HUMANE Paper Review

# RAFT: Adapting Language Model to Domain Specific RAG

**Tianjun Zhang** *
Department of Computer Science
UC Berkeley
Berkeley, CA 94720, USA
{tianjunz}@berkeley.edu

**Shishir G. Patil, Naman Jain, Sheng Shen**
Department of Computer Science
UC Berkeley
Berkeley, CA 94720, USA
{shishirpatil,naman_jain,sheng.s}@berkeley.edu

**Matei Zaharia, Ion Stoica, Joseph E. Gonzalez**
Department of Computer Science
UC Berkeley
Berkeley, CA 94720, USA
{matei,istoica,jegonzal}@berkeley.edu

숭실대학교 문화콘텐츠학과, 석사과정생 이다현

COLM 2024

2025.04.18

# Background

- LLMs, trained on vast public data, have achieved significant advances in general knowledge reasoning tasks
- LLMs are being employed in specialized domains
- General knowledge reasoning < Document-grounded accuracy

How do we adapt pre-trained LLMs for
Retrieval Augmented Generation (RAG) in specialized domains?

# Background

**What is the best method to incorporate information into the pretrained model?**

| Aspect | RAG-based In-Context Learning | Supervised Fine-Tuning |
|---|---|---|
| Function | References documents for Q&A | Learns patterns and aligns with tasks/preferences |
| Strength | Utilizes docs at test time | Leverages the fixed domain setting and early access to test documents f or better learning |
| Limitation | Misses learning opportunities from fixed domains and early doc access | Either lacks RAG during inference or ignores retrieval flaws in training |

# Background

- LLMs for Open-Book Exam

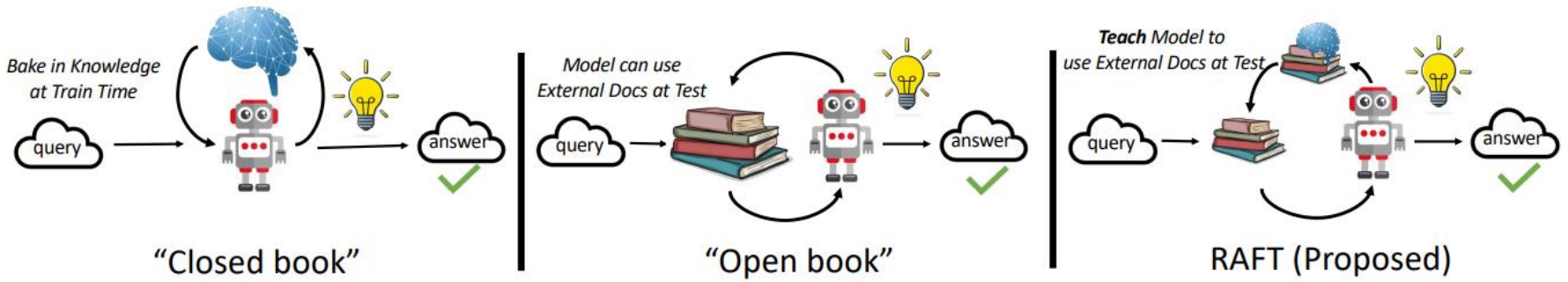


Figure 1: **How best to prepare for an Exam?**(a) Fine-tuning based approaches implement "studying" by either directly "memorizing" the input documents or answering practice QA without referencing the documents. (b) Alternatively, in-context retrieval methods fail to leverage the learning opportunity afforded by the fixed domain and are equivalent to taking an open-book exam without studying. In contrast, our approach (c) RAFT leverages fine-tuning with question-answer pairs while referencing the documents in a simulated imperfect retrieval setting — thereby effectively preparing for the open-book exam setting.

# Background

How do we adapt pre-trained LLMs for
Retrieval Augmented Generation (RAG) in specialized domains?

- Retrieval-Augmented Fine-Tuning (RAFT)
  - Fine-tunes LLMs for domain knowledge + RAG performance
  - Ensures robustness against distracting retrieved information
  - Trains models to understand connections between questions, domain documents, and correct answers

# RAFT

- Supervised Finetuning
  - {Train: Q → A}, {0-shot Inference: Q → A}, {RAG Inference: Q + D → A}
- RAFT: Retrieval Augmented Fine Tuning
  - Train Data is different
  - P % of data: Q + D* + D1 + D2 + . . . + Dk → A*
  - (1 − P) % of data: Q + D1 + D2 + . . . + Dk → A*

# RAFT

- Employ GPT-4-1106 to generate Chain-of-Thought prompts

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

context: [The Oberoi family is an Indian family that is famous for its
involvement in hotels, namely through The Oberoi Group]...[It is located in
city center of Jakarta, near Mega Kuningan, adjacent to the sister JW Marriott
Hotel. It is operated by The Ritz-Carlton Hotel Company. The complex has two
towers that comprises a hotel and the Airlangga Apartment respectively]...[The
Oberoi Group is a hotel company with its head office in Delhi.]

Instruction: Given the question, context and answer above, provide a logical
reasoning for that answer. Please use the format of: ##Reason: {reason}
##Answer: {answer}.

--------------------------------------------------------------------------------

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
```

# RAFT

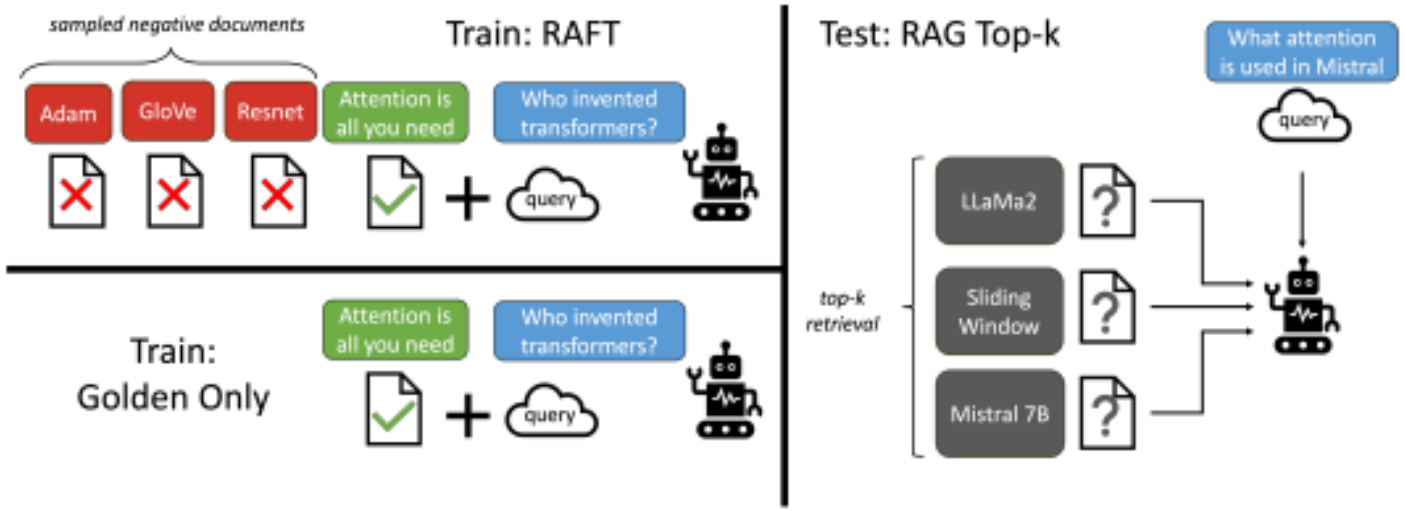


Figure 2: **Overview of our RAFT method.** The top-left figure depicts our approach of adapting LLMs to *reading* solution from a set of positive and distractor documents in contrast to standard RAG setup where models are trained based on the retriever outputs, which is a mixture of both memorization and reading. At test time, all methods follow the standard RAG setting, provided with a top-k retrieved documents in the context.

# Evaluation

- RAFT > domain-specific finetuned model, general-purpose model with RAG
  - Better at reading and extracting information from in-domain documents

Table 1: **RAFT improves RAG performance for all specialized domains**: Across PubMed, HotPot, HuggingFace, Torch Hub, and Tensorflow Hub, we see that Domain-specific Fine-tuning improves significantly of the performance of the base model, RAFT consistently outperforms the existing domain-specific finetuning method with or without RAG. This suggests the need to train the model with context. We compare our model with LLaMA finetuning receipes, and provide GPT-3.5 for reference.

| | PubMed | HotPot | HuggingFace | Torch Hub | TensorFlow |
|---|---|---|---|---|---|
| GPT-3.5 + RAG | 71.60 | **41.5** | 29.08 | 60.21 | 65.59 |
| LLaMA2-7B | 56.5 | 0.54 | 0.22 | 0 | 0 |
| LLaMA2-7B + RAG | 58.8 | 0.03 | 26.43 | 08.60 | 43.06 |
| DSF | 59.7 | 6.38 | 61.06 | 84.94 | 86.56 |
| DSF + RAG | 71.6 | 4.41 | 42.59 | 82.80 | 60.29 |
| RAFT (LLaMA2-7B) | **73.30** | 35.28 | **74.00** | **84.95** | **86.86** |

# Evaluation

- Effect of CoT
  - guides the model to the answer
  - also enriches the model's understanding
  - improve the overall accuracy and prevent overfitting to concise answers

Table 2: **Ablation on Chain-of-Thought**: The numbers of RAFT and RAFT without CoT. Results on various datasets show that adding CoT can significantly improve the performance of the finetuned model. With a gains of 9.66% and 14.93% in the Hotpot QA and HuggingFace datasets respectively.

| | PubMed | HotpotQA | HuggingFace | Torch Hub | TensorFlow |
|---|---|---|---|---|---|
| RAFT w.o CoT | 68.30 | 25.62 | 59.07 | **86.56** | 83.21 |
| RAFT | **73.30** | **35.28** | **74.00** | 84.95 | **86.86** |

# Evaluation

- Employ GPT-4-1106 to generate Chain-of-Thought prompts

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

context: [The Oberoi family is an Indian family that is famous for its
involvement in hotels, namely through The Oberoi Group]...[It is located in
city center of Jakarta, near Mega Kuningan, adjacent to the sister JW Marriott
Hotel. It is operated by The Ritz-Carlton Hotel Company. The complex has two
towers that comprises a hotel and the Airlangga Apartment respectively]...[The
Oberoi Group is a hotel company with its head office in Delhi.]

Instruction: Given the question, context and answer above, provide a logical
reasoning for that answer. Please use the format of: ##Reason: {reason}
##Answer: {answer}.

--------------------------------------------------------------------------------

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
```

# Evaluation

- Qualitative Analysis

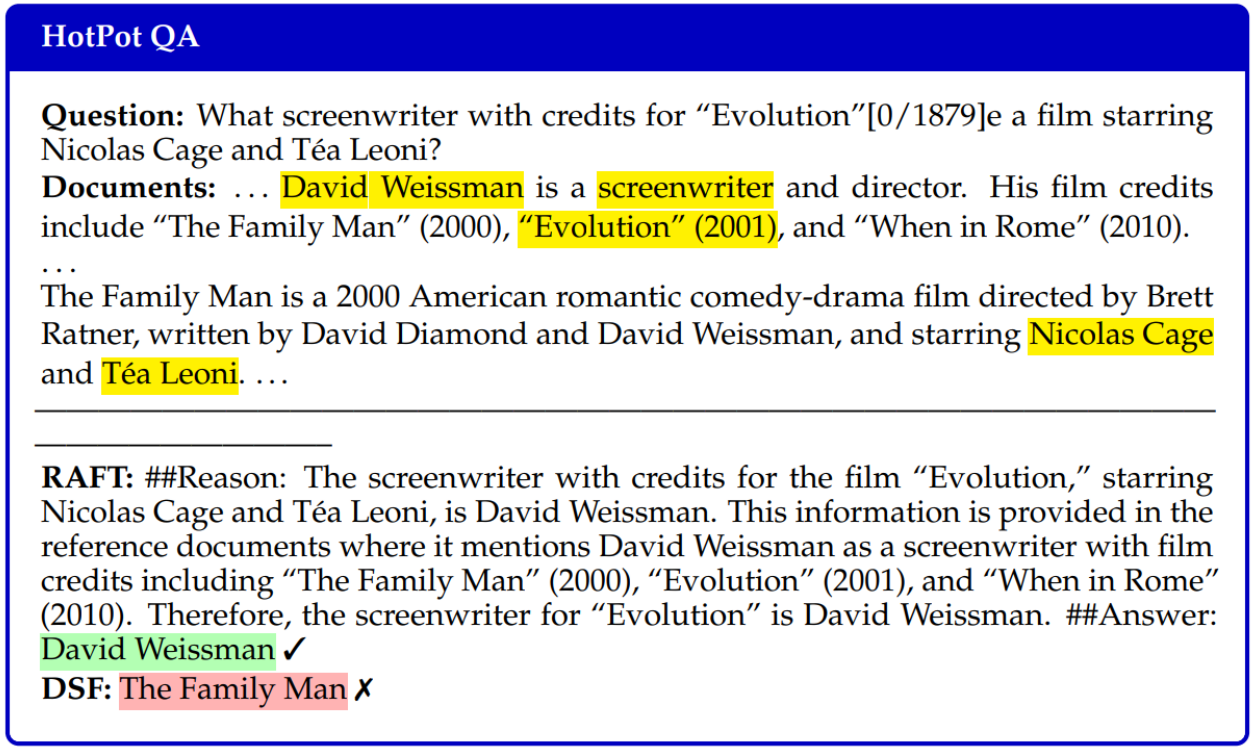


Figure 4: **Comparison of RAFT and DSF**: On the HotPot QA dataset, we can see that DSF model extracts the wrong information from the context when the question is asking who is the screen writer and it answers a film name. RAFT manages to get the accurate results .

# Evaluation

- Should we train the LLM always with the golden context for RAG?

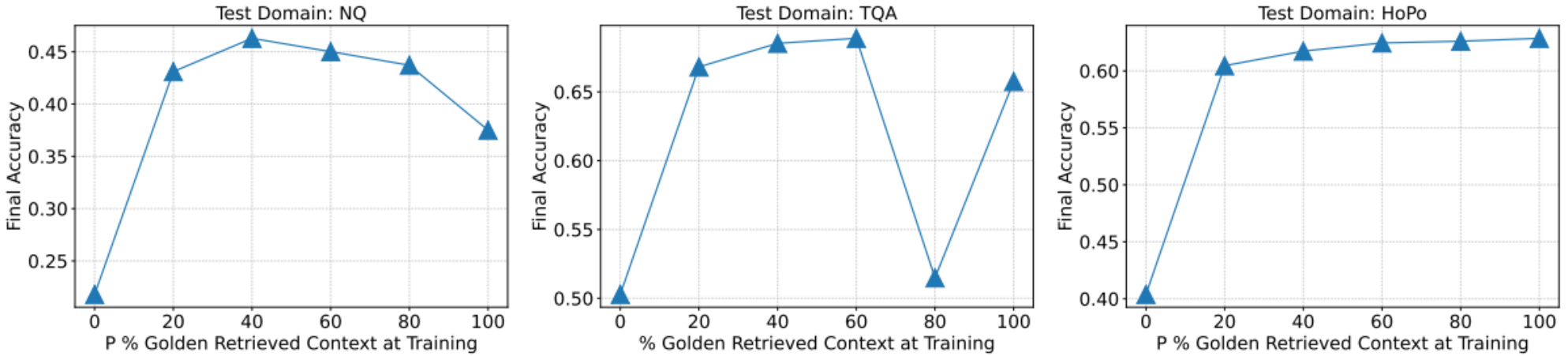


Figure 5: **How many golden documents to involve?** We study the hyperparameter P% where it indicates how much portion of training data is with golden document. Results on NQ, TQA and HotpotQA suggest that mixing some amount of data that the golden document is not put in the context is helpful for in-domain RAG.

# RAFT Generalizes to Top-K RAG

- How does the number of distractor documents in RAFT affect the model's performance when augmented with top-k RAG results during evaluation?

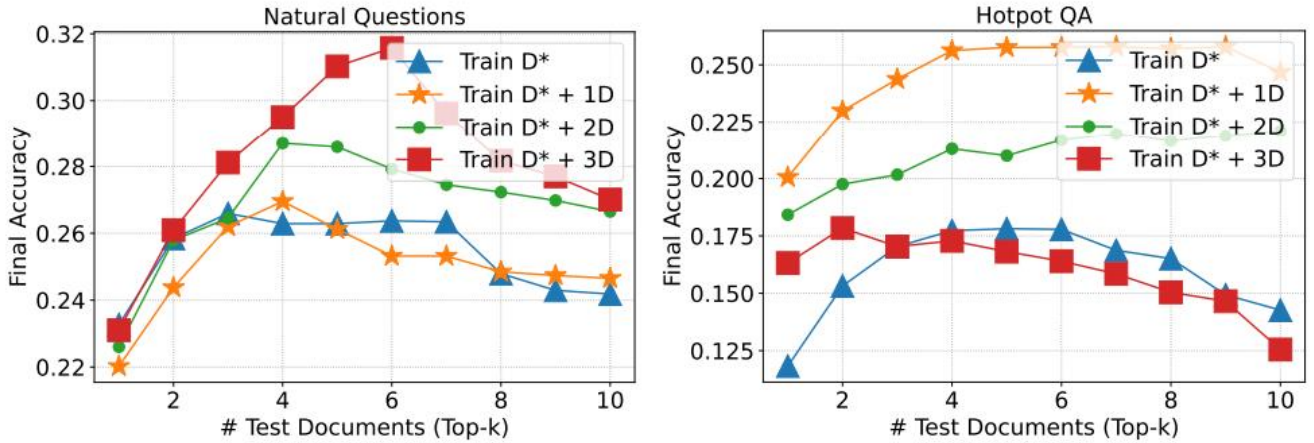


Figure 6: **Test-Time Documents Varying**: To analyze how robust RAFT is to varying number of test-time documents, we study three domains – NQ, Trivia QA and HotPot QA. In NQ, we find that training with 4 documents leads to optimal performance, and this changes to 3 and 2 for for Trivia QA and HotPot QA respectively. However, we see that training with only *golden* documents leads to poor performance.

# Conclusion

- RAFT (Retrieval-Augmented Fine-Tuning) is a training strategy for enhancing LLM performance in domain-specific open-book settings.
- It improves a model's ability to:
  - Identify relevant information while ignoring distractors
  - Extract answers from retrieved documents effectively
- Key design choices include:
  - Training with distractor documents
  - Partially removing golden contexts to improve generalization
  - Incorporating CoT reasoning grounded in the retrieved content
- Experiments on PubMed, HotpotQA, and Gorilla APIBench show that RAFT outperforms both instruction-tuned and domain-finetuned baselines with or without RAG.

# My Review

- 장점
  - 문제 정의가 뚜렷하고 비유를 사용해서 이해하기 쉬움
  - CoT 유무, 골든 문서 비율(P%), 디스트랙터 수(k) 등 핵심 요소를 ablation study를 통해 입증
- 단점
  - 실험은 대부분 Llama-2-7B 단일 크기에 집중 – 큰 모델 실험 부재
  - 디스트랙터 “개수”만 조절, 유사도·주제 다양성이 성능에 미치는 영향은 탐색하지 않음
  - GPT-4로 생성한 CoT를 학습에 사용 → 모델 간 정보 누수(teacher-student leakage) 및 비용 문제 논의 부족