AIFS Multimodal Climate AI Architecture

ECMWF AIFS TimeSeries Tokenizer + Llama 3-8B Integration

Legend:
- AIFS Components
- Llama Components
- Tokenizers
- Fusion Mechanisms
- Data Sources
- Outputs

Input Layer

5D Climate Tensor
B=Batch, T=Time, V=Variables
H=Height, W=Width

ECMWF Climate Data
(GRIB/NetCDF)
5D Tensor: [B, T, V, H, W]
[2, 4, 73, 16, 16]
Variables: T, U, V, Q, Z, etc.

Climate Text Queries
"Analyze temperature patterns"
"Predict extreme weather"
Tokenized: [B, seq_len]
[2, 512 tokens]

AIFS Processing Pipeline

AIFS Data Processor
ECMWF AIFS-Single-1.0
• GraphCast architecture
• 73 variable fields
• Spatial: 16×16 → 512

TimeSeries Tokenization
Sequential → Embeddings

AIFS TimeSeries Tokenizer
• Transformer encoder backbone
• 5D → Token sequence
• Hidden dim: 512
• Temporal modeling: 4 steps

Llama Processing Pipeline

Llama 3-8B Tokenizer
• BPE encoding
• 128k vocabulary
• Special tokens: <|start|>, <|end|>
• Text → [B, seq_len] tokens

Multimodal Fusion Layer

Cross-Modal Attention
• Climate tokens → Llama embeddings
• Multi-head attention (32 heads)
• Hidden size: 4096
• Learnable projection layers

Cross-Modal Fusion
Climate ↔ Language

Fusion Strategies
• Cross-attention
• Token concatenation
• Adapter layers
• Interleaved tokens

Advanced Fusion
• Temperature-scaled attention
• Residual connections
• Layer normalization
• Dropout regularization

Model Integration Layer

Llama 3-8B Language Model
• 32 transformer layers
• 4096 hidden dimensions
• 8-bit quantization support
• Flash attention disabled (compatibility)

AIFS Encoder Architecture
• Pre-trained ECMWF weights
• GraphCast-based design
• 73-variable processing
• Autoregressive decoder

Integrated AIFS-Llama
• Joint embeddings
• Shared attention
• End-to-end training
• Zero-shot capabilities

AIFS Multimodal Specifications:

• AIFS Model: ECMWF AIFS-Single-1.0
• Language Model: Meta-Llama-3-8B
• Climate Variables: 73 fields
• Spatial Resolution: 0.25° × 0.25°
• Temporal Resolution: 3-hourly
• Context Length: 8192 tokens
• Fusion Method: Cross-attention

Output Generation Layer

Multimodal Climate Analysis
• Scientific explanations
• Weather predictions
• Trend analysis
• Uncertainty quantification
• Interactive responses