LexiLingo AI

Tài Liệu Tham Khảo

References & Resources

Comprehensive Reference Guide

Version: 2.0

Last Updated: January 2026

Author: Nguyen Huu Thang

Mục lục

1 Language Models

1.1 Large Language Models

Model	Paper/Link	Description
Qwen2.5	GitHub | Paper	Alibaba's multilingual LLM, strong on English & code
LLaMA 3	Meta AI | Paper	Meta's open LLM, excellent for fine-tuning

Bảng 1: Language Models

2 Speech Models

Model	Paper/Link	Description
Whisper v3	OpenAI | Paper	SOTA speech recognition, multilingual
Faster-Whisper	GitHub	CTranslate2 optimized Whisper (4x faster)
HuBERT	Paper | HuggingFace	Self-supervised speech representation
Piper TTS	GitHub	Fast, offline VITS-based TTS
Silero VAD	GitHub	Voice Activity Detection

Bảng 2: Speech Models

3 Fine-tuning Techniques

Technique	Paper	Description
LoRA	arXiv:2106.09685	Low-Rank Adaptation for efficient fine-tuning
QLoRA	arXiv:2305.14314	Quantized LoRA, enables 65B models on single GPU
PEFT	GitHub	HuggingFace Parameter-Efficient Fine-Tuning
SFT	TRL Docs	Supervised Fine-Tuning Trainer

Bảng 3: Fine-tuning Techniques

4 Datasets

4.1 Grammar Error Correction (GEC)

Dataset	Link	Size	Description
BEA-2019	CodaLab	34K	Write & Improve + LOCNESS shared task
FCE Corpus	Cambridge	1,244	First Certificate in English exam

Dataset	Link	Size	Description
CoNLL-2014	NUS	1,312	Grammatical Error Correction shared task
JFLEG	GitHub	1,501	Fluency-focused GEC benchmark
W&I+LOCNESS	HuggingFace	34K	Combined Write&Improve + LOCNESS
ERRANT	GitHub	Tool	Error annotation toolkit

Bảng 4: Grammar Error Correction Datasets

4.2 English Learner Corpora

Dataset	Link	Size	Description
EFCAMDAT	EF-Cambridge	83M words	EF-Cambridge Open Language Database, CEFR-labeled
TOEFL11	ETS	12,100	TOEFL essays with scores
ICNALE	Official	10K	International Corpus Network of Asian Learners
PELIC	GitHub	46K texts	Pitt English Language Institute Corpus

Bảng 5: English Learner Corpora

4.3 CEFR & Vocabulary

Dataset	Link	Description
CEFR-J Wordlist	Official	Japanese CEFR word frequency lists
English Profile	Cambridge	Official CEFR vocabulary lists
Oxford 3000/5000	Oxford	Core vocabulary lists
EVP	Cambridge	CEFR-graded vocabulary

Bảng 6: CEFR & Vocabulary Resources

4.4 Dialogue & Conversation

Dataset	Link	Size	Description
Intel/orca_dpo_pairs	HuggingFace	13K	DPO training pairs
Anthropic HH-RLHF	HuggingFace	170K	Human preference data
OpenAssistant	HuggingFace	161K	Multilingual conversation
Tatoeba	Official	10M+	Parallel sentences, good for translation

Bảng 7: Dialogue & Conversation Datasets

4.5 Pronunciation & Speech

Dataset	Link	Size	Description
TIMIT	LDC	6,300	Phoneme-aligned speech corpus
LibriSpeech	OpenSLR	1,000 hours	Clean speech for ASR
CommonVoice	Mozilla	17K+ hours	Multilingual speech corpus
L2-ARCTIC	GitHub	26 hours	Non-native English speech

Bảng 8: Pronunciation & Speech Datasets

5 Related Research Papers

5.1 Grammar Error Correction

Paper	Year	Link	Key Contribution
GECToR	2020	arXiv:2005.12592	Efficient sequence tagging for GEC
T5 for GEC	2021	ACL	Transfer learning approach
GrammarT5	2022	arXiv:2203.07442	Grammar pre-training
LLM-GEC	2023	arXiv:2303.13648	LLMs for GEC

Bảng 9: Grammar Error Correction Papers

5.2 Language Assessment

Paper	Year	Link	Key Contribution
CEFR Classification	2018	ACL	Automatic CEFR level prediction
Automated Essay Scoring	2020	arXiv:2012.13958	BERT for essay scoring
Fluency Assessment	2021	Interspeech	Speech fluency metrics

Bảng 10: Language Assessment Papers

5.3 Multi-task Learning

Paper	Year	Link	Key Contribution
MT-DNN	2019	arXiv:1901.11504	Multi-Task Deep Neural Networks
UniLM	2020	NeurIPS	Unified Language Model pre-training
T5	2020	arXiv:1910.10683	Text-to-Text Transfer Transformer

Bảng 11: Multi-task Learning Papers

6 Tools & Libraries

Tool	Link	Purpose
Transformers	HuggingFace	Model loading & inference
PEFT	HuggingFace	Parameter-efficient fine-tuning
TRL	HuggingFace	Transformer Reinforcement Learning
BitsAndBytes	GitHub	8-bit/4-bit quantization
vLLM	GitHub	High-throughput LLM serving
ERRANT	GitHub	Error annotation toolkit
Language Tool	GitHub	Rule-based grammar checking
Sentence-Transformers	GitHub	Sentence embeddings

Bảng 12: Development Tools & Libraries

7 Vietnamese NLP Resources

Resource	Link	Description
VinAI PhoGPT	GitHub	Vietnamese generative model
VietAI ViT5	HuggingFace	Vietnamese T5
UIT-ViNewsQA	GitHub	Vietnamese QA dataset
VLSP	Official	Vietnamese Language and Speech Processing

Bảng 13: Vietnamese NLP Resources

8 Mobile Deployment

Resource	Link	Description
ONNX Runtime	GitHub	Cross-platform inference
TensorFlow Lite	TensorFlow	Mobile deployment
llama.cpp	GitHub	CPU inference for LLMs
whisper.cpp	GitHub	CPU inference for Whisper
MLC LLM	GitHub	Universal LLM deployment

Bảng 14: Mobile Deployment Resources