What is RAFT?

RAFT (Retrieval-Augmented Fine-Tuning) is a training recipe designed to improve how Large Language Models (LLMs) answer questions in specialized domains. The paper analogizes this to an "open-book" exam, where the model has access to a fixed set of documents it can reference. The core innovation of RAFT is teaching the model to not only find the right answer within provided documents but also to actively ignore irrelevant "distractor" documents that do not help. This approach helps improve the model's reasoning abilities. RAFT's effectiveness was consistently demonstrated across several datasets, including PubMed, HotpotQA, and the Gorilla API Bench (Hugging Face, Torch Hub, and Tensorflow Hub).

The RAFT Training Method

The RAFT training process involves creating a unique dataset where each entry contains a question (Q), a set of documents, and a Chain-of-Thought style answer (A*). The documents provided are a mix of:

- **Golden Documents (D∗)**: One or more documents from which the correct answer can be derived.
- **Distractor Documents (Dk)**: Documents that are irrelevant to the question and are included to train the model to ignore unhelpful information.

A key part of the strategy is that for a certain percentage (

P) of the training data, the model receives both golden and distractor documents. For the remaining (

1−P) percent, the golden document is intentionally excluded, forcing the model to learn to identify when the context is insufficient and to sometimes rely on memorization.

For example, to create the high-quality training answers, the model is prompted to generate a logical reasoning chain that explicitly quotes the source text. This is done using

##begin_quote## and ##end_quote## tags, as shown below:

Question: The Oberoi family is part of a hotel company that has a head office in what city?

CoT Answer: ##Reason: The document ##begin_quote## The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group. ##end_quote## establishes that the Oberoi family is involved in the Oberoi group, and the document ##begin_quote## The Oberoi Group is a hotel company with its head office in Delhi. ##end_quote## establishes the head office of The Oberoi Group. Therefore, the Oberoi family is part of a hotel company whose head office is in Delhi. ##Answer: Delhi

---

Performance and Results

RAFT significantly outperforms standard baselines on domain-specific tasks. A RAFT-tuned LLaMA2-7B model achieved an accuracy of

---

The Role of Chain-of-Thought (CoT)

The use of Chain-of-Thought (CoT) is critical to RAFT's effectiveness. Providing just the answer can lead the model to quickly overfit, whereas incorporating a detailed reasoning chain improves understanding and accuracy. The performance boost is significant; adding CoT to RAFT training increased accuracy on HotpotQA from 25.62% to

35.28% and on Hugging Face from 59.07% to **74.00%**.

Optimizing Training and Robustness

The research uncovered two key insights for making the model more robust:

1. **Don't Always Include the Golden Document**: The paper challenges the assumption that the correct document should always be present during training. Experiments show that the optimal percentage of training data containing the golden document varies; for some datasets, performance peaked when the golden document was present only 40% or 60% of the time. This suggests that training with imperfect context is beneficial.
2. **Train with Multiple Distractors**: Training with only the golden document leads to poor performance when faced with distractors at test time. The model becomes more resilient to irrelevant text when it is trained with a mix of one golden document and several distractor documents. For the experiments, a setup of one golden document and four distractors was consistently used.