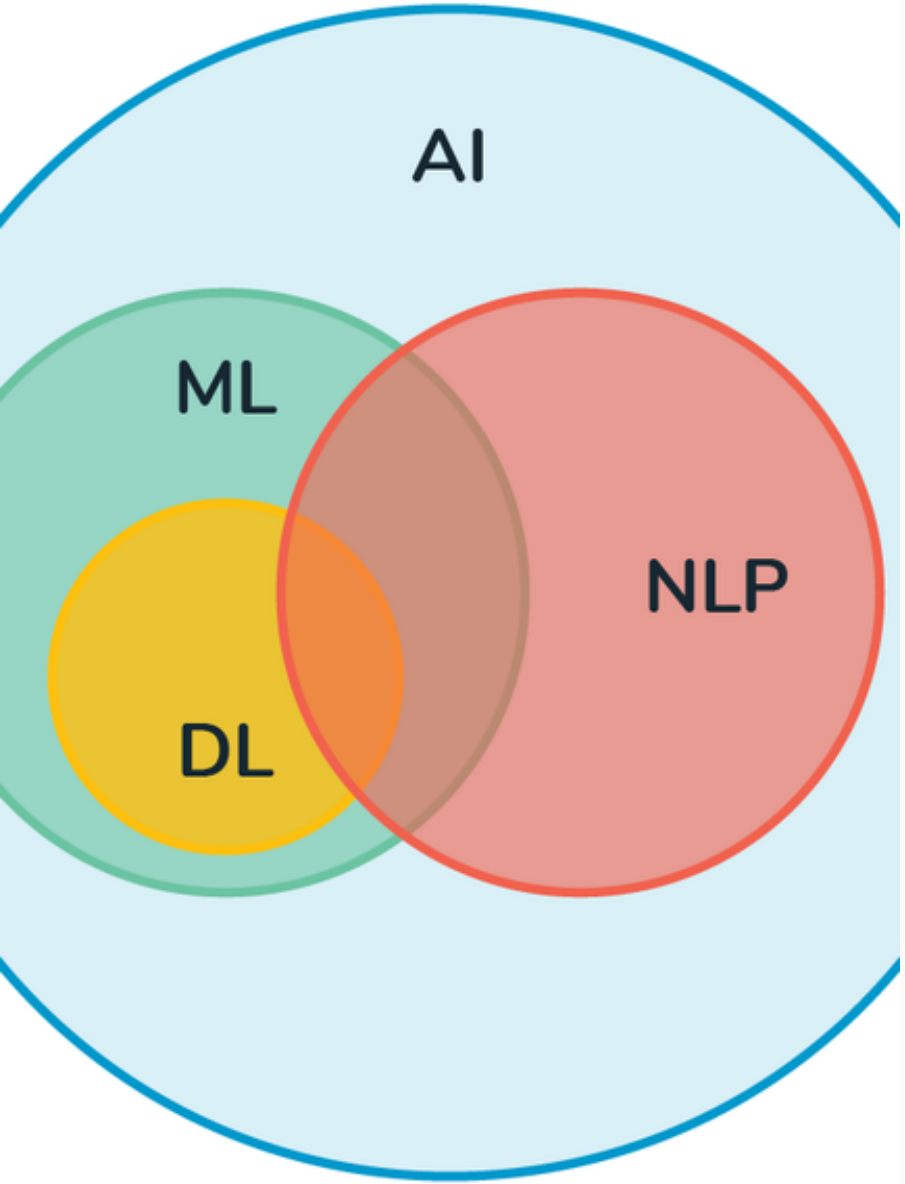
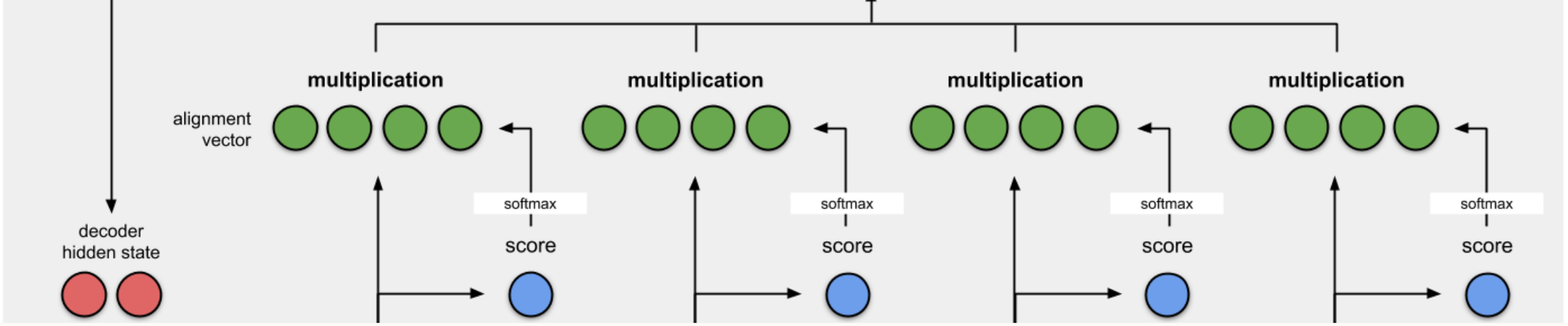


Project Overview

312707026 梁溢彥

This project is based on the training of the Alpaca-2 series open-source model.



Optimized Chinese Vocabulary

Expanded Word List 📖

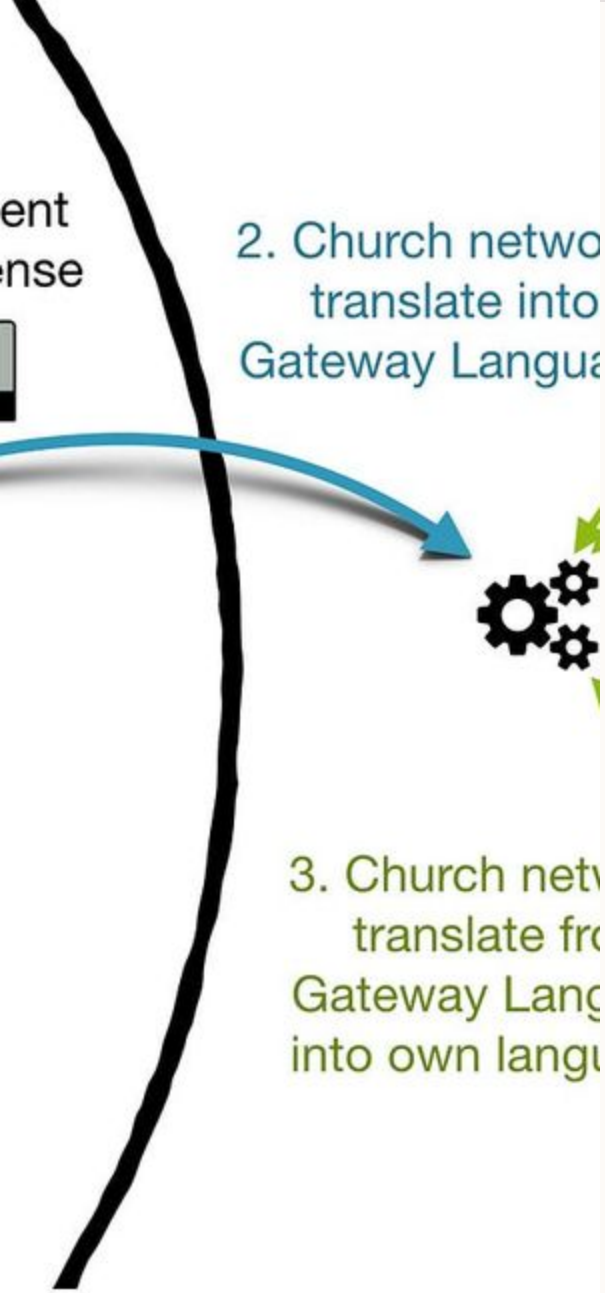
This model expanded the vocabulary of the LLaMA and Alpaca models (size 55296) to enhance coverage of Chinese words. The redesigned word list improves the model's encoding and decoding efficiency for Chinese text.

FlashAttention-2 Mechanism ⚡

This model implements the FlashAttention-2 for an efficient attention mechanism, especially suitable for processing long texts. It provides faster speed and optimizes memory usage.

Extended Long-context with PI and YaRN 🚄

Based on Positional Interpolation (PI) and Yield and Random Noise (YaRN) technologies, this model supports up to 64K context. These advancements enable the model to handle longer text without increasing complexity.



m 5: The Gateway Language p

Simplified Bilingual System Prompts

1

Streamlined System Prompts 👮

Compared to previous models, this version simplifies the system prompts for better adaptability in bilingual environments.

Human Preference Alignment

Reinforcement Learning from Human Feedback 👮

This model enhances its ability to convey correct values through Reinforcement Learning from Human Feedback (RLHF). The Alpaca-2-RLHF series follows the usage pattern of the SFT model.

File Structure

1 **inference.py**

Model inference script for generating answers.

2 **preprocessing.py**

Data preprocessing script to convert data into a format suitable for the model.

3 **datasets/**

Stores original and processed data.

4 **README.md**

Provides project overview and usage instructions.

Usage Instructions

1

Inference

Execute inference.py to generate answers: python inference.py --model_dir <model_path> [--cuda_ids <CUDA_device_IDs>] [--enable_8bit_quantization] [--enable_4bit_quantization]

preprocessing.py Script

Excel Data Reading 📊

Uses pandas to read Excel data from a specified path.

Data Processing 🔄

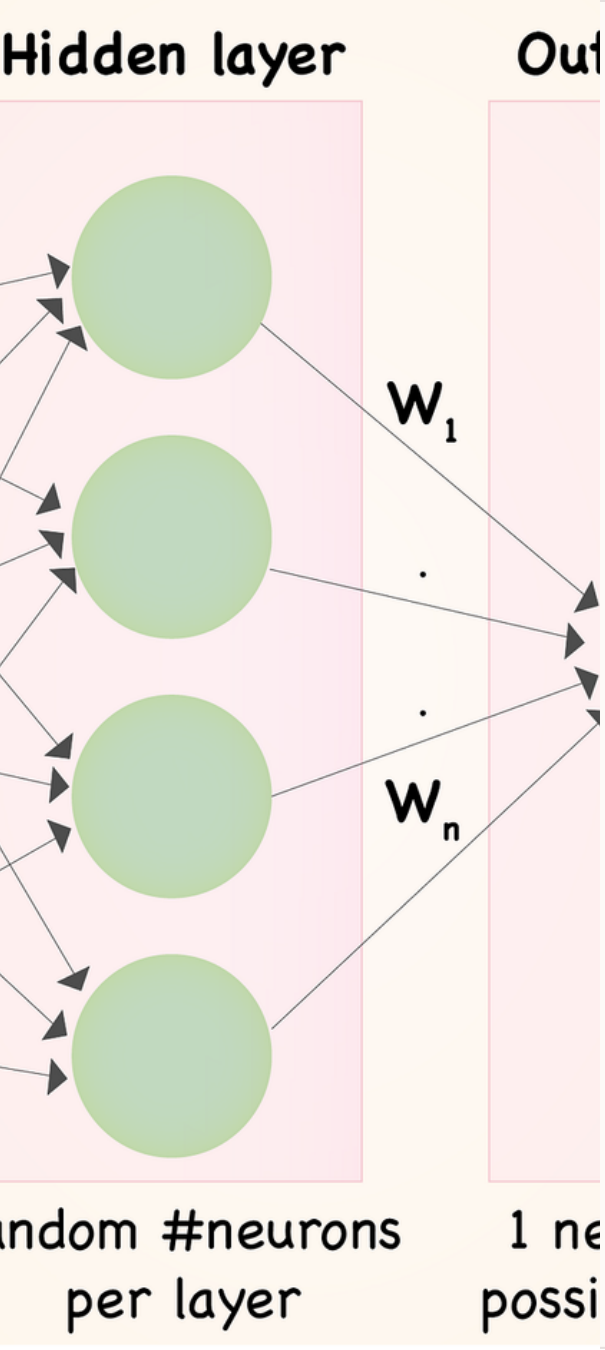
Converts answer format, removes unnecessary columns.

JSON Format 📋

Converts processed data into JSON format.

JSON File Saving 💾

Saves the converted JSON data to a file.



model_config.json Configuration File

1 **Main Parameters**

Defines the parameters and settings used for the LLaMA model.

2 **Pretraining 🧠**

This model underwent incremental training using large-scale unlabeled data based on the original Llama-2.

3 **Command Fine-tuning 🛠️**

Further fine-tuned using annotated command data to obtain the Chinese-Alpaca-2 series of models.