QuestCrafter Project : Day 1 Summary

What We Did Today

Team: Kachallah Fatima, Joseph Fabrice TSAPFACK, Gemima Ondele, Mike-Brady

Today the group met and tackled the first day of the QuestCrafter project. Here's what we accomplished:

---

1. Project Setup & GitHub Organization

- Created the GitHub repository: GemimaOndele/QuestCrafter-AI_Project
- Added the project specification document (projet1-7.pdf)
- Started organizing the codebase structure with folders for data/, scripts/, models/.
- Set up initial .gitignore file to exclude common Python artifacts and large files

---

2. Dataset Selection

We reviewed the three recommended datasets from the project brief:

- WritingPrompts (Kaggle)
- TinyStories (Hugging Face)
- Reddit Jokes (Kaggle)

Decision: Reddit Jokes dataset

Why we chose it:

- Large corpus (1 million jokes) provides good training data diversity
- Simple structure: joke title maps to "prompt", punchline maps to "quest response"
- Natural variation in tone, length, and style — perfect for learning different quest flavors
- Easy to parse from CSV format

---

3. Initial Data Pipeline Script

Started writing download_data.py to handle:

- Loading Reddit Jokes CSV from Kaggle
- Cleaning text (removing entries that are too short or too long)
- Splitting data into train/validation/test sets (80/10/10 split)
- Exporting as JSONL format with schema:

```
{
 "prompt": "...",
 "response": "...",
 "source": "reddit_jokes",
 "metadata": { "score": ..., "author": "..." }
}
```

---

4. Roadmap & Next Steps

Agreed on the 5-week timeline:

- **Week 1 (current):** Data download, cleaning, preprocessing
- **Week 2:** Baseline generation with distilgpt2 (no training), test set design, evaluation rubric
- **Week 3:** Fine-tune model on cleaned data, track training curves
- **Week 4:** Full evaluation (baseline vs fine-tuned), add safety filters
- **Week 5:** Build demo app (Streamlit/Gradio), finalize report, present results

---

Status

Week 1, Day 1 Complete

- Repo created and initial structure in place
- Dataset chosen and justified
- Data pipeline script started
- Team aligned on the 5-week plan

Next milestone: Complete data preprocessing by Friday, January 31