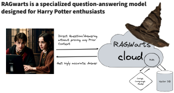
RAGwarts: Harnessing RAG for Harry Potter Question Answering

Project Wizards :: Mitesh Jalan, Pritish Thombare, Smit Kumbhani

Introduction:

RAGwarts is a specialized question-answering model designed for Harry Potter enthusiasts


Direct Question/Answering without proving any Prior Context
Get higly accurate Answer
RAGWarts cloud
RAG
Large Language Model
Vector DB

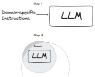
- RAGwarts is a specialized question-answering model designed for Harry Potter enthusiasts, offering accurate responses to enrich engagement with the franchise.
- It integrates RAG[1] (Retrieval-Augmented Generation) to merge retrieval-based and generative methods for more insightful answers.
- Multiple small language models, ranging from 1-3 billion parameters, are fine-tuned using Instruction Tuning to enhance their performance in answering Harry Potter trivia.
- The project addresses the challenge of understanding domain-specific language and context within the Harry Potter universe.

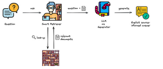
Background:

Pipeline
Vector DB

1. **Vector DB stores vector representations. (Enables efficient similarity search tasks.)**

Instruction Tuning: Fine-tunes language models for specific tasks by providing explicit instructions or examples, enhancing their ability to understand domain-specific language and context effectively.





RAG Framework: Combines retrieval-based and generative models for nuanced responses by leveraging a retriever to select relevant text passages and a generator to produce coherent answers.

Data

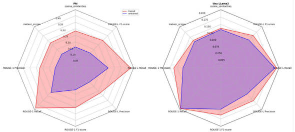
Task 1 (Vector Database Creation):

- Purpose: To provide text data for the retriever part of RAG.
- Data Sources: Harry Potter novels, screenplays, and movie transcripts sourced from the internet. (2.05 million tokens)

Task 2 (Instruction Tuning of Language Models):

- Purpose: To fine-tune small language models on Harry Potter trivia.
- Data Source: HuggingFace dataset of Harry Potter trivia questions and answers.
  (3,000 pairs in the train set and 500 pairs in the test set.)

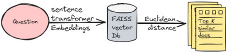
Results:



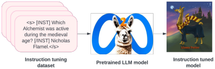
- The instruction tuned Phi-2[3] and Tiny-Llama[4] outperforms pretrained models.
- Statistical metrics: BLEU, METEOR, ROUGE, and cosine similarity.
- Qualitative Human evaluation.

Methods:

1) Retriever: Find top k similar documents from vector DB.

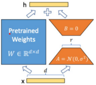


2) Large Language Model (Tuning): Instruction tuning using LoRA[2].



<s> [INST] Which Alchemist was active during the medieval age? [/INST] Nicholas Flamel.</s>

- **LoRA:**
  - Approximates parameters for faster, efficient inference.
  - Adjust only subset of parameters.
  - Reduces resource demands for training


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Conclusion:

- Developed a tailored question-answering model for Harry Potter enthusiasts using advanced NLP techniques like RAG and instruction tuning.
- Experimented with various fine-tuning methods and evaluated model performance using BLEU, METEOR, ROUGE, and Cosine Similarity metrics.
- Achieved valuable insights and enhancements in model capabilities, but faced challenges in balancing computational demands and effectiveness.
- Our project lays a foundation for continued exploration and improvement at the intersection of NLP and fandom engagement.
- Moving forward, we aim to refine our approach and address computational limitations to further enhance model performance and user experience.

References:

1. https://arxiv.org/abs/2005.11401
2. https://arxiv.org/abs/2106.09685
3. https://huggingface.co/microsoft/phi-2
4. https://github.com/jzhang38/TinyLlama
5. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.wizardingworld.com%2Fnews%2Fscholastic-celebrate-25-years-of-sorcerers-stone&psig=AOvVaw0gHXyQe1LuD_9cMihoFMTd&ust=1715237887065000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCLiR1qS9_YUDFQAAAAAdAAAAABAE
6. https://assets-global.website-files.com/63f3993d10c2a062a4c9f13c/659c0133ea40dffd68b7c977_1*hIosRqLDyhVERG7Twv5Jxg.png
7. https://www.google.com/url?sa=i&url=https%3A%2F%2Fharrypotter.fandom.com%2Fwiki%2FSorting_Hat&psig=AOvVaw2nuekqzhr7uQu54KrLeuXV&ust=1715240102583000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCODPorrF_YUDFQAAAAAdAAAAABAE