Math10K**: com 10 mil amostras de raciocínio matemático geradas com auxílio do ChatGPT.

  - **Commonsense170K**: com 170 mil exemplos formatados de tarefas de raciocínio de senso comum.

- Estudo empírico detalhado sobre posicionamento, configuração e impacto de adaptadores em tarefas ID e OOD.

Métodos Avaliados Foram avaliadas quatro principais categorias de PEFT:

- **Prompt-based**: Prompt Tuning, Prefix-Tuning.
- **Reparametrização**: LoRA, KronA.
- **Series Adapter**: Houlsby Adapter, AdaMix, Compacter.
- **Parallel Adapter**: Parallel Adapter, Ladder-Side Tuning.

Melhor Posicionamento dos Adaptadores

- **Series Adapter**: após camadas MLP (59.5% acurácia média).
- **Parallel Adapter**: dentro das camadas MLP (61.7%).
- **LoRA**: após Attention e MLP simultaneamente (60.0%).

Configurações Otimizadas

- **Prefix-Tuning**: 10 tokens virtuais.
- **Series/Parallel**: bottleneck size = 256.
- **LoRA**: rank = 32.

Resultados em Raciocínio Matemático

- **LLaMA-13B + LoRA**: 65.4% de acurácia média (superando GPT-3.5 em MultiArith, AddSub, SingleEq).
- **GPT-3.5 (Zero-shot CoT)**: 70.4% (baseline).

Resultados em Raciocínio de Senso Comum

- **LLaMA-13B + Parallel Adapter**: 81.5% de acurácia média.
- **ChatGPT (Zero-shot CoT)**: 77.0%.
- Adaptadores superaram todos os modelos base, inclusive PaLM e GPT-3.

Análise In-Distribution vs Out-of-Distribution

- Desempenho é superior quando o treinamento ocorre com dados ID, como em raciocínio de senso comum.
- Em tarefas OOD simples, como AddSub, modelos pequenos com PEFT superam GPT-3.5.
- Há ainda uma lacuna de desempenho em tarefas complexas como GSM8K.

Conclusão Através da LLM-Adapter, os autores demonstram que métodos PEFT, especialmente LoRA e adaptadores paralelos e em série, podem oferecer desempenho competitivo em tarefas específicas mesmo utilizando modelos significativamente menores. O estudo destaca a importância da configuração correta dos adaptadores e da seleção de dados de treinamento para maximizar os benefícios da abordagem PEFT.

Utilidade Esse artigo também realiza parte de nossa proposta, exibindo um comparativo entre diferentes tipos de adaptadores em um método de finetuning específico, o PEFT (parameter-efficient finetuning). Pode servir também, como referência para pensarmos em como validar o desempenho pós finetuning, e apresenta também ideias de outros dados importantes que podemos tirar da validação de performance (melhor posição dos adaptadores na rede, quais os melhores parâmetros de configuração, etc.).

Table 1: Variação de desempenho entre Fewshot (F) e Fewshot com explicações (FE)

Modelo	**Std(|F-FE|)**	**Média F**	**Média FE**
OPT	2.31	40.68	41.82
OPT-R	0.84	43.44	43.68
OPT-RE	0.78	44.49	44.86