Multi-Hop Question Answering

CSE 587 Spring 2025
Final Project
by Lakshay Piplani (PSU ID: ljp5718) and Durva Dev (PSU ID: dbd5616)

Github Repository Link

Introduction

For our final project, we are tackling the problem of multi-hop reasoning.

Multi-Hop Reasoning

Multi-hop reasoning is a systematic approach to answering questions that are complex, that is, implicitly contain (and can be broken down into) sub-questions. Due to this composite nature of multi-hop questions, answering them requires the agent to execute multiple steps of inference, and at each step utilize information from different sources. Therefore, salient features of multi-hop reasoning are:

- **Complex Question:** A complex question that cannot be answered using a single piece of fact
- **Intermediate Inferences:** Multiple inferences are needed to be chained together
- **Multiple Facts:** Each step requires some new pieces of information

Formally, we can represent multi-hop reasoning as:

$$\text{Answer} = f_n(\cdots f_2(f_1(\text{Query}, \text{Set of Facts}_1), \text{Set of Facts}_2), \ldots, \text{Set of Facts}_n)$$

where f_i represents reasoning operations at hop i.

Examples

Example 1: Simple Two-Hop Question

Question: "Where was the inventor of the telephone born?"

1. *Hop 1*: Identify the inventor of the telephone
   Inventor = Alexander Graham Bell (from Fact 1)
2. *Hop 2*: Find birth place of Alexander Graham Bell
   BirthPlace = Edinburgh, Scotland (from Fact 2)

Example 2: Temporal Reasoning

Question: "Was the CEO of Company X born before the company was founded?"

1. *Hop 1*: Find founding year of Company X (1995)
2. *Hop 2*: Find birth year of current CEO (1980)
3. *Hop 3*: Compare years ($1980 < 1995 \Rightarrow$ Yes)

Machine Learning for Multi-Hop Reasoning

Multi-hop reasoning serves as an integral benchmark for testing high-order natural language reasoning abilities of machine learning models.

Classical Deep Learning Approaches

Below are some example from the literature of architectures that have been tested for multi-hop reasoning:

- Non transformer-based: Schlichtkrull et al. (2018) posited an architecture that uses Gated Recurrent Units that are entity-aware to perform iterative reasoning over text passages. Sabour, Frosst, and Hinton (2017) capture relationships between entities across documents using convolution layers. They also use a routing-by-agreement mechanism which they have shown to enable multi-hop inference

- Transformer-based: Fang et al. (2020) propose Multi-Hop Graph Relation Network that uses Graph Convolutional Network with edge-type-specific message passing.

- Hybrid architecture: Qiu et al. (2019) have shown that we can use Recurrent Neural Networks to capture the context and Graph Convolutional Network to capture relation between entities.

Retrieval-Augmented Methods

Retrieval Augmented Generation (RAG) uses two components for multi-hop reasoning:

- Iterative Retriever: At each hop, the retriever uses the question and the accumulated relevant context to fetch further context

- Language Model: The question and the final retrieved context are used by the LM to generate the answer and optionally the reasoning chain-of-thought. Alternatively, the retriever and LM can perform alternating turns of retrieve-then-reason-then-retrieve. This method has shown to be superior to the regular retrieve once-reason once architecture.

Language Model Fine-Tuning Methods

Latest approaches fine-tune an LM to answer complex questions requiring multi-hop reasoning. Key steps include:

- Prompt Engineering: Use an instruction template (e.g., Llama 3's chat format) to explicitly prompt the question, answer, supporting context, along with correct step-by-step reasoning, and divergent reasoning negative examples to improve robustness.

- Supervised Fine-Tuning (SFT): Train the LM via causal language modeling, where the input is the prompt and labels are the gold reasoning chains and answers.

- Evaluation: of final answer and gold chain using Exact Match, F1 scores or human review.

Common LLM training approaches include:

- **Supervised**: We give the model a prompt including instructions, the question, the facts, and the label as the final answer and the supporting chain of thought.
- **Weakly Supervised**: It is similar to supervised, with the exception that we give only the final answer as the label. Therefore, the ”golden chain” of thought is implicit and the model should learn to create such a chain of thought implicitly.
- **Reinforcement Learning**: We provide positive rewards to the language model when it generates correct reasoning paths.

Challenges

Key challenges in ML approaches:

- **Explainability**: Deep Learning architectures are a black-box, that is, it is not easy to explain why a particular output was generated by the network.
- **Compositionality**: Generalizing to unseen data points can be a challenge.
- **Efficiency**: Computational complexity of training or fine-tuning language models is very high and infeasible to perform on commodity compute resources.

Sate-of-the-Art on Multi-Hop Question Answering

Table 1: Top 3 Models on HotpotQA Leaderboard (Full-Wiki Setting)

Rank	**Model**	**EM/F1**
1	Zhang et al. (2024)	72.69 / 85.04
2	PipNet (TenCent)	72.26 / 84.86
3	Yin et al. (2023)	72.07 / 84.34

Dataset

We use the HotpotQA dataset Yang et al. (2018) for our use-case of multi-hop question answering (QA), which in our opinion is well-suited for evaluating models that need to reason using multiple threads of evidence. Please note that we have used the version of dataset that contains the supporting facts and context paragraphs.

The HotpotQA dataset features natural, multi-hop questions, with strong supervision for supporting facts to enable more explainable QA systems. It is collected by a team of NLP researchers at Carnegie Mellon University, Stanford University, & Université de Montréal.

Unlike other benchmarks, the dataset was crowdsourced using Wikipedia pages as the ”context” and crowd workers were shown multiple context documents and asked explicitly to come up with questions requiring reasoning about all of the documents, ensuring that the dataset covers multi-hop questions that are more natural, and are not designed with any pre-existing knowledge base schema in mind.

HotpotQA is divided into two main settings: distractor and full wiki. In this work, we focus on the distractor setting, which is easier to fit with our objective of enhancing multi-hop QA using knowledge graphs.

Since we had resource constraints that limited the scope of our experiments, we only used the train and dev splits out of all the splits (shown in Table 1) used in the actual model. We sampled and mixed both the datasets (train, dev) to get a total of 22000 examples.

Name	Desc.	Usage	# Examples
train-easy	single-hop	training	18,089
train-medium	multi-hop	training	56,814
train-hard	hard multi-hop	training	15,661
dev	hard multi-hop	dev	7,405
test-distractor	hard multi-hop	test	7,405
test-fullwiki	hard multi-hop	test	7,405
Total			112,779

Table 1: Data split. The splits *train-easy*, *train-medium*, and *train-hard* are combined for training. The distractor and full wiki settings use different test sets so that the gold paragraphs in the full wiki test set remain unknown to any models.

Below is an example from the dataset.


```
{
    "_id": "5a7286865542991f9a20c4cb",
    "answer": "Kind Hearts and Coronets",
    "question": "Which 1949 British black comedy film does the Skindles hotel appear on? ",
    "supporting_facts": [
        [
            "Skindles",
            4
        ],
```

```
    [
        "Kind Hearts and Coronets",
        0
    ]
],
"context": [
    [
        "Skindles",
        [
            "Skindles was a hotel in Maidenhead, England, on the Buckinghamshire bank of the
                ↪ River Thames by Maidenhead Bridge.",
            " Formerly the Orkney Arms, built in 1743, it was turned from a coaching inn into
                ↪  a fashionable hotel by William Skindle in 1833.",
            " In the 20th century, it became notorious as a place for adulterous assignations
                ↪ .",
            " Its guests included Winston Churchill and Princess Margaret, and musicians who
                ↪ performed there included The Rolling Stones and The Strawbs.",
            " The hotel appears in the film Kind Hearts and Coronets.",
            " Skindles is mentioned in the play Journey's End by R. C. Sherriff: 'We danced a
                ↪  bit at Skindles, and drank a lot of port and muck'."
        ]
    ],
    [
        "Dead Clever",
        [
            "Dead Clever is a British black comedy film, first screened on ITV on New Year's
                ↪ Day, 2007.",
            " Written by Sally Wainwright, it stars Suranne Jones, Helen Baxendale and Dean
                ↪ Lennox Kelly.",
            " Although officially titled \"Dead Clever\" it was subtitled \"The Life and
                ↪ Crimes of Julie Bottomley\".",
            " The music was written by BAFTA nominated TV & film music composer Sheridan
                ↪ Tongue."
        ]
    ],
    [
        "Burn Burn Burn",
        [
            "Burn Burn Burn is a 2015 British black comedy film, the directorial debut of
                ↪ Chanya Button.",
            " The film is a coming-of-age tale, inspired by the Jack Kerouac novel \"On the
                ↪ Road\" published in 1957.",
            " The fictional plot follows the story of two girls, Seph (Laura Carmichael) and
                ↪ Alex (Chloe Pirrie), taking a road trip to follow the instructions of their
                ↪  close friend Dan, who has died and given them instructions where to
                ↪ scatter his ashes.",
            " The ashes (stored in tupperware in the glove compartment) keep diminishing in
                ↪ quantity as the trip progresses.",
            " The film had its World premiere at the BFI London Film Festival 2015."
        ]
    ],
    [
        "Bigga than Ben",
        [
            "Bigga than Ben is a 2008 British black comedy film written and directed by Suzie
                ↪  Halewood.",
            " The film is based on the 1999 Russian novel of the same name."
```

```
    ]
  ],
  [
    "Burke &amp; Hare (2010 film)",
    [
      "Burke & Hare is a 2010 British black comedy film, loosely based on the Burke and Hare murders.",
      " Directed by John Landis, the film stars Simon Pegg and Andy Serkis as William Burke and William Hare respectively.",
      " It was Landis's first feature film release in 12 years, the last being 1998's \"Susan's Plan\".",
      " The film was released in the United Kingdom on 29 October 2010."
    ]
  ],
  [
    "A Long Way Down (film)",
    [
      "A Long Way Down is a 2014 British black comedy film directed by Pascal Chaumeil, loosely based on author Nick Hornby's 2005 novel, \"A Long Way Down\".",
      " It stars Pierce Brosnan, Toni Collette, Imogen Poots, and Aaron Paul as four strangers who happen to meet on the roof of a London building on New Year's Eve, each with the intent of committing suicide.",
      " Their plans for death in solitude are ruined when they meet as they decide to come down from the roof alive \u2014 however temporary that may be."
    ]
  ],
  [
    "Kind Hearts and Coronets",
    [
      "Kind Hearts and Coronets is a 1949 British black comedy film.",
      " It features Dennis Price, Joan Greenwood, Valerie Hobson and Alec Guinness; Guinness plays nine characters.",
      " The plot is loosely based on the novel \"Israel Rank: The Autobiography of a Criminal\" (1907) by Roy Horniman.",
      " It concerns Louis D'Ascoyne Mazzini, the son of a woman disowned by her aristocratic family for marrying out of her social class.",
      " After her death Louis decides to take revenge on the family, and to take the dukedom, by murdering the eight people ahead of him in succession to the title."
    ]
  ],
  [
    "Entertaining Mr Sloane (film)",
    [
      "Entertaining Mr Sloane is a 1970 British black comedy film directed by Douglas Hickox.",
      " The screenplay by Clive Exton is based on the 1964 play of the same title by Joe Orton.",
      " This was the second adaptation of the play, the first having been developed for British television and telecast by ITV on 15 July 1968."
    ]
  ],
  [
    "The Ruling Class (film)",
    [
      "The Ruling Class is a 1972 British black comedy film.",
      " It is an adaptation of Peter Barnes' satirical stage play \"The Ruling Class\"
```

```
                ↪ which tells the story of a paranoid schizophrenic British nobleman (played
                ↪ by Peter O'Toole) who inherits a peerage.",
            " The film co-stars Alastair Sim, William Mervyn, Coral Browne, Harry Andrews,
                ↪ Carolyn Seymour, James Villiers and Arthur Lowe.",
            " It was produced by Jules Buck and directed by Peter Medak."
        ]
    ],
    [
        "Just Jim (2015 film)",
        [
            "Just Jim is a 2015 British black comedy film written and directed by Craig
                ↪ Roberts in his directorial debut.",
            " The film stars Roberts as a lonely Welsh teenager who is given the chance to
                ↪ increase his popularity when a cool American (Emile Hirsch) moves in next
                ↪ door."
        ]
    ]
  ],
  "type": "bridge",
  "level": "hard"
}
```


Each example includes a question along with,

1. Two **Gold** (supporting paragraphs) – these are the paragraphs and the line number that contain the context required to answer the question i.e., the truth
2. Eight **distractor** paragraphs – these are paragraphs selected to be similar, but irrelevant to answering the question, increasing the difficulty.

LLM Selection and Training

Base Language Model

We have fine-tuned the open–weights **Llama-3.2-3B**, a 3-billion-parameter decoder-only transformer distilled from the official Llama-3 family proposed by Grattafiori et al. (2024)

Compared to larger Llama-3 models, this 3B version can fit into low-grade GPUs, making training and inference more accessible for us.

LoRA Adapters

We have used rank-$r = 8$ LoRA adapters (`lora_alpha=16`, dropout 0.05) to every `q_proj` and `v_proj` in the attention blocks. Furthermore, the base weights are quantized to 4-bit NF4 via `bitsandbytes`; only ≈0.3 M trainable parameters are updated.

Supervised Fine-Tuning Pipeline

Figure 1 illustrates the complete data flow. Steps are described below.

1. **Train-Eval-Test split**: After we retrieve the raw JSON HotPotQA data, we take 70% of the records for training and reserve 30% for final evaluation. We serialize the raw JSON data into Huggingace's Arrow Dataset for the interim. The training data is further broken into actual training and evaluation data, which is 20% of the training set and is used during training evaluation.

2. **Per-data point graphization**: Each data point is converted into a graph $G = (V, E)$ where nodes v_i represent context paragraphs and edges e_i encode inter–paragraph links (golden chain edges, and common tokens edges). Therefore, there is a graph for each data point, and we use it to generate golden and divergent examples while creating the prompt.

3. **Prompt construction**: For every example, we generate *system*, *user* and *assistant* blocks using the official Llama-3 chat template:

<begin_of_text>
<start_header_id>system<end_header_id>
You are a careful multi-hop reasoner. Think step-by-step.Your task is to:
1. First identify the most relevant paragraph(s) to the question
2. Extract key information from those paragraphs
3. Combine the information to form a final answer
4. Present your reasoning clearly with [STEP X] tags before each step
EXAMPLE QUESTION: Which company is older, EOG Resources or General Mills?
PARAGRAPHS:
0: EOG Resources —— Founded in 1999...
1: General Mills —— Founded in 1866...
[STEP 1] Check founding dates in paragraph 0 and 1
[STEP 2] EOG founded in 1999 (paragraph 0)
[STEP 3] General Mills founded in 1866 (paragraph 1)
FINAL: General Mills is older <eot_id>
<start_header_id>user<end_header_id>

Here is the QUESTION you are supposed to answer: QUESTION: ...
PARAGRAPHS:
0: Title_0 || Sentences_0
1: Title_1 || Sentences_1
...
These are the relevant paragraph numbers:
[STEP 1] p_0
...
FINAL: *answer*
The other candidate(s) are not relevant to the question. These are wrong paths.
CANDIDATE 2: (divergent)
...
CANDIDATE 3: (divergent)
...
<eot_id>
<start_header_id>assistant<end_header_id>
[STEP 1] p_0
...
FINAL: *answer*
assistant<end_header_id>

We use each data-point's graph to generate three candidate chains: Candidate 1 is the *golden reasoning* chain, and two additional *distractor* chains (random non-gold hops) are provided as in-context negatives.

4. **Label generation**: Only the *assistant* section containing golden reason is used as target text. HuggingFace's tokeniser is called with `text_target` set to this slice, so all preceding tokens are masked to -100 and ignored by the cross-entropy loss.

5. **Left padding**: We set

   `tokenizer.pad_token = tokenizer.eos_token`, `tokenizer.padding_side = "left"`.

   The `DataCollatorForSeq2Seq` pads inputs and labels on the *left*, following the best practice for causal LM training.

6. **Parameter-Efficient Fine-Tuning**: As discussed earlier, we apply LoRA adapters to the layers of the base model, which itself remains unchanged. The LoRA weights themselves are 4-bit quantized to accelerate training at the cost of decreased performance.

7. **Optimization**. Training parameters are:

   - Epochs = 2
   - batch = 1, gradient accumulation = 12. We keep batch size as 1 because larger batch sizes were not fitting onto our hardware
   - AdamW optimizer
   - Mixed-precision FP16
   - The evaluation split (20%) monitors perplexity at each epoch

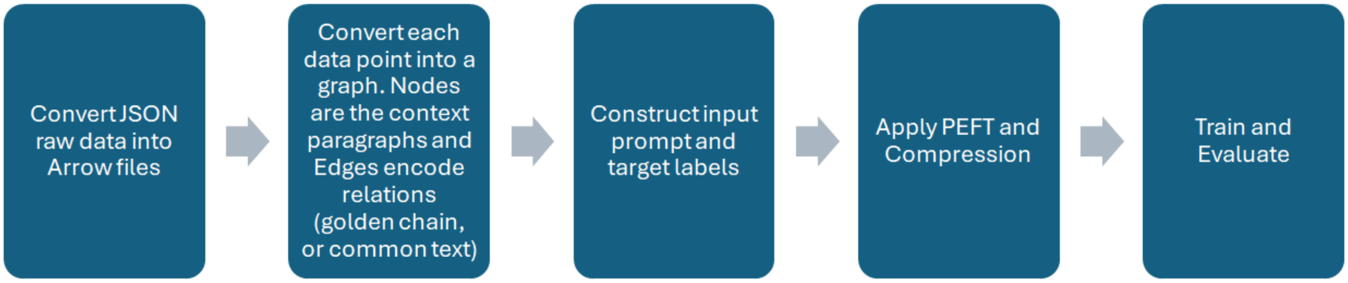

Figure 1: Training Pipeline

Experiments and Evaluation

Experimental Setup

Dataset splits: As discussed earlier, HotpotQA JSON data is randomly partitioned 70:30 into **train** and **test** folds, preserving the label distribution. During fine-tuning we further carve the training slice into 80:20 *train / eval* using `Dataset.train_test_split` (seed = 42). The final outcomes of the experiments below are reported on the held-out *test* split.

Inference controller: We use the below inference methods:

1. **Greedy (1-step).** The model receives the prompt (question + context paragraphs) and generates a single `FINAL:` answer without explicit reasoning steps.
2. **Adaptive Graph-of-Thought (AGoT).** We also use AGoT inference strategy, as proposed by Pandey et al. (2025) Our controller expands a reasoning DAG, branching only when the Shannon entropy of the next-hop logits exceeds $\tau = 1.3$ bits. We cap the search at four hops to prevent loops.

All inference runs are performed with the 4-bit base weights and rank-8 LoRA adapters; decoding uses `do_sample=False`.

Inference Prompt

We gave the following prompt to the model for inference.

```
    "<|begin_of_text|>",
"<|start_header_id|>system<|end_header_id|>",
"You are a careful multi-hop reasoner. Think step-by-step.",
"Your task is to:"
"1. First identify the most relevant paragraph(s) to the question",
"2. Extract key information from those paragraphs",
"3. Combine the information to form a final answer",
"4. Present your reasoning clearly with [STEP X] tags before each step",
"5. End with FINAL: ",
"""
```

```
EXAMPLE QUESTION: Which company is older, EOG Resources or General Mills?
PARAGRAPHS:
0: EOG Resources || Founded in 1999...
1: General Mills || Founded in 1866...

[STEP 1] Check founding dates in paragraph 0 and 1
[STEP 2] EOG founded in 1999 (paragraph 0)
[STEP 3] General Mills founded in 1866 (paragraph 1)
FINAL: General Mills is older
"""
"<|eot_id|>"
    prompt.extend([
"<|start_header_id|>user<|end_header_id|>",
f"Here is the QUESTION you are supposed to answer: {question}",
"PARAGRAPHS:"
<|eot_id|>"<|start_header_id|>assistant<|end_header_id|>\n")
"FINAL"
```

Experiments

We ran multiple experiments to study the performance of the HotpotQA dataset on multiple models.Figures show the loss, gradient norm and learning rate.

1. Llama-3.2-3B quantized (4-bit) – finetuned vs pre-trained
2. Llama-3.2-1B quantized (4-bit) – finetuned vs pre-trained

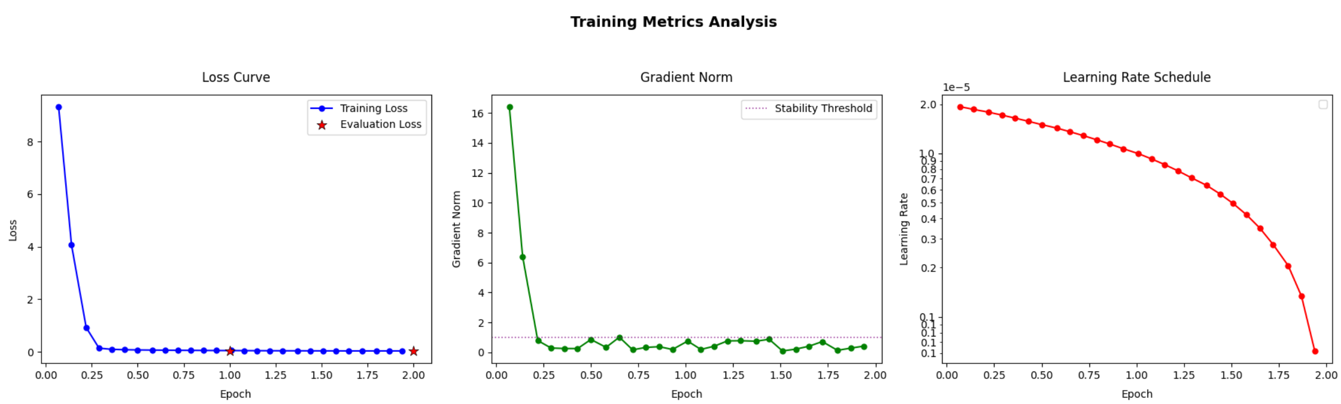

Figure 2: Plots for LLama-3.2-3B 4-bit quantized

We tried to fine-tune the model with the 8-bit quantized option but faced significant difficulties in training due to memory constraints.

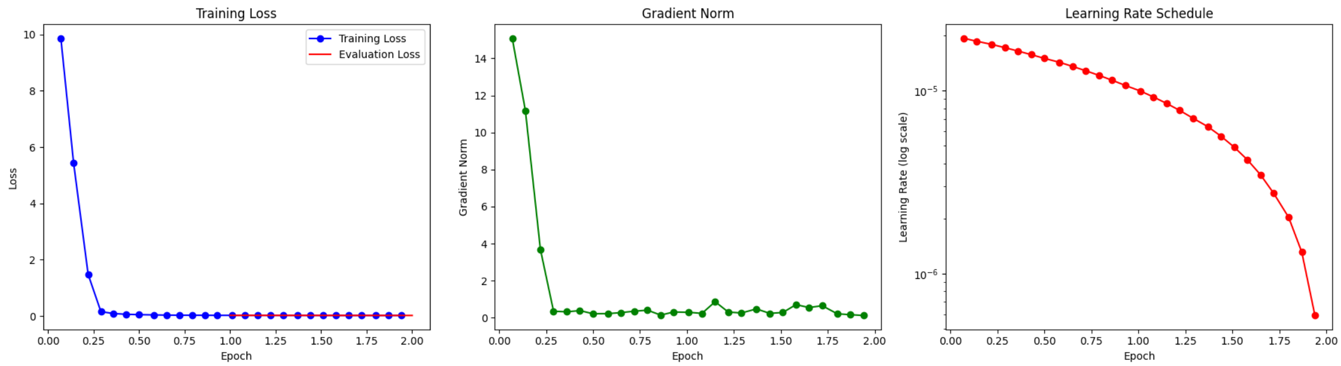

Figure 3: Plots for LLama-3.2-1B 4-bit quantized

Llama-3.2-3B

The figure 4 shows that fine-tuning the model on our dataset drastically increases the number of exact matches and also improves the F1 score.

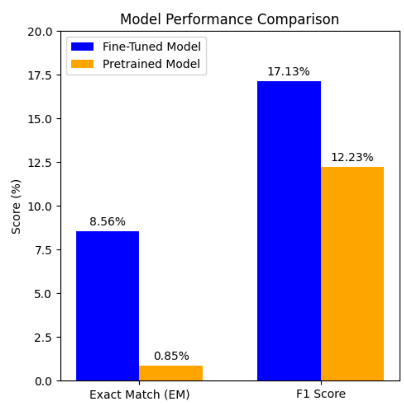

Figure 4: Fine-tuning improves Matches

The figures below show the prompts give by the fine-tuned model, the "RAW GENERATED TEXT" is the answer given by the model. For manual verification, we checked the first few prompts.For the first prompt in the figure below, the model gives the correct output "SOHA ALI KHAN".

```
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
Here is the QUESTION you are supposed to answer: Which lead actor/actress in War Chhod Na Yaar has also acted in Bengali and English-language films?
PARAGRAPHS:
  0: David Dencik || Karl David Sebastian Dencik (born 31 October 1974) is a Swedish-Danish actor.  He has acted in both Swedish and Danish films, and has also had major rol
in English-language films including "Tinker, Tailor, Soldier, Spy" and " The Girl with the Dragon Tattoo".
  1: Gita Dey || Gita Dey (5 August 1931 – 17 January 2011) was a renowned actress in Bengali Tollywood films, theatre and Bengali folk theater.  She became a stage artist a
he age of 6 years.  She came to the film industry in 1943.  Her first film release was "Ahuti" (1941) as a child actress.  She acted in over two hundred Bengali language fil
and over two thousand stage shows.  She acted in the movie "Teen Kanya" under the direction of Satyajit Roy and Rittik Ghatak's "Meghe Dhaka Tara, Subarnarekha, Komal Gandha
Kato Ajanare".  She also acted in Hindi movies such as "Parineeta" (2005) with Vidya Balan and Sanjay Dutt and other movies.  She was associated with All India Radio for a l
 time doing Shruti Natok.  She received the Presidential Award for Lifetime Achievement from Dr. A.P.J. Abdul Kalam and many other awards during her lifetime.
  2: Debashree Roy filmography || Debashree Roy is an Indian actress who has performed in over hundred films.  She is a National Award winner actress and known for her work
Bengali cinema.  She has been a highly successful leading actress in Bengali cinema throughout eighties and nineties as well as a critically acclaimed actress.  She started
 career as a child artist in a Bengali film named "Pagal Thakur" (1966) directed by Hiranmoy Sen. She was launched as an adult actress in Arabinda Mukhopadhyay's Bengali fli
"Nadi Theke Sagare" (1978).  In 1980, Tarun Majumdar cast her in his romantic flick "Dadar Kirti" which literally made her a star.  She acted in Aparna Sen's directorial deb
36 Chowringhee Lane (1981).  She appeared in Hindi films like Kanak Mishra's "Jiyo To Aise Jiyo" (1981), Desh Gautam's "Bura Aadmi" (1982), Kovelamudi Raghavendra Rao's "Jus
e Chaudhury" (1983), Mukul Dutt's "Phulwari" (1984), Akash Jain's "Seepeeyan" (1984), Vijay Singh's "Kabhi Ajnabi The" (1985), Bhabendra Nath Saikia's "Kaal Sandhya" (1997).
er "Kabhi Ajnabi The" was assumed to be a huge hit at box office but it turned out to be a major flop at box office causing a doom to her Bollywood career.  In 1985, she act
in Tarun Majumdar's romantic flick Bhalobasa Bhalobasa which was a major success at box office.  This film ensured Roy's pairing with Tapas Paul as the leading on-screen pai
g of nineteen eighties.  Her other major hits with Paul are "Lalmahal" (1986), "Uttar Lipi" (1986), "Arpan" (1987), "Shankhachur" (1988), "Surer Sathi" (1988), "Surer Akashe
1988), "Nayanmani" (1989), "Chokher Aloy" (1989), "Shubha Kamana" (1991), "Mayabini" (1992), "Phire Paoa" (1993), "Tobu Mone Rekho" (1994), "Putrabadhu" (1998).
  3: War Chhod Na Yaar || War Chhod Na Yaar (English translation: "Quit the War, Dude") is a 2013 Hindi Indian war comedy film directed and written by Faraz Haider and dialo
s are written by Deepak Kingrani.  The film was announced on 30 March 2013 with a launch party at Novotel Hotel in the presence of the lead actors, Sharman Joshi, Soha Ali K
, Javed Jaffery, Sanjai Mishra, Dalip Tahil and Mukul Dev who play important characters in the film.
  4: Anil Chatterjee || Anil Chatterjee (Chattopadhyay) Bengali: অনিল চ্যাটার্জী (চট্টোপাধ্যায়) (25 October 1929 – 17 March 1996) was a Bengali actor in the Bengali cinema durin
he early fifties through the mid-nineties and is mostly remembered as a character actor.  He acted or performed in about 150 movies, including a few in Hindi.  He played dif
ent shades, though mostly as a character actor, as well as in leading roles and at times as an antagonist, despite the limited opportunities he received.  Irrespective of th
oles, he left an indelible impression on the viewers and the critics.  He also acted in the leading role in a tele-serial named "Naqab" in the national network of Doordarsha
 He is one of the very few selected actors who worked with Satyajit Ray, Ritwik Ghatak, Tapan Sinha and Mrinal Sen; with the first three directors, he performed in more than
number of occasion.  His performance in the title role of the film "Deshbandhu Chittaranjan" [a biopic on the noted and revered freedom-fighter and barrister Chittaranjan Da
1870–1925) who was conferred the title of "Deshbandhu (Friend of the Country)"] won the appreciation of Basanti Devi, the wife of the "Deshbandhu" (she is said to have perso
ly blessed the actor).
  5: Soha Ali Khan || Soha Ali Khan Khemu (born Soha Ali Khan , 4 October 1978) is an Indian film actress, known primarily for her work in the Hindi film industry, although
 has also acted in Bengali and English-language films.
  6: Siddhartha Jadhav || Siddhartha Ramchandra Jadhav is an Indian actor and comedian.  He is popular on television and has acted in Marathi and Hindi movies making him one
 the most talented and successful actors in both Marathis films and television.  He has acted in several Bollywood movies like "Golmaal" and "Golmaal Returns" but he asserts
at his first love is Marathi film, TV and stage.Jadhav also acted in a bengali movie named Ami Subhah Bolchi having Mithun chakraborty as the lead.  He is currently paired w
 Bharti Singh on TV show Comedy Circus Ke Ajoobe and has been working with her for 3 seasons consecutive.
  7: Maharshi Raghava || Maharshi Raghava (born "Raghava Gogineni") is an Indian actor from Andhra Pradesh.  He acted in more than 170 films in various roles.  He shot into
e as a lead actor with his first film "Maharshi".  The same film earned him the title "Maharshi".  He also acted in TV serials.
  8: Utpal Dutt || Utpal Dutta (Bengali: "উৎপল দত্ত" "Utpôl Dôtto",    ) (29 March 1929 – 19 August 1993) was an Indian actor, director, and writer-playwright.  He was primar
 an actor in Bengali theatre, where he became a pioneering figure in Modern Indian theatre, when he founded the "Little Theater Group" in 1949.  This group enacted many Engl
, Shakespearean and Brecht plays, in a period now known as the "Epic theater" period, before it immersed itself completely in highly political and radical theatre.  His play
ecame an apt vehicle for the expression of his Marxist ideologies, visible in socio-political plays such as "Kallol" (1965), "Manusher Adhikar", "Louha Manob" (1964), "Tiner
loar" and "Maha-Bidroha".  He also acted in over 100 Bengali and Hindi films in a career spanning 40 years, and remains most known for his roles in films such as Mrinal Sen'
Bhuvan Shome" (1969), Satyajit Ray's "Agantuk" (1991), Gautam Ghose's "Padma Nadir Majhi" (1993) and Hrishikesh Mukherjee's breezy Hindi comedies such as "Gol Maal" (1979) a
"Rang Birangi" (1983).
  9: Farida Jalal || Farida Jalal (born 1949) is an Indian actress, who has acted in over two hundred Bollywood films.  Known primarily for her work in Hindi cinema, she has
so worked in Telugu, Tamil and English-language films.
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

FINAL...

Tokenization complete.
Generated token IDs: tensor([    25,   2100,   4317,  14925,  25273,    198, 128001],
       device='cuda:0')
New tokens count: 7

=== RAW GENERATED TEXT ===
': Soha Ali Khan\n<|end_of_text|>'
```

However, for this prompt, the model is unable to give the correct answer, the correct answer according to the dataset being 1993. Instead of giving the correct answer, the year the Dealey Plaza was given a "Historic Landmark" status, it ends up answering the year in which President John F. Kennedy's motorcade passes the plaza which is 1963.

```
=== PROMPT BUILT ===
Prompt preview:
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
You are a careful multi-hop reasoner. Think step-by-step.
Your task is to:1. First identify the most relevant paragraph(s) to the question
2. Extract key information from those paragraphs
3. Combine the information to form a final answer
4. Present your reasoning clearly with [STEP X] tags before each step
5. End with FINAL:

        EXAMPLE QUESTION: Which company is older, EOG Resources or General Mills?
        PARAGRAPHS:
        0: EOG Resources || Founded in 1999...
        1: General Mills || Founded in 1866...

        [STEP 1] Check founding dates in paragraph 0 and 1
        [STEP 2] EOG founded in 1999 (paragraph 0)
        [STEP 3] General Mills founded in 1866 (paragraph 1)
        FINAL: General Mills is older
        <|eot_id|>
<|start_header_id|>user<|end_header_id|>
Here is the QUESTION you are supposed to answer: The Zapruder film shows President John F. Kennedy's motorcade as it passed through a plaza that was named aNational Historic Landmark in what year?
PARAGRAPHS:
  0: Umbrella man (JFK assassination) || The "umbrella man", identified by the United States House Select Committee on Assassinations in 1978 as Louie Steven Witt, is a name given to a figure who appears in the Zapruder film, and several other films and photographs, near the Stemmons Freeway sign within Dealey Plaza during the assassination of John F. Kennedy.
  1: Warren Commission || The President's Commission on the Assassination of President Kennedy, known unofficially as the Warren Commission, was established by President Lyndon B. Johnson through Executive Order on November 29, 1963 to investigate the assassination of United States President John F. Kennedy that had taken place on November 22, 1963.  The U.S. Congress passed Senate Joint Resolution 137 authorizing the Presidential appointed Commission to report on the assassination of President John F. Kennedy, mandating the attendance and testimony of witnesses and the production of evidence.  Its 888-page final report was presented to President Johnson on September 24, 1964 and made public three days later.  It concluded that President Kennedy was assassinated by Lee Harvey Oswald and that Oswald acted entirely alone.  It also concluded that Jack Ruby also acted alone when he killed Oswald two days later.  The Commission's findings have proven controversial and have been both challenged and supported by later studies.
  2: Zapruder film || The Zapruder film is a silent, color motion picture sequence shot by private citizen Abraham Zapruder with a home-movie camera, as U.S. President John F. Kennedy's motorcade passed through Dealey Plaza in Dallas, Texas, on November 22, 1963, thereby unexpectedly capturing the President's assassination.
  3: Orville Nix || Orville Orhel Nix (April 16, 1911 – January 17, 1972) was a witness to the assassination of U.S. President John F. Kennedy in Dallas, Texas, on November 22, 1963.  His filming of the event, capturing only the last few seconds of the drama, is considered nearly as important as the Abraham Zapruder film.
  4: Marilyn Sitzman || Marilyn Sitzman (December 14, 1939 – August 11, 1993) was an American receptionist and a witness to the assassination of U.S. President John F. Kennedy in Dallas, Texas on November 22, 1963.  She was with her boss, Abraham Zapruder, as he made the Zapruder film, the most studied record of the assassination.
  5: Abraham Zapruder || Abraham Zapruder (May 15, 1905 – August 30, 1970) was a Russian-born American clothing manufacturer who witnessed the assassination of President John F. Kennedy in Dallas, Texas, on November 22, 1963.  Zapruder unexpectedly captured the event in a home movie while filming the presidential limousine and motorcade as it traveled through Dealey Plaza.
  6: Lee Harvey Oswald || Lee Harvey Oswald (October 18, 1939 – November 24, 1963) was an American former U.S. Marine who assassinated United States President John F. Kennedy on November 22, 1963.  According to four federal government investigations and one municipal investigation, Oswald shot and killed Kennedy as the President traveled by motorcade through Dealey Plaza in the city of Dallas, Texas.
  7: William Greer || William Robert Greer (September 22, 1909 – February 23, 1985) was an agent of the U.S. Secret Service, best known for having driven President John F. Kennedy's presidential limousine in the motorcade through Dealey Plaza in Dallas on November 22, 1963, when the president was assassinated.
  8: Assassination of John F. Kennedy || John F. Kennedy, the 35th President of the United States, was assassinated on Friday, November 22, 1963 at 12:30 p.m. Central Standard Time in Dallas, Texas while riding in a motorcade in Dealey Plaza.  Kennedy was fatally shot by a former U.S Marine, Lee Harvey Oswald, while he was riding with his wife, Jacqueline, Texas Governor John Connally, and Connally's wife, Nellie, in a presidential motorcade.  A ten-month investigation by the Warren Commission from November 1963 to September 1964 concluded that Oswald acted alone in shooting Kennedy, and that Jack Ruby also acted alone when he killed Oswald before he could stand trial.  Kennedy's death marked the fourth (following Lincoln, Garfield, and McKinley) and most recent assassination of an American President.  Vice President Lyndon B. Johnson became President upon Kennedy's death.
  9: Dealey Plaza || Dealey Plaza , in the historic West End district of downtown Dallas, Texas (U.S.), is the location of the assassination of John F. Kennedy on November 22, 1963.  The Dealey Plaza Historic District was named a National Historic Landmark in 1993 to preserve Dealey Plaza, street rights-of-way, and buildings and structures by the plaza visible from the assassination site, that have been identified as witness locations or as possible assassin locations.
<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

FINAL...

Tokenization complete.
Generated token IDs: tensor([    25,    220,   5162,     18,    198, 128001], device='cuda:0')
New tokens count: 6

=== RAW GENERATED TEXT ===
': 1963\n<|end_of_text|>'
```

LLama-3.2-1B

The 1B fine-tuned model is not able to answer most prompts correctly as shown.

Most examples are gibberish and end up repeating the questions.

Evaluation Metrics

- **Exact Match (EM)**: A prediction is considered correct if its lower-cased, punctuation-stripped span matches HotpotQA's reference answer string.
- **Token-level F1**: The harmonic mean of token precision and recall, again using the official HotpotQA normaliser.

Metrics are computed with the `evaluate` implementation of the `squad_v2` script; the required input schema is `{"id","prediction_text","no_answer_probability"}` for predictions and `{"id","answers"` for references.

Findings

1. Bigger models handle divergence better; less prone to overfitting
2. More structured prompts yield clearer, more complete steps and training with positive and negative-examples leads to better model performance

Future Directions and Learnings

We believe we could further improve the architecture through the below methods.

- **Critique fine tuning**: The architecture can be improved by implementing a critique fine tuning module after the supervised fine tuning. For this, we can use any LLM (the same Llama model or any other LM) to critique the output generated by our model when it is prompted to perform multi-hop reasoning given only the question and context paragraphs. Thus, the labels on which the model will be trained are the same as before, that is, gold facts and final answer, the only difference being we do not give candidates or the gold facts to the LM; instead, it learns to generate the gold chain and answer. Therefore, our curent fine tuning module could teach the model to differentiate between good and bad chains, while this additional module can teach it to generate the correct chain.
- **Increased Epochs**: Training with multiple passes over data can help our model to better learn the underlying representation .
- **Model Tuning**: We relied on a conservative model in order to ensure we could train our LM under hardware constraints. However, we could try slightly heavier models like 7B variant, no LoRA, or no quantizations in order to unlock greater performance.

We also learned new techniques as part of the project

- **Implementing PEFT and Compression**: Having read the theory about PEFT and Compression techniques, we challenged to use these tools to reduce training time.
- **Prompting Templates**: We learned that different LLMs are trained to accept different prompt templates and tokens. So, while Chat Markup Language may suit GPT-models, Llama has its own markup language that we used to optimize generation.
- **Hardware resourced needed**: We realized that fine-tuning our model, even with PEFT and Compression, was not possible on commodity CPU. Therefore, we explored different options like cloud compute and on-premise GPU to run our workloads. Furthermore, we ensured that padding was done to generate token sequences that have length as multiples of 8. This helped map the data more efficiently onto GPU for matrix operations.

References

Fang, Yuwei et al. (2020). "Multi-hop Graph Relation Network for Question Answering". In: *EMNLP*, pp. 884–895.

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models.* arXiv: `2407.21783 [cs.AI]`. URL: `https://arxiv.org/abs/2407.21783`.

Pandey, Tushar et al. (2025). *Adaptive Graph of Thoughts: Test-Time Adaptive Reasoning Unifying Chain, Tree, and Graph Structures.* arXiv: `2502.05078 [cs.AI]`. URL: `https://arxiv.org/abs/2502.05078`.

Qiu, Linping et al. (2019). "Dynamically Fused Graph Network for Multi-hop Reasoning". In: *ACL*.

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton (2017). "Dynamic Routing Between Capsules". In: *NeurIPS*, pp. 3856–3866.

Schlichtkrull, Michael et al. (2018). "Modeling Relational Data with Graph Convolutional Networks". In: *ECSW*.

Yang, Zhilin et al. (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yin, Zhangyue et al. (2023). *Rethinking Label Smoothing on Multi-hop Question Answering.* arXiv: `2212.09512 [cs.CL]`. URL: `https://arxiv.org/abs/2212.09512`.

Zhang, Jiahao et al. (2024). *End-to-End Beam Retrieval for Multi-Hop Question Answering.* arXiv: `2308.08973 [cs.CL]`. URL: `https://arxiv.org/abs/2308.08973`.