Whitepaper: Customizing DeepSeek R1 for Scientific Discovery

Alexander Migdal Overleaf GPT

January 26, 2025

Abstract

This whitepaper outlines a comprehensive plan to develop a domain-specific large language model (LLM) based on Hugging Face's DeepSeek R1. The goal is to create a specialized AI system capable of parsing, reasoning, and interacting with scientific content in fields such as turbulence, number theory, and quantum physics. Additionally, this project includes an AI-powered collaboration facilitator to streamline communication and inspire innovation among team members. By fine-tuning the model on domain-specific datasets and integrating it with platforms like Slack and Zoom, the project aims to build a dynamic tool for scientific discovery and team collaboration.

Contents

1 Introduction

Large Language Models (LLMs) have revolutionized natural language understanding and reasoning, but their generic nature often limits their utility in niche scientific fields. This project aims to customize DeepSeek R1 into a domain-specific LLM capable of:

- Parsing and reasoning about turbulence equations and their duality to string theory [2, 1].
- Extracting and processing content from scientific papers.
- Facilitating collaboration among researchers by collecting, analyzing, and synthesizing ideas during virtual meetings.

2 Dataset Preparation

2.1 Data Sources

The model will be trained on:

- Papers cited in publications by Alexander Migdal [2, 1].
- Recursive references from the initial papers, focusing on turbulence and number theory.
- Authoritative works in number theory, quantum physics, and related fields.

2.2 Automated Data Collection

The dataset will be curated using APIs and libraries:

- **Semantic Scholar API**: For retrieving citation networks.
- **arXiv API**: To gather preprints in turbulence and related domains.
- **CrossRef API**: To access metadata and references from journals.

2.3 Data Preprocessing

Data preprocessing will involve:

- Text extraction from PDFs using `PyPDF2` or `Grobid`.
- Cleaning equations and tokenizing mathematical expressions.
- Formatting data for fine-tuning (prompt-response pairs).

3 Fine-Tuning DeepSeek R1

3.1 Model Overview

DeepSeek R1 is an open-source LLM optimized for tunability and symbolic reasoning. Fine-tuning will adapt the pre-trained model to the scientific domain.

3.2 Training Pipeline

The fine-tuning process includes:

1. Tokenizing the curated dataset using Hugging Face's `transformers` library.
2. Training the model on GPU instances using transfer learning.
3. Evaluating performance on domain-specific tasks.

4 AI-Powered Collaboration Facilitation

4.1 Overview

To enhance teamwork, the AI agent will:

- Collect ideas, formulas, and code snippets from researchers asynchronously.
- Organize virtual meetings using Slack or Zoom to cross-fertilize ideas.
- Actively summarize and document discussions, providing actionable insights.

4.2 Workflow

1. **Collect Input**: The AI agent collects ideas from team members via Slack or a web interface.
2. **Curate Contributions**: Ideas are clustered and categorized using DeepSeek R1's reasoning capabilities.
3. **Organize Virtual Meetings**:
   - Schedule meetings via Zoom API.
   - Share agenda and discussion points with team members.
4. **Facilitate Meetings**:
   - Present curated insights and highlight connections between contributions.
   - Take real-time notes and generate summaries.
5. **Follow-Up**: Summarize discussions, assign tasks, and track progress.

4.3 Implementation Details

The AI agent will integrate with:

- **Slack API**: For asynchronous idea collection and notifications.
- **Zoom API**: To schedule and facilitate meetings.
- **DeepSeek R1**: For clustering ideas and reasoning about contributions.

4.4 Python Implementation: Slack Bot

Listing 1: Slack Bot for Idea Collection

```
import slack_sdk
from transformers import pipeline

# Initialize Slack client
slack_client = slack_sdk.WebClient(token="YOUR_SLACK_TOKEN")

# Initialize summarizer
summarizer = pipeline("summarization", model="huggingface/open-r1")

# Listen for user inputs
def process_slack_message(event):
    if "text" in event:
        user_input = event["text"]
        response = summarizer(user_input, max_length=100, min_length=30, do_sample=
        return response[0]["summary_text"]

# Send follow-up messages
def send_slack_message(channel, text):
    slack_client.chat_postMessage(channel=channel, text=text)

# Example usage
send_slack_message(channel="#team-ideas", text="Please␣share␣your␣latest␣ideas␣or␣c
```

4.5 Virtual Meeting Scheduling with Zoom API

Zoom API integration will automate meeting scheduling, ensuring that all team members are notified and provided with an agenda. The AI agent will generate the agenda based on clustered contributions.

5 Infrastructure Requirements

The collaboration facilitator requires:

- **Cloud Hosting**: To manage Slack and Zoom API interactions.
- **Compute Resources**: For real-time summarization and clustering using DeepSeek R1.
- **APIs**: Slack and Zoom for team communication and meeting organization.

6 Conclusion

By integrating a collaboration facilitator with DeepSeek R1, this project aims to revolutionize teamwork in scientific research. The AI agent will streamline idea collection, inspire cross-fertilization, and enhance productivity, ultimately driving innovative solutions in turbulence and beyond.

References

[1] Alexander Migdal. Fluid dynamics duality and solution of decaying turbulence, 2024.

[2] Alexander Migdal. Quantum solution of classical turbulence: Decaying energy spectrum. *Physics of Fluids*, 36(9):095161, 2024.