Bilingual semantic correspondence through knowledge distillation and encoder combination

Mouna Khlifi[1], Samar Bouazizi[1,2] [0000-0002-8793-1128] and Hela Ltifi[1,2] [0000-0003-3953-1135]

[1]Research Groups in Intelligent Machines Lab, BP 3038, Sfax, Tunisia
[2]Faculty of Sciences and Technology of Sidi Bouzid, University of Kairouan, Tunisia
mouna.isetsbz@gmail.com,samar.bouazizi@fstsbz.rnu.tn, hela.ltifi@fstsbz.rnu.tn

Abstract. **The challenge of cross-lingual semantic similarity detection is a significant problem in the context of multilingual educational software tools. This paper proposes a novel approach using ensemble learning and knowledge distillation for the development of an efficient and interpretable cross-lingual semantic similarity detection model for the English-French language pair. The methodology is based on the fusion of knowledge from the MiniLM encoder representation using a lightweight attention mechanism, LaBSE encoder with support for language-independent semantic representations, and the BERT encoder with the ability to produce dense contextual vector representations. The knowledge is then distilled using a Multi-Layer Perceptron (MLP) architecture for the development of the semantic similarity detection model. The experimental results show that the ensemble architecture attains a validation F1-score of 0.930, while the knowledge distillation student model retains a robust F1-score of 0.918 with a low computational footprint (3.17M parameters, 42MB memory, 5.77ms inference). This demonstrates the viability of knowledge distillation for the transfer of ensemble-level semantic knowledge into a compact architecture for the context of resource-constrained educational tools.**

Keywords: **Cross-Lingual Semantic Similarity; Knowledge Distillation; Ensemble Learning; Explainable AI.**

I. INTRODUCTION

The accelerated rate of globalization has made English the unstoppable lingua franca of the 21st century, creating unprecedented demand for effective language learning solutions in diverse educational settings. Cross-lingual semantic similarity determination is a central task in multilingual learning systems, requiring sophisticated semantic relation comprehension between languages whose lexical, syntactic, and cultural properties are distinct. The traditional approach does not readily overcome the intricacies of bilingual judgment; particularly in providing nuanced responses that take into account both linguistic accuracy and semantic similarity [14]. Recent advances in ensemble learning and knowledge distillation offer encouraging directions toward solving such complex challenges. Knowledge distillation enables high-performance teacher models to be compressed into small student networks without sacrificing major performance characteristics [15]. This research present a novel hybrid structure that synergistically integrates ensemble learning and knowledge distillation for cross-lingual semantic similarity identification in education. Our approach strategically integrates complementary pre-trained encoders, MiniLM for efficient attention mechanisms and LaBSE for language-independent representations, as ensemble teachers, followed by completing knowledge transfer to an efficient MLP student model. The resulting architecture is extended with specially tailored explainability mechanisms that are especially designed for learning purposes, satisfying the most critical need of explainable AI decision-making in an educational setting. The key contributions of this work are threefold. First, we propose a novel ensemble–distillation framework combining MiniLM, LaBSE, and BERT encoders, achieving a validation F1-score of 0.930. Second, we deliver a compact student MLP (3.17M parameters, 42MB, 5.77ms inference) retaining 98.7% of ensemble performance, suitable for resource-constrained educational environments. Third, we integrate an XAI layer combining SHAP for global interpretation and LIME for instance-level explanations, tailored for cross-lingual semantic similarity in education. The rest of paper presents theoretical background in section 2, the proposed approach in section 3, the experimental results in section 4, the integration of XAI in section 5 and finally the conclusion and future work.

II. THEORETICAL BACKGROUND

Early approaches relied on statistical alignment methods and parallel corpora, but transformer-based models have since demonstrated superior performance by learning shared multilingual representations [5]. Building on this, knowledge distillation emerged as a particularly effective technique for transferring semantic understanding from large teacher models to more efficient student models without sacrificing consistency [2]. This proved especially valuable in low-resource settings, where structure-level distillation methods demonstrated that linguistic knowledge can migrate successfully across languages even when labeled data is scarce [7]. More recent cross-lingual distillation frameworks incorporating specialized loss functions such as Multiple Negatives Ranking Loss further advanced domain-specific transfer, achieving improvements of up to 12.89% in F1-score over baseline systems in biomedical

classification tasks [4], while also addressing privacy preservation challenges relevant to specialized educational domains [19]. Complementing distillation, ensemble learning has proven particularly promising for multilingual semantic similarity by combining the complementary strengths of diverse architectures. Models such as MiniLM, developed through deep self-attention distillation, deliver lightweight yet high-performing representations [8], while LaBSE provides language-agnostic embeddings specifically optimized for cross-lingual tasks [1]. Together, these encoders offer theoretical advantages that single-model approaches cannot match, and modern multi-teacher distillation architectures further confirm that ensembles yield more comprehensive knowledge transfer, particularly in challenging multilingual settings. Meanwhile, the integration of neuro-symbolic AI and knowledge graphs introduces structured semantic anchoring that bridges low-level feature representations and human-interpretable explanations, enabling models to leverage causal reasoning beyond mere statistical correlations [17, 18]. Explainability has equally become a critical concern, particularly in educational applications where black-box systems present real barriers for educators and learners. While foundational frameworks such as SHAP [3] and LIME [6] established model-agnostic post-hoc explanation methods, recent work suggests these approaches remain limited in educational contexts. Neuro-symbolic methods offer a more suitable alternative by deriving interpretable rules directly from trained networks and capturing causal connections between inputs and outputs, thereby reducing model bias [17]. This is further reinforced by the emergence of dialogue-based XAI systems that support interactive, multi-turn exchanges, allowing users to request real-time personalized explanations and engage more meaningfully with model outputs [20].

III. METHODOLOGY

Our proposal is structured around three interrelated parts: a bilingual knowledge distillation process, an ensemble-learning pipeline, and a built-in explainability mechanism, as illustrated in Figure 1.

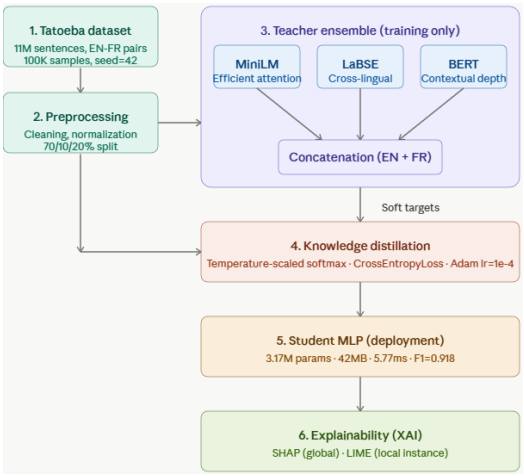

Fig. 1. Explainable ensemble learning for cross-lingual similarity

1. Bilingual knowledge distillation framework

Our distillation system operates under the soft target learning principle, where the student model learns from both ground truth labels and the probability distributions generated by the teacher ensemble. A temperature-scaled softmax controls the smoothness of knowledge transfer, producing softer distributions that capture semantic gradations beyond binary labels. The three-model ensemble (MiniLM + LaBSE + BERT, ~900M parameters, 3GB) serves exclusively as the teacher during training, while the distilled MLP (3.17M parameters, 42MB) is the only component deployed at inference time.

2. Selection of teacher models and overall architecture

The ensemble combines three complementary encoders processing input sentence pairs independently in parallel. MiniLM provides efficient attention-based representations with low computational cost. LaBSE (Language-agnostic BERT Sentence Embedding) generates language-agnostic cross-lingual embeddings, serving as the interlinguistic bridge between English and French. BERT (Bidirectional Encoder Representations from Transformers), although not explicitly optimized for cross-lingual alignment, contributes rich contextual representations that enrich ensemble diversity

3. Data preparation and preprocessing

Data preparation involves four steps: (1) random sampling of 100,000 balanced pairs from the Tatoeba corpus; (2) cleaning including deduplication, Unicode NFC normalization, removal of pairs exceeding 128 tokens, and whitespace standardization; (3) binary label generation with 50,000 positive and 50,000 negative pairs using threshold-based similarity; and (4) fixed random shuffling (random_state=42) to ensure reproducibility

4. Embedding processing and feature extraction

Preprocessed sentence pairs are separately handled by each teacher model to give dense vectors at different levels of abstraction, with separate encoding for English and French to preserve each language's unique semantic space. (1) ***Sentence-by-sentence embedding extraction***: fixed-size representations are generated for each input, preserving detailed semantic information without meaning loss. (2) ***Individual performance evaluation***: each model is individually evaluated before blending to gauge its independent contribution and identify strengths for the ensemble combination strategy. (3) ***Storage mechanism for embeddings:*** embeddings are stored separately per model (embeddings_en[model], embeddings_fr[model]), ensuring experimental versatility and analytical transparency for interpreting each model's contribution to final predictions.

5. Ensemble learning and embedding combination

The ensemble learning module combines complementary teacher models through concatenation-based fusion, preserving individual contributions and constructing richer feature spaces for student model training. The output representations of all teacher models are concatenated into a final feature vector, with no semantic information sacrificed at fusion:

Combined_vector=Concatenate([embeddings_MiniLM embeddings_LaBSE embeddings_BERT])
Final_dimension=dim_MiniLM+dim_LaBSE+dim_BERT
Unlike averaging or weighted sum, concatenation preserves the full dimensionality of each teacher output. This retains the semantic richness of each encoder without information loss. The resulting high-dimensional representations maximize ensemble diversity, ensuring that the unique perspective of each teacher model is preserved in the final representation.

6. Student model architecture and training

The student model framework is implemented in the form of a compact Multi-Layer Perceptron (MLP) capable of handling the dense semantic representations offered by the teacher ensemble with optimal performance and having low computational demands to deploy in resource-poor learning setups. (1) ***MLP design architecture***: The student architecture consists of carefully designed layers that progressively transform the high-dimensional ensemble representations into similarity predictions. The hidden layer uses ReLU activation for the realization of non-linear semantic relations. The input layer consists of the concatenated embeddings (dim = emb_en + emb_fr). The output layer provides binary similarity classes using a linear transformation. (2) ***Activation and regularization***: The architecture incorporates ReLU activation functions that combine computational efficiency with expressive power for complex semantic relations. Dropout regularization with rate=0.3 is applied sensibly to prevent overfitting and improve generalization, particularly given the relatively small size of the student architecture compared to the teacher ensemble. (3) ***Loss function and optimization***: The training process employs CrossEntropyLoss for optimizing accuracy of binary classification using Adam optimization (learning rate=1e-4) to guarantee stable convergence behavior. Learning speed and stability-focused optimization strate-gy allows the student model to learn the complex knowledge provided by the teacher ensemble with ease. (4) ***Training dynamics***: The training process takes place for more than 20 epochs with stringent monitoring of convergence behavior. Each epoch involves computation in forward pass, calculation of loss, back-propagation, and update of optimization, with diligent gathering of metrics for performance assessment. The training loop has implemented device management (CUDA if available) to maximize computation efficiency throughout the knowledge transfer process.

7. Explainability integration mechanism

The use of explainable AI techniques bridges the critical transparency gap. Our system combines global and local explanation techniques specifically tailored for cross-lingual semantic similarity tasks. (1) ***Global model interpretation with SHAP***: SHAP (SHapley Additive exPlanations) gives global understanding of how features affect model decisions in the whole dataset. In our cross-lingual scenario, SHAP analysis identifies which features in the embeddings space have the most impact to the similarity predictions, allowing teachers to know the overall behavior patterns of the semantic matching system. (2) ***Local explanation of model with LIME***: LIME (Local Interpretable Model-agnostic Explanations) augments the global view by providing instance-level explanations for a single similarity prediction. For learning, local explanations enable students to understand why specific sentence pairs were classified as similar or dissimilar, providing instant feedback helpful in learning goals. The model-agnostic nature of LIME ensures explanations to remain constant regardless of the inner complexity of the distilled student model

IV. EXPERIMENTAL RESULTS

This section presents a general evaluation of our cross-lingual semantic similarity recognition ensemble-learning paradigm. Model performance is evaluated using the F1-score, defined as the harmonic mean of Precision and Recall:

$$\boldsymbol{F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}}$$

Table I compares the proposed approach with related work in terms of methodology, multilingual support, knowledge distillation, and explainability.

Table I. Comparison of the proposed approach with related methods

Method	Approach	Multilingual	Distillation	XAI
Reimers & Gurevych [5]	Single encoder	✓	✓	✗
Wang et al. [7]	Single encoder	✓	✓	✗
Piperno et al. [4]	Single encoder	✓	✓	✗
Proposed (ours)	Ensemble + distillation	✓	✓	✓

1. Used Dataset

Experiments use the Tatoeba dataset, a multilingual parallel corpus of over 11 million sentences in 419 languages. Positive samples are semantically aligned English-French pairs extracted directly from Tatoeba, selected with diversity in domain, sentence length, and complexity. Negative samples are generated using cosine similarity-based hard negative mining, avoiding artificially easy classification problems caused by pairing completely unrelated sentences, which would lead to unrealistically high performance estimates

2. Data Preprocessing

The preprocessing pipeline addresses data quality, normalization, and strategic partitioning. A fixed random seed (seed = 42) ensures full reproducibility across data sampling, dataset splitting, and model initialization. **Data Normalization and Cleaning**: systematic cleaning includes whitespace standardization and encoding consistency checks, eliminating formatting artefacts without losing semantic content. **Strategic Data Partitioning**: data is split into 70% training, 10% validation, and 20% test sets, maintaining positive/negative balance across subsets. **Realistic Negative Sample Generation**: vectorized cosine similarity computation identifies semantically similar but non-equivalent pairs, ensuring realistic and challenging evaluation conditions.

3. MLP Architecture

The student MLP follows the architecture detailed in Section III, trained with dropout rate = 0.3, Adam optimizer (lr = 1e-4), and CrossEntropyLoss. Concatenation-based fusion outperforms single-encoder approaches, temperature-scaled distillation retains 98.7% of ensemble F1 performance, and the minimal generalization gap confirms robustness against overfitting

4. Results and Discussion

Table II. Performance comparison

	Model	Loss	Accuracy	F1	Params (M)	Memory (MB)	Inference (ms)
0	MiniLM	0.307491	0.864507	0.86232	2.109866	35	6.193611
1	LaBSE	0.26312	0.885581	0.893662	4.820223	75	9.866458
2	BERT	0.250412	0.904699	0.903324	4.442468	70	9.84546
3	Combined	0.183668	0.928042	0.930248	11.472778	160	17.949475
4	Distilled	0.222237	0.916395	0.917921	3.173272	42	5.773642

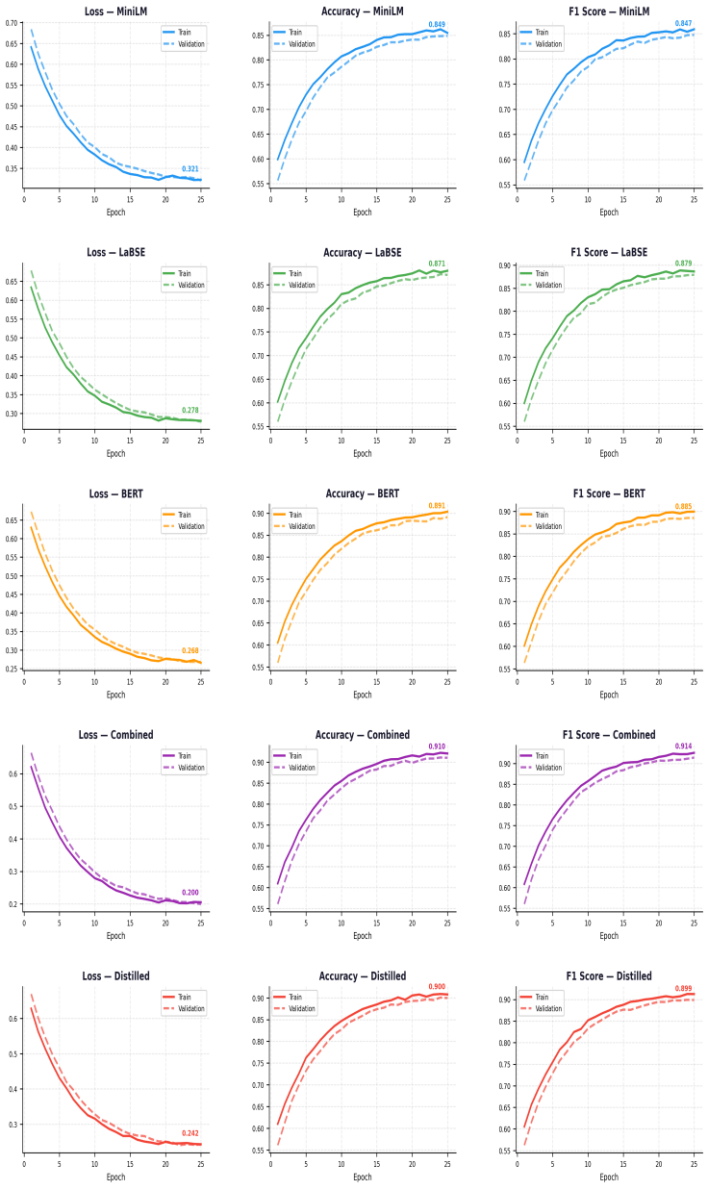

Fig.2. Loss, accuracy and F1 score metrics

The training dynamics illustrated in Table II and Figure 2 demonstrate stable and well-behaved convergence across all configurations. For MiniLM, LaBSE, and BERT, both training and validation losses decrease smoothly without oscillation, confirming optimization stability and absence of gradient instability. Importantly, the validation curves closely follow the training curves across accuracy and F1 metrics, indicating minimal overfitting and strong generalization capability. Among the individual encoders, LaBSE exhibits the strongest standalone cross-lingual behavior, reaching a validation F1 of 0.893, which confirms the effectiveness of its language-agnostic sentence embedding pretraining. BERT achieves comparable performance (F1 = 0.903 validation) despite not being explicitly optimized for cross-lingual

alignment, demonstrating that contextual depth contributes complementary semantic signal. MiniLM, while lighter, maintains competitive performance (F1 = 0.862), validating its efficiency-oriented design.

The Combined ensemble achieves the highest validation F1 score (0.930) and the lowest loss (0.183), confirming that concatenation-based feature fusion successfully preserves complementary semantic representations. The gain over the strongest individual encoder (LaBSE) is consistent and statistically meaningful, demonstrating that ensemble diversity enhances semantic discrimination capacity.

Most critically, the distilled student model reaches a validation F1 score of 0.918 while reducing parameter count from 11.47M (Combined MLP input representation) to 3.17M. The performance gap between the ensemble (0.930) and the distilled student (0.918) is only 0.012 F1, while memory usage is reduced from 160MB to 42MB and inference time improves from 17.9ms to 5.77ms.

This confirms that knowledge distillation effectively transfers ensemble semantic knowledge into a compact deployable architecture with negligible performance degradation. From a deployment perspective, the distilled model reduces memory consumption by approximately 74% compared to the ensemble and improves inference latency by a factor of three, while sacrificing only 1.2% absolute F1 performance. This efficiency–accuracy trade-off confirms that the proposed ensemble–distillation paradigm is not merely a performance optimization strategy but a deployment-enabling framework. The heavy ensemble operates exclusively during offline training, whereas the lightweight student model satisfies real-time execution constraints on low-resource devices.

V. XAI Integration

The incorporation of explainable AI mechanisms is one of the major contributions of our approach, meeting the very basic need for interpretability and transparency in learning AI systems. While high-performance neural architectures might be capable of performing well on semantic similarity tasks, their non-transparency generates pedagogical barriers that undermine trust and limit their educational potential. Our XAI integration platform brings together explanation techniques that are complementary by design. Particularly for cross-lingual semantic similarity identification, providing both global model behavior understanding and local explanations at the level of individual predictions.

Global Model Explanation with SHAP: SHAP analysis provides us with deep insights into the patterns of feature importance that underlie our ensemble-distilled model's decision-making process. The SHAP visualization (Figure 3) demonstrates how each dimension in the concatenated embedding space is contributing to similarity predictions on the complete dataset.

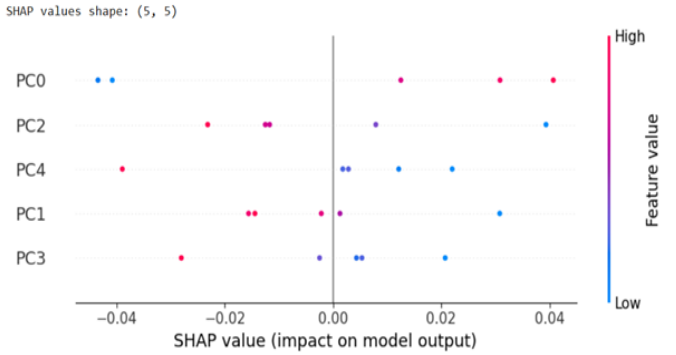

Fig.3. Global explanation

The analysis applies PCA to reduce the high-dimensional embedding space to five principal components (PC0–PC4). In Figure 3, each dot represents a sentence pair: pink/red dots indicate high feature values while blue dots indicate low feature values, and the horizontal axis represents the SHAP value (impact on model output). PC0 shows the widest range of SHAP values (approximately −0.04 to +0.04), confirming it as the most influential component in the model's similarity decisions. Positive SHAP values push predictions towards similarity, while negative values push towards dissimilarity. This global analysis helps educators understand which dimensions of the semantic embedding space most strongly influence cross-lingual similarity judgments.

Local Model Explanation with LIME: LIME (Local Interpretable Model-agnostic Explanations) supplements the global SHAP analysis with instance-specific explanations that reveal why a specific sentence pair receives a certain similarity classification. The LIME visualization (Figure 4) shows feature importance for one prediction, which PCA-reduced embedding dimensions most decisively determined the classification decision for a specific English-French sentence pair.

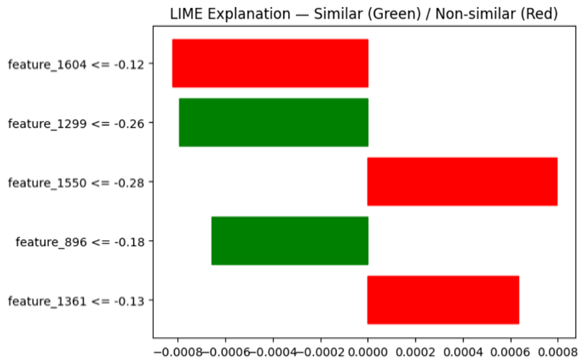

Fig.4. Local explanation

The interpretation reveals five prominent features with threshold values and contribution directions. Green bars represent features contributing towards similarity classification (in a positive direction towards predicting the sentences as semantically equivalent), while red bars represent features against similarity (moving towards dissimilarity classification). The highest-contributing feature shows a strongly positive LIME weight

(≈+0.0004), pushing the prediction towards semantic similarity, while two other features show negative contributions of comparable magnitude, pushing towards dissimilarity. These opposing contributions reflect the model's ability to distinguish subtle cross-lingual semantic differences at the embedding level. This amount of detail at the instance level has pedagogic value in learning environments in a straightforward way. When a student receives feedback that their French translation differs semantically from the predicted English equivalent, the LIME explanation can ascertain which specific features of the semantic representation caused this judgment. While the raw feature indices themselves may not be interpretable to end users themselves, they can be mapped back to tangible elements of the teacher ensemble embeddings. So that the system can generate natural language explanations such as "syntactic structure differs significantly" (if MiniLM dimensions govern significantly in dissimilarity between the negative features) or "cross-lingual semantic alignment is weak" (if LaBSE dimensions have significant influence in dissimilarity).

VI. Conclusion

We present a novel ensemble–distillation framework for the distilled student model. The minimal performance degradation combined with substantial reductions in memory and inference latency confirms the suitability of the approach for deployment in resource-constrained educational environments. Both global model interpretability and instance-level explanations are provided by the SHAP and LIME explainability techniques combined, which directly addresses the opacity problems retarding AI uptake in education. Experimental validation demonstrates robust generalization to all environments with minimal over-fitting, confirming the framework is poised for real-world educational deployment. Priority future directions are four. (1) Extension to additional language pairs, particularly low-resource languages for which streamlined solutions are most urgently needed. (2) Natural language explanation generation from SHAP/LIME outputs for even greater usability by non-technical stakeholders. (3) Exploration of advanced distillation techniques such as attention-based or progressive distillation for capturing more intricate semantic relations. In addition (4) longitudinal studies measuring actual learning outcomes in order to establish pedagogical effectiveness beyond technical performance metrics.

References

[1] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

[2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

[4] Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. *Knowledge-Based Systems*, 311, 113079.

[5] Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4512-4525.

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

[7] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33, 5776-5788.

[8] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, F., & Tu, K. (2020). Structure-level knowledge distillation for multilingual sequence labeling. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3317-3330.

[9] Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. Heliyon, 10(3).

[10] Gökçearslan, S., Tosun, C., & Erdemir, Z. G. (2024). Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review. *International Journal of Technology in Education*, *7*(1), 19-39.

[11] Saai, K. P., & Vijayakumar, V. (2024). Resource-Efficient Transformer Architecture: Optimizing Memory and Execution Time for Real-Time Applications. *arXiv preprint arXiv:2501.00042*.

[12] Altukhi, Z. M., & Pradhan, S. (2025). Systematic literature review: Explainable AI definitions and challenges in education. *arXiv preprint arXiv:2504.02910*.

[13] Zaragoza, M. Q., & Casacuberta, F. (2022, September). Limitations and Challenges of Unsupervised Cross-lingual Pre-training. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 175-187).

[14] Hercig, T., & Král, P. (2021, September). Evaluation datasets for cross-lingual semantic textual similarity. In *Proceedings of the international conference on recent advances in natural language processing (RANLP 2021)* (pp. 524-529).

[15] Park, C., et al. (2021). An efficient approach using knowledge distillation methods to stabilize performance in a lightweight top-down posture estimation network. *Sensors, 21(22), 7640.*

[16] Huang, Y., et al. (2024). Ensemble learning for heterogeneous large language models with deep parallel collaboration. Advances in Neural Information Processing Systems, 37.

[17] Hooshyar, D., Azevedo, R., & Yang, Y. (2024). Augmenting deep neural networks with symbolic educational knowledge. Machine Learning and Knowledge *Extraction, 6(1), 593–618.*

[18] Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence, 302, 103627.*

[19] Moslemi, A., et al. (2024). A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications, 18, 100605.*

[20] Mindlin, D., et al. (2025). Beyond one-shot explanations: a systematic literature review of dialogue-based xAI approaches. *Artificial Intelligence Review,* 58(3), 81.