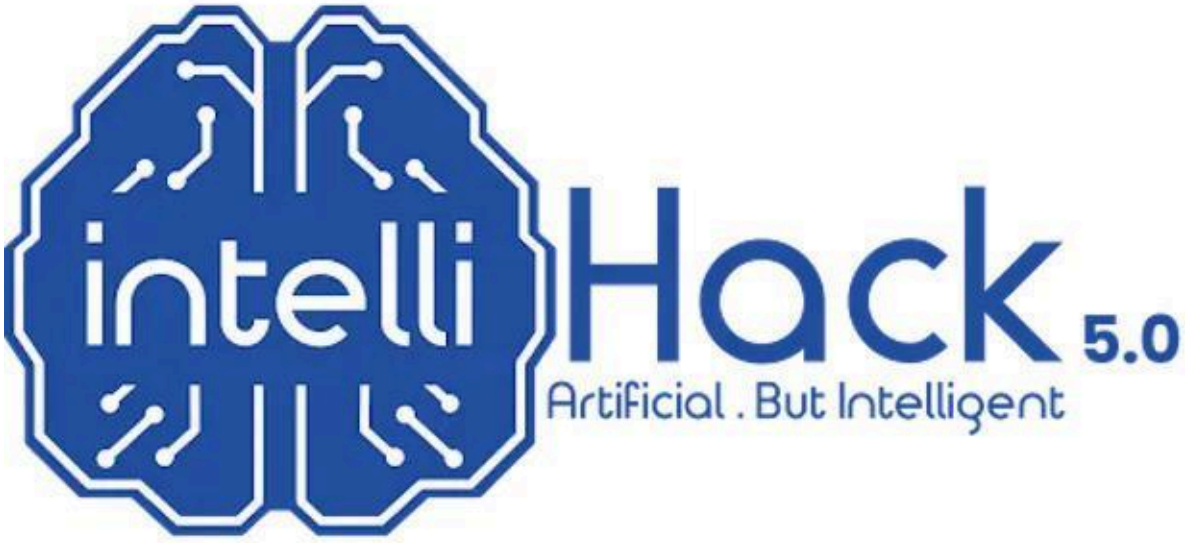
INTELLIHACK 5.0

Task 03 - Evaluation Report



Team Cognic AI

Loading the Fine-Tuned Model

After fine-tuning, we load the model using the **llama_cpp** library for efficient inference. This allows us to perform real-time question answering on research-based queries.

```
from llama_cpp import Llama

llm = Llama.from_pretrained(
    repo_id="AkinduH/Qwen2.5-3B-Instruct-Fine-Tuned-on-Deepseek-Research-Papers",
    filename="unsloth.Q4_K_M.gguf",
)
```

Evaluation Methodology

To assess the performance of our fine-tuned model, we implemented a **quantitative evaluation** using three key metrics:

1. Cosine Similarity (Semantic Similarity)

- We use the **all-MiniLM-L6-v2** model from **Sentence Transformers** to compute embeddings for both the **generated response** and the **ground truth**.
- The **cosine similarity** between these embeddings determines how semantically similar the responses are.

2. ROUGE Score (Text Overlap)

- The **ROUGE-L** metric evaluates **overlapping sequences** between the generated response and the ground truth.

3. BERTScore (Contextual Similarity)

- **BERTScore** computes similarity by leveraging contextual embeddings from **BERT**, rather than exact text matching.

Evaluation Process

1. **Sample Selection**: We selected **50 samples** from the **validation dataset**.
2. **Response Generation**: The model generates answers for each **unseen question**.
3. **Metric Computation**: We compute **cosine similarity, ROUGE-L, and BERTScore** for each sample.
4. **Results Analysis**: The averaged scores across all samples provide insight into the model's **effectiveness and reliability**.

Evaluation Results

Metric	Score
Cosine Similarity	0.8255
BERT Score	0.8968
ROUGE Score	0.2754

Observations & Insights

1. The model **generalizes well to unseen data**, showing strong semantic understanding.
2. While the **ROUGE score is lower**, it suggests the model tends to **rephrase rather than match exact words**.
3. The **high BERT Score** confirms that the model preserves meaning effectively.