CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Two-Stage Early Exiting From Globality Towards Reliability

Jianing He | Qi Zhang | Hongyun Zhang | Duoqian Miao

School of Computer Science and Technology, Tongji University, Shanghai, China

Correspondence: Duoqian Miao (dqmiao@tongji.edu.cn)

Received: 8 November 2024 | **Revised:** 1 February 2025 | **Accepted:** 6 March 2025

Funding: This work was supported by the National Natural Science Foundation of China (No. 62376198), the National Key Research and Development Program of China (No. 2022YFB3104700), and the Shanghai Baiyulan Pujiang Project (No. 08002360429).

Keywords: early exiting | inference acceleration | pre-trained language model | principal component analysis | three-way decisions

ABSTRACT

Early exiting has shown significant potential in accelerating the inference of pre-trained language models (PLMs) by allowing easy samples to exit from shallow layers. However, existing early exiting methods primarily rely on local information from individual samples to estimate prediction uncertainty for making exiting decisions, overlooking the global information provided by the sample population. This impacts the estimation of prediction uncertainty, compromising the reliability of exiting decisions. To remedy this, inspired by principal component analysis (PCA), the authors define a residual score to capture the deviation of features from the principal space of the sample population, providing a global perspective for estimating prediction uncertainty. Building on this, a two-stage exiting strategy is proposed that integrates global information from residual scores with local information from energy scores at both the decision and feature levels. This strategy incorporates three-way decisions to enable more reliable exiting decisions for boundary region samples by delaying judgement. Extensive experiments on the GLUE benchmark validate that the method achieves an average speed-up ratio of 2.17× across all tasks with minimal performance degradation. Additionally, it surpasses the state-of-the-art E-LANG by 11% in model acceleration, along with a performance improvement of 0.6 points, demonstrating a better performance-efficiency trade-off.

1 | Introduction

Recently, pre-trained language models (PLMs) have achieved remarkable improvements in various natural language processing (NLP) tasks [1–7]. However, PLMs are notorious for high computational costs and long inference latency, posing great challenges to their deployment in resource-constrained devices and real-time applications. Additionally, the overthinking problem [8] is another challenging issue for the application of PLMs. Specifically, for most easy samples, the model's shallow-layer representations are sufficient for generating correct predictions. In contrast, due to PLMs' over-parameterisation, the deep-layer representations often become overly complex or contain excessive noise and class-irrelevant redundant information, leading to incorrect predictions (per Figure 1). Hence, the overthinking problem not only impairs the task performance but also affects the inference efficiency of PLMs.

To address these issues, a branch of literature focuses on accelerating the inference of PLMs via early exiting Xin et al. [9]; Zhou et al. [10]; Liao et al. [11]; Xin et al. [12]; Balagansky and Gavrilov [13]; Zeng et al. [14]; He et al. [15, 16]. Early exiting is an important dynamic inference strategy. As shown in Figure 2, an internal classifier is added at each intermediate

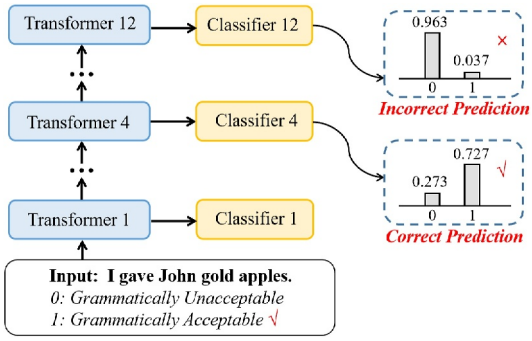


FIGURE 1 | The overthinking problem in BERT. On an easy sample labelled as 1 in the CoLA task, the fourth internal classifier is sufficient for making a correct prediction, while the final classifier produces an incorrect prediction due to over-parameterisation.

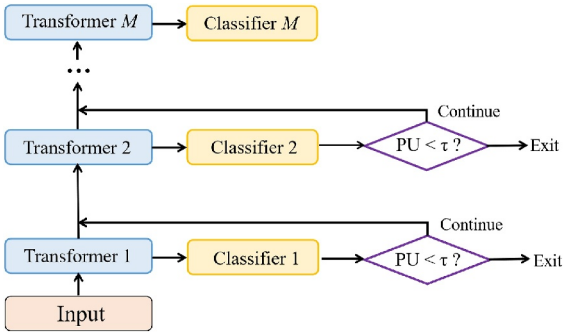


FIGURE 2 | Classic early exiting framework for PLMs. PU denotes the prediction uncertainty. τ denotes the threshold.

layer of the PLM to provide an early prediction during inference, allowing samples to exit from the current layer once the prediction uncertainty falls below a predefined threshold τ. This enables a sample-wise dynamic inference procedure to process easy samples with shallow layers and to predict hard samples with deep layers, avoiding passing all samples through the entire model. Early exiting can effectively improve the inference efficiency of PLMs without sacrificing accuracy and also mitigate the overthinking problem.

A typical implementation of early exiting involves devising an exiting signal to estimate prediction uncertainty, thus determining whether samples should exit from early layers. There are mainly three exiting strategies according to the types of exiting signals. The first is the score-based exiting strategy, for example DeeBERT Xin et al. [9], Right-Tool Schwartz et al. [17] and E-LANG Akbari et al. [18], which determines exiting for each sample by calculating the entropy, softmax score, or energy score of the prediction probability distribution on that sample. The exiting criterion is met once the entropy or energy score (softmax score) falls below (exceeds) a predefined threshold. This strategy allows for continuous adjustments to the speed-up ratio, while also enabling samples to exit immediately once the prediction uncertainty of an intermediate layer is sufficiently low. However, research indicates that it also suffers from the issue of overconfidence Li et al. [19], that is, an underestimation of prediction uncertainty, which can lead to the premature exiting of samples with incorrect predictions, thereby compromising the model's task performance. The second is the patience-based exiting strategy, for example PABEE Zhou et al. [10] and F-PABEE Gao et al. [20], which allows a sample to exit early once a sufficient (i.e. reaching the threshold) number of consecutive internal classifiers provide consistent predictions on that sample. Compared to the score-based strategy, this strategy effectively mitigates overconfidence by ensembling the outputs of multiple classifiers, thus delivering more robust early exiting. However, it fails to provide continuous adjustments to the speed-up ratio due to the discrete nature of its exiting signals. It may also delay the exiting of correctly predicted samples at higher thresholds, resulting in redundant computations and extended inference time. The last is the learning-based exiting strategy, for example BERxiT Xin et al. [12], PALBERT Balagansky and Gavrilov [13] and ConsistentEE Zeng et al. [14], which leverages neural networks to produce exiting signals for each sample based on its intermediate-layer representations. Unlike the first two heuristic exiting strategies, which enable the direct computation of exiting signals and are easy to implement, the learning-based strategy requires a learning process to formulate its exiting signals, resulting in additional training overhead. Additionally, its performance is highly influenced by the quality and diversity of the training data. Nevertheless, this strategy can adaptively identify intricate high-dimensional features from the data without manual intervention, thus producing more effective exiting signals.

However, most existing works rely on the model outputs for individual samples (i.e., local information) to formulate their exiting signals but overlook the global information from the sample population that is closely related to prediction uncertainty. Notably, samples that deviate from the population distribution are more likely to be out-of-distribution (OOD) samples, and the model's predictions for these OOD samples typically exhibit higher uncertainty levels than those for in-distribution (ID) samples. Therefore, the deviation of samples from the population distribution can effectively reflect the prediction uncertainty, which is ignored by the aforementioned exiting strategies, degrading the accuracy of prediction uncertainty estimation and leading to unreliable decisions regarding exiting.

In this paper, we propose to enhance prediction uncertainty estimation by considering the deviation of samples from the population distribution, thus delivering more reliable exiting decisions. To this end, inspired by principal component analysis (PCA), we formalise the key patterns of the sample population distribution using the principal space, that is, the subspace spanned by the principal components of the training data. We then introduce a residual score that captures the deviation of features from the principal space to estimate prediction uncertainty from a global perspective. Intuitively, the residual score reflects the likelihood of a sample being OOD, with higher values indicating greater prediction uncertainty. Accordingly, we propose a two-stage exiting strategy that integrates global information from residual scores with local information from energy scores at both the decision and feature levels (per Figure 3). By incorporating three-way decisions Yao [21], this strategy introduces the second decision stage for boundary

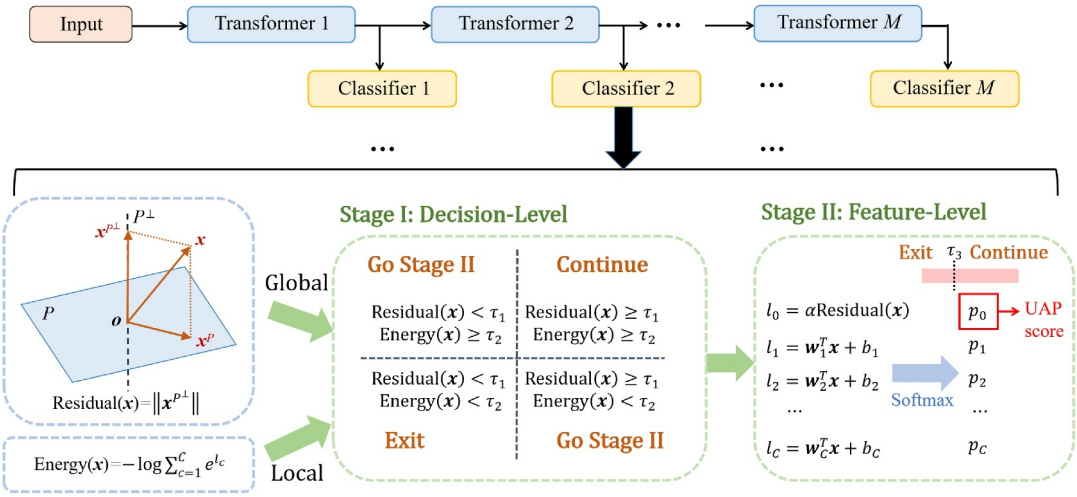


FIGURE 3 | Method overview. Our method employs a two-stage exiting strategy based on principal component analysis and three-way decisions. This strategy integrates global information from residual scores with local information from energy scores at both the decision level (in stage I) and the feature level (in stage II). $\boldsymbol{x}$ represents the sample feature. The principal space P and the scaling parameter α are determined by the training data. $\boldsymbol{w_1} \sim \boldsymbol{w_C}$ and $b_1 \sim b_C$ represent the weights and biases of the internal classifier, respectively. τ_1, τ_2 and τ_3 are the thresholds for residual, energy and UAP scores, respectively.

region samples, enabling a deeper integration of local and global information to enhance prediction uncertainty estimation for more reliable exiting decisions. Compared to our similar work DE^3-BERT He et al. [15], which caters to classification tasks and utilises a single-level integration of local and global information, our method is suitable for non-classification tasks as it calculates the residual score based on sample features. Moreover, our dual-level information integration strategy further reduces information loss and offers varied perspectives for exiting decision-making.

Our contributions are summarised as follows.

- We reveal the limitations of current methods that integrate global and local information, particularly in their scalability and information integration strategies.
- We define a residual score to capture the deviation of features from the principal space, estimating prediction uncertainty from a global perspective.
- We propose a two-stage exiting strategy based on three-way decisions, aiming to enhance the reliability of exiting decisions by integrating global information from residual scores with local information from energy scores at both the decision and feature levels.

Extensive experiments on the GLUE benchmark demonstrate that our method outperforms the state-of-the-art E-LANG by an average of 11% in inference speed and 0.6 points in task performance, with negligible additional computational or storage overhead. An in-depth analysis further verifies the generality and interpretability of our method.

The rest of this paper is organised as follows. Section 2 provides an overview of related works. Section 3 details our proposed early exiting method. Section 4 presents the experiments and in-depth analysis. Finally, Section 5 concludes this paper.

2 | Related Works

In this section, we review related works in three aspects: exiting strategy design, architecture/loss design, and the three-way decision.

2.1 | Exiting Strategy Design

Existing exiting strategies for PLMs can be roughly divided into three categories: score-based exiting strategies, patience-based exiting strategies, and learning-based exiting strategies. Score-based strategies use scoring functions to estimate prediction uncertainty based on the logits or probability distributions offered by internal classifiers. The existing scoring functions include the entropy in DeeBERT Xin et al. [9] and FastBERT Liu et al. [22], the softmax score in Right-Tool Schwartz et al. [17] and the energy score in E-LANG Akbari, Banitalebi-Dehkordi, and Zhang [18]. The exiting condition is met once the entropy or energy score (softmax score) falls below (exceeds) the threshold. Patience-based strategies use cross-layer consistency to estimate prediction uncertainty for exiting decision-making. In PABEE Zhou et al. [10], early exiting is triggered when a sufficient number of consecutive internal classifiers produce identical predictions. LECO Zhang et al. [23], BADGE Zhu et al. [24] and F-PABEE Gao et al. [20] introduce softer cross-layer comparison strategies to deliver more flexible early exiting. PCEE-BERT Zhang et al. [25] utilises a hybrid exiting signal that combines entropy and patience to jointly enhance the reliability and flexibility of exiting decisions. Learning-based strategies learn to make exiting decisions. BERxiT Xin et al. [12], PALBERT Balagansky and Gavrilov [13] and ConsistentEE Zeng et al. [14] use neural networks to generate exiting signals. HASHEE Sun et al. [26] and BE3R Mangrulkar , MS, and Sembium [27] train neural networks to predict the exiting layer for each sample or token, requiring no layer-by-layer exiting judgements.

The strategies mentioned above solely focus on local information from individual samples, neglecting the global information suggested by the sample population. This affects the estimation of prediction uncertainty, thus compromising the reliability of exiting decisions. DE^3-BERT He et al. [15] attempts to integrate distance-based global information to enhance the entropy-based early exiting. However, the extraction of global information and the strategies for integrating local and global information remain underexplored. In this paper, inspired by PCA and three-way decisions, we introduce a novel two-stage exiting strategy that integrates global information from residual scores with local information from energy scores at both the decision and feature levels, aiming for more reliable exiting decisions.

2.2 | Architecture/Loss Design

Some studies focus on the architecture design of early exiting networks. CascadeBERT Li et al. [19] employs multiple cascaded complete networks rather than a single multi-exit network to facilitate comprehensive representations for accurate predictions. GPFEE Liao et al. [11] integrates both past and future states to enhance early predictions. LECO Zhang et al. [23] formalises the architecture design of early exiting networks as a neural architecture search problem. BADGE Zhu et al. [24] incorporates block-wise bypasses to mitigate cross-layer optimisation conflicts. DisentangledEE Ji et al. [28] introduces adaptors to disentangle generic language representations from task-specific representations and puts forward a non-parametric classifier for enhancements. Other studies focus on improving the training objectives of early exiting networks. CascadeBERT Li et al. [19] introduces a difficulty-aware regularisation to calibrate the model outputs. LeeBERT Zhu [29] and GAML-BERT Zhu et al. [30] introduce the distillation loss to foster cross-layer mutual learning among classifiers.

Different from our method that focuses on refining early exiting strategies, the methods mentioned above enhance the architectures or training objectives for early exiting networks. Integrating our method with these orthogonal works is worth further research and exploration.

2.3 | Three-Way Decision

The three-way decision was proposed by Yao [21] to address the problem of region partition in rough sets. This theory aligns with the human decision-making process, enabling immediate decisions for highly certain items while delaying judgement on less certain items to enhance decision reliability. Since 2010, significant advancements have been made in the theoretical research of three-way decisions Yao [31, 32]. Recently, the three-way decision has demonstrated successful applications not only in the field of rough sets and granularity computing Wang and Zhu [33]; Gou and Zhang [34]; Yuan et al. [35, 36]; Li et al. [37, 38]; Wang et al. [39], but also in classification tasks Han et al. [40] and clustering tasks Guo et al. [41].

In this paper, we introduce a two-stage exiting strategy based on three-way decisions, aiming to more effectively address the early exiting of boundary region samples by delaying judgement.

3 | Methods

In this section, we illustrate the problem definition and the proposed early exiting method in detail.

3.1 | Problem Definition

As shown in Figure 3, we adopt a BERT-style PLM with M layers as the backbone model. $h^{(m)}$ denotes the hidden states at the m-th layer. Given a classification task involving C classes, an internal classifier F_m where $m \in \{1, 2, ..., M-1\}$ is attached to each intermediate layer to produce an early prediction $p^{(m)}$ by mapping the hidden states $h^{(m)}$ into a probability distribution over the C classes: $p^{(m)} = F_m(h^{(m)})$, allowing samples to exit early during inference when the estimated prediction uncertainty is sufficiently low.

3.2 | Method Overview

We propose a novel early exiting method for PLMs based on principal component analysis (PCA) and three-way decisions, which integrates local and global information at both the decision and feature levels to enhance the reliability of exiting decisions. Figure 3 provides an overview of our method. Firstly, through principal component analysis on the training set, we first define a residual score in Equation (4) that captures the deviation of features from the principal space to provide a global perspective for prediction uncertainty estimation. On this basis, we further propose a two-stage exiting strategy using three-way decisions, which integrates global information from residual scores with local information from energy scores at both the decision and feature levels to ensure reliable exiting decisions. Specifically, integration occurs at the decision level in the first stage. If a consistent exiting decision is reached using energy scores and residual scores separately, this outcome stands as the final decision; otherwise, the decision-making process advances to the second stage. In the second stage, we introduce a novel exiting signal through feature-level integration, that is the Uncertainty-Aware Probability (UAP) score. This score ensures a more accurate estimation of prediction uncertainty for boundary region samples, thereby enhancing the reliability of exiting decisions for those samples.

3.3 | Principal Space and Residual Score

Principal component analysis (PCA) is a statistical technique used to reduce the dimensionality of data while retaining critical information. Inspired by PCA, the principal space P is defined as the subspace generated by the first D principal components of the training data, aiming to capture the major patterns of the sample population distribution. Intuitively, samples that deviate from the population distribution are more likely to be OOD

samples, typically resulting in higher prediction uncertainty. Correspondingly, we devise a residual score that captures the deviation of features from the principal space to indicate the possibility of a sample being OOD, thereby providing prediction uncertainty estimation from a global perspective. The computation of the residual score is detailed as follows.

Firstly, we compute the covariance matrix of the training set:

$$\mathbf{\Sigma} = \frac{1}{N-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}}), \tag{1}$$

where N denotes the number of training samples, $\boldsymbol{X} \in \mathbb{R}^{N \times H}$ denotes the training data whose rows are H-dimensional features, and $\overline{\boldsymbol{X}}$ denotes the mean of sample features. Next, we perform orthogonal diagonalisation on $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \boldsymbol{Q}\mathbf{\Lambda}\boldsymbol{Q}^{-1}, \tag{2}$$

where the eigenvalues in $\mathbf{\Lambda}$ are sorted decreasingly, and the eigenvectors in $\boldsymbol{Q}$ form a set of standard orthogonal basis. Then, we define the principal space P as the D-dimensional subspace generated by the first D columns of $\boldsymbol{Q}$. Let $P^{\perp}$ denote the orthogonal complement space of P, which equals to the $(H - D)$-dimensional subspace generated by the last $H - D$ columns of $\boldsymbol{Q}$. Accordingly, the feature $\boldsymbol{x} \in \mathbb{R}^N$ can be orthogonally decomposed:

$$\boldsymbol{x} = \boldsymbol{x}^P + \boldsymbol{x}^{P^{\perp}}, \tag{3}$$

where $\boldsymbol{x}^P$ and $\boldsymbol{x}^{P^{\perp}}$ are the projections of $\boldsymbol{x}$ onto P and $P^{\perp}$, respectively. The component $\boldsymbol{x}^{P^{\perp}} = \boldsymbol{R}\boldsymbol{R}^T\boldsymbol{x}$ represents the reconstruction error in PCA, where $\boldsymbol{R} \in \mathbb{R}^{H \times (H-D)}$ consists of the last $H - D$ columns of $\boldsymbol{Q}$ in Equation (2). Finally, the residual score is defined as the L2 norm of $\boldsymbol{x}^{P^{\perp}}$:

$$\text{Residual}(\boldsymbol{x}) = \|\boldsymbol{x}^{P^{\perp}}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{x}}. \tag{4}$$

The residual score reflects the deviation of $\boldsymbol{x}$ from P, and a higher value indicates greater prediction uncertainty. In residual-based early exiting, the exiting condition is satisfied when the residual score drops below the predefined threshold. In contrast to current exiting signals that rely solely on local information from individual samples, the residual score provides a global perspective for estimating prediction uncertainty.

3.4 | Inference Stage

In this subsection, we propose a two-stage exiting strategy using three-way decisions to enhance the reliability of exiting decisions. This strategy integrates global information from residual scores with local information from energy scores at two levels: the decision level in the first stage and the feature level in the second stage.

Energy-based Early Exiting. We employ classical energy scores to provide a local sample-specific perspective for estimating prediction uncertainty.

E-LANG Akbari et al. [18] first applied the energy score to early exiting, demonstrating its superiority over entropy and softmax scores. The energy score is defined as follows:

$$\text{Energy}(\boldsymbol{x}) = -\log \sum_{c=1}^{C} e^{l_c}, \tag{5}$$

where C denotes the number of classes, and l_c denotes the logit value of sample $\boldsymbol{x}$ on the c-th class suggested by the internal classifier. A higher energy score indicates greater prediction uncertainty. Per Equation (5), the energy score is computed based on the logits of individual samples, offering a local perspective for estimating prediction uncertainty. For energy-based early exiting, the inference process is terminated once the energy score falls below the predefined threshold.

Stage I. In the first stage, the residual score and the energy score are used independently to make exiting decisions. Given a sample, there are three potential decision outcomes at each intermediate layer, corresponding to the positive, negative, and boundary regions in three-way decisions:

- Positive Region. If a consistent 'exit' decision is made using either the energy score or the residual score, the sample exits from the current layer.
- Negative Region. If a consistent 'continue' decision is made using either the energy score or the residual score, the sample continues to execute the next layer.
- Boundary Region. If inconsistent decisions arise from the energy score and the residual score, an additional second-stage decision process is introduced to enhance prediction uncertainty estimation for boundary region samples, thus enabling more reliable exiting decisions.

Stage II. We introduce a novel exiting signal for the second stage, called the Uncertainty-Aware Probability (UAP) score. This score integrates global and local information at the feature level to better handle boundary region samples. Specifically, to enable feature-level integration, we convert the residual score into a new logit by scaling and then output the softmax probability corresponding to the new logit as the UAP score. The new logit is calculated as below:

$$l_0 = \alpha\|\boldsymbol{x}^{P^{\perp}}\| = \alpha\sqrt{\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{x}} \tag{6}$$

is the scaled residual score where the scaling parameter α is a constant. Note that the residual score cannot be directly used as a new logit, as the subsequent softmax operation is highly sensitive to the logit scale. If the residual score is excessively large, it will dominate the final UAP score; conversely, if the residual score is too small, it will be buried by the noise in the original logits. To match the scales of the new logit and the original logits, we set the α value as the ratio of the means of the maximum original logit and the residual score:

$$\alpha = \frac{\sum_{k=1}^{K} \max_{c=1,\ldots,C}\{l_{k,c}\}}{\sum_{k=1}^{K} \|\boldsymbol{x}_k^{P^{\perp}}\|}, \tag{7}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_K$ are K samples uniformly sampled from the training set, and $l_{k,c}$ denotes the logit of sample $\boldsymbol{x}_k$ for the c-th class. In this way, the scale of the new logit matches that of the maximum original logit on average.

We add this new logit to the original logits. Unlike the original logits which indicate the similarity of features to each original class (local information), the new logit indicates the deviation of features from the principal space (global information). The UAP score is finally defined as the softmax probability corresponding to the new logit:

$$\mathrm{UAP}(\boldsymbol{x}) = \frac{e^{\alpha\sqrt{\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{x}}}}{\sum_{c=1}^{C} e^{l_c} + e^{\alpha\sqrt{\boldsymbol{x}^T\boldsymbol{R}\boldsymbol{R}^T\boldsymbol{x}}}}, \tag{8}$$

where l_c denotes the original logit of sample $\boldsymbol{x}$ for the c-th class. The UAP score lies in (0, 1). Its value represents the probability of the sample belonging to the constructed virtual Out-Of-Distribution (OOD) class, indicating the model's inability to provide high-certainty predictions. Hence, the UAP score can serve as a proxy for prediction uncertainty in the second stage, and the exiting criterion is met once its value falls below a predefined threshold.

To analyse the information sources of UAP, we have its equivalent expression by applying a monotonic increasing function $f(x) = -\ln\left(\frac{1}{x} - 1\right)$:

$$\alpha\|\boldsymbol{x}^{P^\perp}\| - \ln\left(\sum_{c=1}^{C} e^{l_c}\right), \tag{9}$$

where the first term is the scaled residual score and the second term is the energy score, both of which are highly correlated with prediction uncertainty. We notice that samples with higher residual and energy scores exhibit elevated UAP scores, indicating increased prediction uncertainty. Obviously, the proposed UAP score integrates global information from residual scores with local information from energy scores at the feature level by adding the scaled residual score as a new logit. This facilitates prediction uncertainty estimation for boundary region samples, enabling more reliable exiting decisions.

3.5 | Training Stage

Following previous studies [10, 11, 29], the training objective of our method is formulated as the weighted sum of cross-entropy losses across all classifiers:

$$L = \frac{\sum_{m=1}^{M} m \times L^{(m)}}{\sum_{m=1}^{M} m}, \tag{10}$$

where $L^{(m)}$ denotes the cross-entropy loss at the m-th classifier. Considering that shallow-layer parameters receive more supervision signals from internal classifiers than deep-layer parameters, the loss weight assigned to each classifier is proportional to its layer number. This weighting strategy balances the parameter updates across shallow and deep layers. Importantly, internal classifiers are jointly trained with the backbone model without parameter sharing across layers.

4 | Experimental Results and Analysis

In this section, we evaluate our method on six classification tasks in the GLUE benchmark [42] using BERT-base Devlin et al. [1] as the backbone model. We first briefly introduce the datasets, followed by a description of the baseline methods and experimental settings. The experimental results and analysis are at the end.

4.1 | Tasks and Datasets

We conduct experiments on six classification tasks in the GLUE benchmark [42], including SST-2, MRPC, QNLI, RTE, QQP, and MNLI. We exclude the STS-B task since it is a regression task. We also exclude the WNLI and CoLA tasks following previous studies [9, 11, 12, 14, 19, 29]. The dataset statistics are provided in Table 1.

4.2 | Baselines

We select three groups of representative and state-of-the-art baselines for convincing comparisons.

Backbone. Firstly, we adopt the widely used pre-trained model BERT-base [1] as the backbone, offering a performance reference with a 1.00× speed-up ratio for each task.

Budget Exiting. We choose BERT-6L, where the first 6 layers of the BERT-base and a classifier are jointly fine-tuned to provide the final predictions. BERT-6L achieves static model acceleration with a 2.00× speed-up ratio, setting a lower bound for early exiting models as no specific exiting strategies are employed.

Early Exiting. We choose DeeBERT [9], Right-Tool Schwartz et al. [17], E-LANG Akbari et al. [18], PABEE [10], F-PABEE Gao et al. [20], DE^3-BERT He et al. [15] and PCEE-BERT Z. Zhang et al. [25], which encompass all existing heuristic exiting strategies, including score-based strategies, patience-based strategies, and hybrid strategies. To validate the effectiveness of integrating local and global information, we also present the

TABLE 1 | Dataset statistics.

Dataset	Classes	|Train|	|Dev|	|Test|	Task
SST-2	2	67k	0.9k	1.8k	Sentiment
MRPC	2	3.7k	0.4k	1.7k	Paraphrase
QQP	2	364k	40k	391k	Paraphrase
MNLI	3	393k	20k	20k	NLI
QNLI	2	105k	5.5k	5.4k	QA/NLI
RTE	2	2.5k	0.3k	3k	NLI

Abbreviations: QA, Question Answering task; NLI, Natural Language Inference task.

experimental results of Residual, which relies solely on the residual score for exiting decision-making.

For fair comparisons, the methods employing learning-based exiting strategies Xin et al. [12]; Balagansky and Gavrilov [13]; Zeng et al. [14]; Mangrulkar et al. [27]; Sun et al. [26] are not included since they require additional parameters and training costs in formulating their exiting criteria. In contrast, our method employs a heuristic exiting strategy for direct exiting signal computation, facilitating implementation but also constraining the model acceleration performance. Besides, early exiting methods that focus on improving training objectives or network architectures Li et al. [19]; Ji et al. [28]; Liao et al. [11]; Zhu [29]; Zhu et al. [30] are also excluded, as our method primarily focuses on enhancing exiting strategies.

4.3 | Experimental Settings

Speed Measurement. Since the actual inference time is unstable across different runs, following previous studies Zhang et al. [25]; Liao et al. [11], the speed-up ratio for inference is measured as the ratio of the total number of layers in the model to the layers actually executed during forward propagation:

$$\text{Speed-up Ratio} = \frac{\sum_{m=1}^{M} M \times N^m}{\sum_{m=1}^{M} m \times N^m}, \tag{11}$$

where M denotes the total number of layers in the model and N^m denotes the number of samples exiting from the m-th layer. This metric has been demonstrated approximately proportional to the actual inference time (see Section 4.5).

Training. Our implementation is based on Hugging Face's Transformers library [43]. We add a single-layer fully connected network as the internal classifier at each intermediate layer of the PLM. All internal classifiers are jointly trained with the backbone model. Following Zhou et al. [10] and Zhang et al. [25], we perform a grid search over learning rates of {1e-5, 2e-5, 3e-5, 5e-5}, and batch sizes of {16, 32, 128}. The maximum sequence length is set to 128. We employ a linear decay learning rate scheduler and the AdamW optimiser [44]. All experiments are performed on two RTX4090 GPUs with 24GB.

Inference. Following prior research [20, 25], we set the batch size to 1 for the inference stage. This setting emulates a common industry situation where requests from various users are received sequentially. We set the number of principal components D for each intermediate layer as the minimum value that ensures the cumulative variance contribution rate exceeds 85%. This threshold is widely accepted in the field of PCA Gerst [45]; Zhao and Zhang [46], as it effectively retains the main patterns of the data distribution while minimising noise and redundant information. We set K in Equation (7) to be half of the total number of samples in the training set. For fair comparisons, we manually adjust the thresholds for each task to attain a similar speed-up ratio to the baseline methods (approximately 2.00×), and further compare their task performance.

4.4 | Overall Performance Comparison

In this subsection, we compare the performance-efficiency trade-off of our method with those of the baseline methods.

Table 2 presents the test results of each early exiting method on the GLUE benchmark using BERT-base as the backbone model. The speed-up ratio is approximately 2.00×. Overall, our method consistently outperforms all baseline methods across all tasks, achieving an average speed-up ratio of 2.17× with minimal performance degradation compared to the backbone model. It also surpasses the state-of-the-art baseline E-LANG by 11% in inference speed while simultaneously delivering an average performance improvement of 0.6 points. These observations demonstrate the effectiveness of integrating local and global information in facilitating inference efficiency while maintaining task performance. Notably, our method significantly outperforms the similar DE^3-BERT on most tasks, further confirming the superiority of the proposed residual score and

TABLE 2 | Test results on the GLUE benchmark using BERT-base as the backbone model.

Method	RTE	MRPC	QQP	SST-2	QNLI	MNLI	AVG
BERT-base‡	66.4 (1.00×)	88.9 (1.00×)	71.2 (1.00×)	93.5 (1.00×)	90.5 (1.00×)	84.6 (1.00×)	82.5 (1.00×)
BERT-×	63.9 (2.00×)	85.1 (2.00×)	69.7 (2.00×)	91.0 (2.00×)	86.7 (2.00×)	80.8 (2.00×)	79.5 (2.00×)
DeeBERT†	64.3 (1.95×)	84.4 (2.07×)	70.4 (2.13×)	90.2 (2.00×)	85.6 (2.09×)	74.4 (1.87×)	78.2 (2.02×)
Right-tool	64.6 (1.92×)	84.2 (2.04×)	70.5 (2.04×)	89.3 (1.92×)	86.2 (1.96×)	77.6 (2.04×)	78.7 (1.99×)
PABEE†	64.0 (1.81×)	84.4 (2.01×)	70.4 (2.09×)	89.3 (1.95×)	88.0 (1.87×)	79.8 (2.07×)	79.3 (1.97×)
PCEE-BERT	67.1 (1.89×)	86.4 (2.13×)	70.9 (1.96×)	92.3 (1.92×)	88.8 (2.17×)	82.2 (1.80×)	81.3 (1.98×)
E-LANG	67.2 (1.96×)	87.0 (1.98×)	71.0 (1.89×)	92.2 (2.05×)	89.6 (1.89×)	83.0 (1.96×)	81.7 (1.96×)
F-PABEE	67.3 (1.85×)	87.5 (2.16×)	70.7 (1.92×)	92.3 (1.96×)	89.2 (2.14×)	82.2 (2.08×)	81.5 (2.02×)
DE^3-	65.7 (1.99×)	86.6 (1.98×)	**71.2 (2.16×)**	92.5 (2.02×)	90.0 (2.07×)	83.2 (2.04×)	81.5 (2.04×)
Residual	67.0 (1.93×)	87.1 (2.18×)	70.6 (2.07×)	92.2 (2.04×)	88.9 (1.98×)	82.5 (1.92×)	81.4 (2.02×)
Ours	**68.2 (1.95×)**	**88.2 (2.59×)**	71.0 (2.05×)	**92.9 (2.14×)**	**90.2 (2.35×)**	**83.3 (1.96×)**	**82.3 (2.17×)**

Note: † and ‡ denote the baseline results taken from GPFEE [11] and the original papers, respectively. Other baseline results are based on our implementation. We report F1-score for MRPC and QQP and accuracy for other tasks. The best results are marked in bold.

dual-level information integration strategy. Interestingly, we observe that our method achieves more significant performance improvements over the baseline methods on small datasets. For instance, on the small datasets RTE and MRPC, our method surpasses the competitive E-LANG by 1.0 and 1.2 points, respectively, while improvements on the remaining larger datasets are limited to a maximum of 0.7 points. We attribute this to the model's overfitting on small datasets, which results in an excessively optimistic estimation of prediction uncertainty when relying solely on the model outputs for individual samples, potentially undermining the reliability of exiting decisions. In this scenario, leveraging prior knowledge from the sample population indicated by residual scores is particularly crucial for accurately estimating prediction uncertainty. Hence, the performance gains from incorporating residual scores become increasingly significant. Additionally, the performance of our method even outperforms that of the original BERT-base on the RTE task. This suggests that our method effectively mitigates the overthinking problems in PLMs by allowing easy samples to exit at appropriate intermediate layers. It enables accurate predictions using high-quality shallow-layer representations rather than excessively complicated deep-layer representations, enhancing task performance while minimising redundant computations.

We also compare the performance-efficiency trade-off curves of our method and three competitive baseline methods using the same fine-tuned multi-exit PLM. Figure 4 shows the experimental results on a representative subset of GLUE. Each point on the curve corresponds to a set of selected thresholds, whose horizontal and vertical coordinates represent the corresponding speed-up ratio and task performance. Our method consistently outperforms the baseline methods that rely solely on local information across nearly all tasks and speed-up ratios. Notably, as the speed-up ratio increases, our method exhibits a more gradual decline in task performance compared to baselines, demonstrating its superiority in high acceleration scenarios. This effectiveness stems from our integration of local and global information at both the decision and feature levels, which enhances the reliability of exiting decisions and facilitates a better performance-efficiency trade-off.

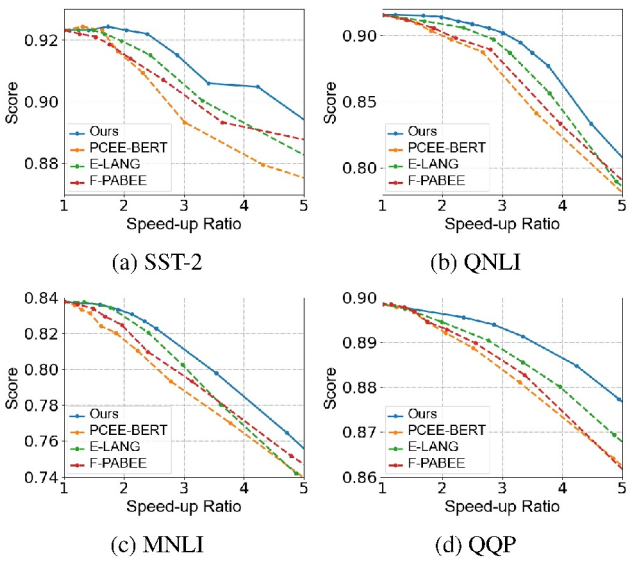


FIGURE 4 | Performance-efficiency trade-offs using different early exiting methods on four GLUE development sets.

4.5 | In-Depth Analysis

We conduct an in-depth analysis to illustrate the working mechanism of our method, further validating the interpretability of our proposed dual-level information integration strategy. To ensure consistent and convincing results, we conduct experiments on a representative subset of GLUE, including SST-2, QNLI, MNLI, and QQP.

Ablation Studies. As mentioned in Section 3.4, our method integrates global information from residual scores with local information from energy scores at both the decision and feature levels to enable reliable exiting decisions. To explore the contribution of local and global information in exiting decision-making, and validate the necessity of conducting dual-level information integration, we conduct ablation studies on local information, global information and decision-level information integration, respectively.

Figure 5 shows the performance-efficiency trade-off curves using different exiting strategies on a representative subset of GLUE. For each task, the experimental results are based on the same model trained as illustrated in Section 3.5. According to Figure 5, our two-stage hybrid exiting strategy and the one-stage hybrid exiting strategy UAP consistently outperform any single strategy using either residual scores or energy scores across all tasks. This observation confirms the effectiveness of integrating global information from residual scores with local information from energy scores, indicating that these two types of

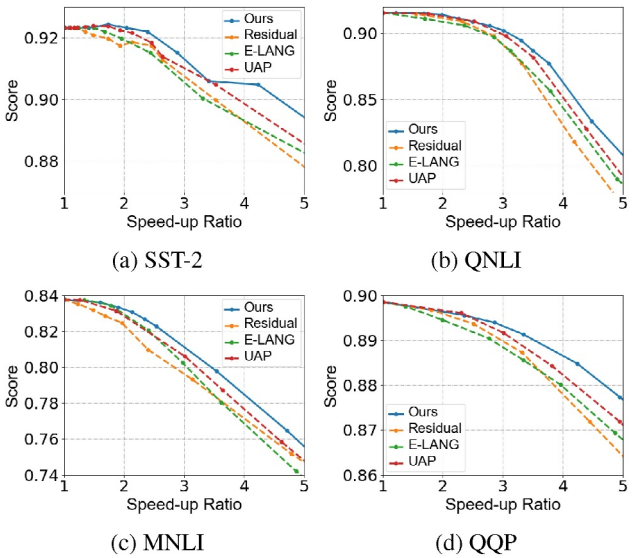


FIGURE 5 | Performance-efficiency trade-offs using different exiting strategies on four GLUE development sets. Ours denotes the proposed two-stage exiting strategy, while Residual, E-LANG and UAP represent one-stage exiting strategies based on the residual score, energy score and UAP score, respectively.

information can mutually correct each other to enhance the reliability of exiting decisions.

Furthermore, the proposed two-stage hybrid exiting strategy demonstrates a superior trade-off between performance and efficiency compared to the one-stage hybrid strategy UAP, which integrates local and global information solely at the feature level. We attribute this to the following reasons. In contrast to feature-level information integration, decision-level information integration can avoid information loss and preserve the integrity of each information source, providing diverse perspectives for exiting decision-making. Hence, the effect of decision-level information integration cannot be replaced by feature-level integration. This demonstrates the necessity of integrating local and global information at both the decision and feature levels, which facilitates their mutual correction and further enhances the reliability of exiting decisions.

DIS Analysis for Exiting Signals. Difficulty Inversion Score (DIS) was first introduced by CascadeBERT Li et al. [19] to evaluate the capability of exiting signals in distinguishing easy samples from hard samples. A higher DIS indicates greater consistency between the exiting signal and sample difficulty, which facilitates prediction uncertainty estimation and leads to more reliable exiting decisions.

To investigate the impact of integrating local and global information on the discriminative capability of exiting signals, we compare the DIS of various exiting signals on the development sets of SST-2 and QNLI, as shown in Table 3. The experimental results are based on the outputs of layers 2, 6 and 10. We see that, compared to exiting signals utilising a single information source, the proposed UAP score exhibits a higher DIS in nearly all cases. This suggests that, by integrating global information from residual scores with local information from energy scores at the feature level, the proposed UAP score demonstrates a stronger discriminative capability across various layers and tasks, which is crucial for making reliable exiting decisions.

Statistics of Exiting Decisions. We adopt two types of error rates, that is *Premature Exiting Rate* and *Delayed Exiting Rate*, to evaluate the reliability of exiting decisions.

- The *Premature Exiting Rate* is defined as the percentage of 'exit' decisions made under the condition that incorrect early predictions are provided by internal classifiers.

TABLE 3 | DIS analysis for each exiting signal at different layers on the development sets of SST-2 and QNLI.

	SST-2			QNLI		
Method	$L = 2$	$L = 6$	$L = 10$	$L = 2$	$L = 6$	$L = 10$
PCEE-BERT	66.3	73.2	75.2	54.5	65.9	70.7
F-PABEE	71.2	78.3	81.0	55.6	68.3	71.6
E-LANG	72.5	78.8	**82.8**	57.3	75.1	76.0
Residual	71.8	80.5	80.9	56.2	77.9	77.2
UAP	**75.6**	**82.2**	81.0	**59.4**	**79.8**	79.1

Note: The best results are marked in bold.

- The *Delayed Exiting Rate* is defined as the percentage of 'continue' decisions made under the condition that correct early predictions are provided by internal classifiers.

On one hand, we employ the Premature Exiting Rate to indicate the model's tendency to emit samples prematurely with incorrect early predictions, which may hinder the task performance of early exiting models. On the other hand, we utilise the Delayed Exiting Rate to reflect the model's tendency to delay the exiting of samples with correct early predictions, which may cause redundant computations and prolong the inference time. These two error rates offer insights into the reliability of exiting decisions from different perspectives. The exiting decisions are considered sufficiently reliable only when both the Premature Exiting Rate and the Delayed Exiting Rate are adequately low.

Figure 6 presents the two error rates for exiting decisions across various early exiting methods. We observe that methods solely using global information consistently exhibit lower Premature Exiting Rates and higher Delayed Exiting Rates compared to those solely using local information in nearly all cases. This suggests that the singleness of information sources limits the reliability of exiting decisions: solely relying on local information overlooks the global insights from the sample population,

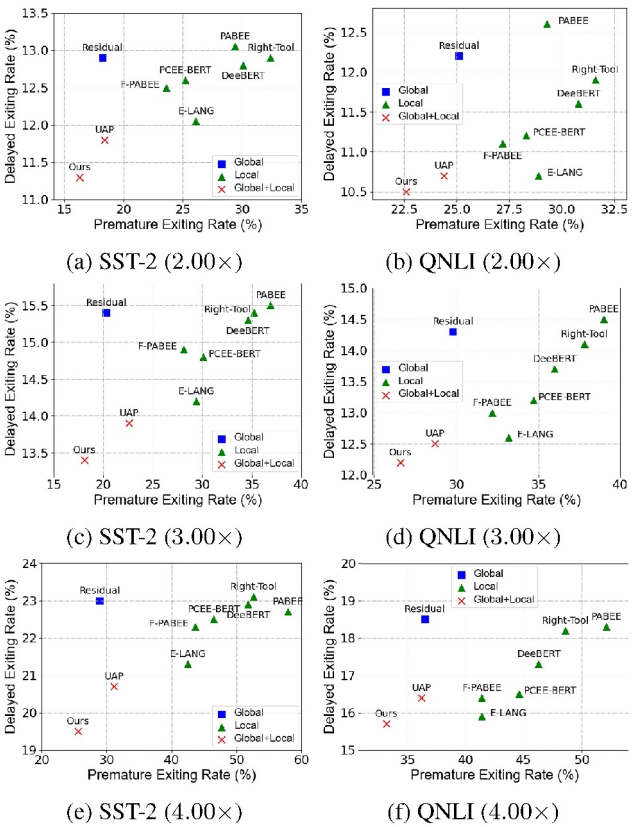
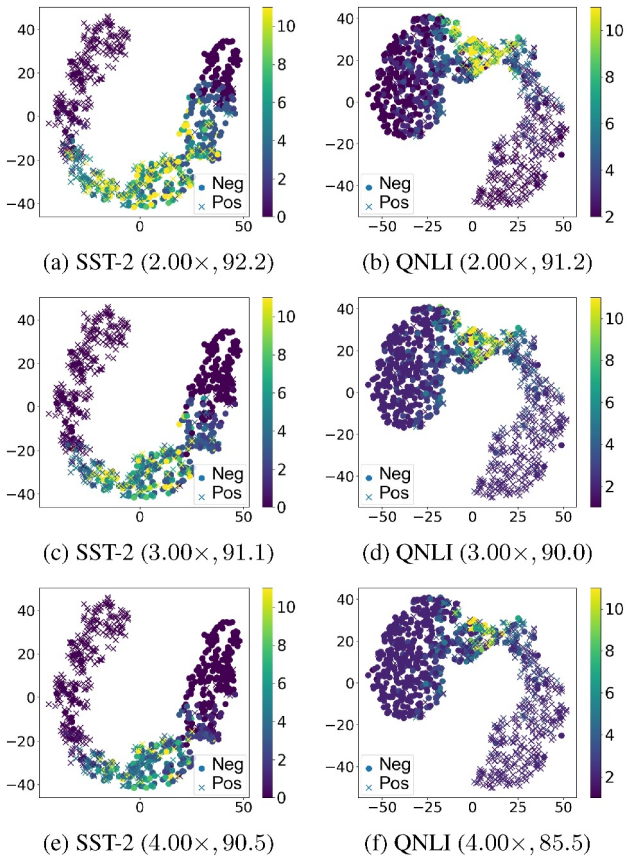


FIGURE 6 | Error rate statistics for exiting decisions across various early exiting methods. Results are reported on the development sets of SST-2 and QNLI at speed-up ratios of 2.00×, 3.00× and 4.00×. We collect the prediction results and exiting decisions for each sample across all executed layers, meaning that each sample provides a sampling for every layer it processed. The Premature and Delayed Exiting Rates are then calculated based on this data.

exacerbating the premature exiting of samples with incorrect early predictions, while exclusive use of global information disregards individual sample-specific details, which exacerbates the delayed exiting of samples with correct early predictions. In contrast, our method effectively reduces both the Premature Exiting Rate and the Delayed Exiting Rate by jointly considering local and global information, facilitating the reliability of exiting decisions.

Additionally, it is noteworthy that both error rates in our method are significantly lower than those in UAP, which adopts a one-stage exiting strategy integrating local and global information only at the feature level. This finding is consistent with the observations shown in Figure 5. We attribute this to the decision-level information integration, which effectively preserves the advantages of individual information sources and facilitates their mutual correction. This demonstrates the necessity of integrating local and global information at both the decision and feature levels to enhance the reliability of exiting decisions.

Visualisation of Sample Exiting Layers. To explore the distribution of exiting layers in the feature space, we visualise the sample's exiting layers using t-SNE projection van der Maaten and Hinton [47] based on the outputs of the sixth layer. Figure 7 shows the visualisation results and the corresponding task



(a) SST-2 (2.00×, 92.2) (b) QNLI (2.00×, 91.2)

(c) SST-2 (3.00×, 91.1) (d) QNLI (3.00×, 90.0)

(e) SST-2 (4.00×, 90.5) (f) QNLI (4.00×, 85.5)

FIGURE 7 | Exiting layer distribution on the development sets of SST-2 and QNLI under various speed-up ratios. The corresponding task performance is provided in parentheses. Neg and Pos denote negative and positive samples, respectively. The colour represents the sample's exiting layer.

performance under various speed-up ratios for the SST-2 and QNLI tasks. Each point represents a sample, with its colour indicating the corresponding exiting layer. It is intuitive that samples far from the classification boundary are generally considered easy, while those near the boundary are often deemed more challenging. As shown in Figure 7, easy samples tend to exit at shallow layers, whereas hard samples are more likely to exit at deeper layers. Additionally, as the speed-up ratio continues to increase, the exiting layer of samples gradually decreases, leading to improved inference efficiency but also causing a certain level of performance degradation. These observations are consistent with our intuition, further validating the effectiveness of our method.

Computational and Storage Costs. Table 4 shows the computational complexity of each module in our model. Note that the computational complexity introduced by $\boldsymbol{R}$ and α in Equation (6) are excluded, as they are computed only once before the inference stage, and their results can be shared across all samples during inference. We observe that the computational overhead incurred by making exiting decisions is only 1.2M per layer for each sample, which is negligible compared to the 1813.5M of an encoder block. This confirms that the exiting decision-making process is computationally efficient and will not impact the model's inference speed. Moreover, since the model's computational costs are primarily driven by the encoder blocks, they are approximately proportional to the number of layers executed. This further validates the use of saved layers to measure model acceleration as shown in Equation (11).

Table 5 compares the parameter volumes of our early exiting model with those of the original backbone BERT-base. We find that for a task involving 2 (or 3) classes, our method introduces only 16.92K (or 25.38K) parameters by incorporating an internal classifier at each intermediate layer of the backbone model, leading to a minimal 0.015% (or 0.023%) increase in the model's storage costs. Additionally, the exiting decision-making module in this paper is parameter-free, further demonstrating the storage efficiency of our method.

TABLE 4 | Computational complexity of each module in our model.

	FLOPs	
Module	$\boldsymbol{C} = 2$	$\boldsymbol{C} = 3$
Embedding	786.4K	786.4K
Encoder	1813.5M	1813.5M
Pooler	1.2M	1.2M
Classifier	3.1K	4.6K
Exiting Decision-Making*	1.2M	1.2M

Note: C denotes the number of classes. The $*$ indicates the module introduced by our method.

TABLE 5 | Comparison of parameter volumes.

	#Params	
Model	$\boldsymbol{C} = 2$	$\boldsymbol{C} = 3$
BERT-base	109.48M	109.48M
Ours	+16.92K	+25.38K

Based on the above analysis, our method is efficient in terms of both storage and computation.

Analysis of Speed Measurements. To investigate whether the speed measurement based on saved layers in Equation (11) can accurately reflect model acceleration, we collect the total number of executed layers and inference time for all samples across various UAP thresholds on the SST-2 development set as shown in Table 6. We observe that the model's actual inference time is approximately proportional to the total number of executed layers, with a Pearson correlation coefficient of up to 0.99 between them. This aligns with our observation in Table 4 that the model's computational overhead during inference is approximately proportional to the number of layers executed. These observations confirm that the speed measurement in Equation (11) can serve as a proxy for the model's actual runtime or computational costs, accurately reflecting model acceleration. Given that the actual inference time is typically unstable across different runs, we use saved layers as the metric for model acceleration in this paper.

Analysis of Cross-Domain Generalisation. To validate the generalisation capability of our method on cross-domain data, we establish the source and target datasets that share the same labels but have different distributions. We train the proposed model using the source dataset and then test it on the target dataset. Preciously, we choose the IMDb and SNLI datasets as the targets, while the SST-2 and MNLI datasets serve as their corresponding sources. The IMDb and SST-2 are sentiment classification tasks, containing film reviews labelled as *positive* or *negative*. The SNLI and MNLI are natural language inference tasks, comprising sentence pairs labelled as *entailment*, *contradiction* or *neutral* regarding their logical relationships.

TABLE 6 | Total executed layers and inference time for all samples in the SST-2 development set across various thresholds for the UAP score.

Threshold	Executed layers	Inference time (s)
0.00	10,464	16.95
0.02	9560	16.57
0.04	7051	15.13
0.06	5819	13.45
0.08	4858	12.86
0.10	4146	12.43
0.13	3475	11.71
0.16	2896	11.62
0.20	2296	10.60
0.23	1906	9.70
0.26	1651	9.55
0.30	1350	9.29
0.32	1174	9.67
0.34	1002	9.75
0.36	885	8.82
0.38	872	9.12
0.40	872	9.12
1.00	872	9.11

In Table 7, we report the experimental results on the target datasets for each early exiting method. We choose the state-of-the-art cross-domain early exiting method CeeBERT Bajpai and Hanawal [48] as the baseline. CeeBERT employs a score-based exiting strategy, with the softmax score as its exiting signal. It introduces an unsupervised online learning algorithm to determine the optimal thresholds for the target datasets, thus facilitating efficient unsupervised cross-domain inference for PLMs.According to Table 7, our method exhibits slightly inferior performance on the target datasets compared to the baseline CeeBERT.This is attributed to the fact that our method relies on the prior knowledge acquired from the training set, specifically the principal space spanned by the first D principal components of the training data. While this prior knowledge can enhance the reliability of exiting decisions for in-domain data by providing a global perspective, its benefits may be diminished in cross-domain scenarios due to distributional differences between training and test data. These distribution differences can even affect the reliability of exiting decisions, leading to sub-optimal performance-efficiency trade-offs of PLMs. This issue warrants further research and exploration.

Analysis of Statistical Significance. To evaluate the statistical significance of our method's improvements over the baseline methods in low acceleration scenarios, we conduct one-sided t-tests. Specifically, we collect experimental results from models trained with 5 different seeds for both our method and the baselines under speed-up ratios of 1.30×, 1.60× and 1.90×, respectively. We then calculate the mean and standard deviation of the performance across different models for each method and speed-up ratio as shown in Table 8. The one-tailed t-tests at a significance level of 0.05 further confirm the statistically significant improvements of our method over the baselines, providing strong evidence for the effectiveness and superiority of our proposed method in low acceleration scenarios.

Generality on Other Backbones. To explore the generalisation capability of our method across different backbones, we conduct experiments with ALBERT Lan et al. [2], which is a more lightweight and efficient variant of BERT. In Table 9, we report the test results of our method compared to the competing baselines on a representative subset of GLUE using the ALBERT-base backbone. The results suggest that our method consistently outperforms the baseline methods by a clear margin on all tasks, demonstrating its strong generalisation capability across various PLMs.

TABLE 7 | Experimental results on the target datasets for each method.

Method	SST-2_IMDb	MNLI_SNLI
CeeBERT	81.0 (2.95×)	79.4 (2.63×)
Ours	−0.3 (2.94×)	−0.4 (2.61×)

Note: The datasets are formatted as (source_target).

TABLE 8 | The mean and standard deviation of the performance for each method over 5 runs at speed-up ratios of 1.30×, 1.60× and 1.90×.

	Method	SST-2	QNLI	MNLI	QQP
∼1.30×	PCEE-BERT	92.3 ± 0.10	91.3 ± 0.12	83.3 ± 0.05	89.7 ± 0.06
	F-PABEE	92.2 ± 0.13	91.3 ± 0.13	83.6 ± 0.07	89.7 ± 0.09
	E-LANG	92.3 ± 0.10	91.4 ± 0.11	83.7 ± 0.06	89.8 ± 0.04
	Ours	**92.5 ± 0.11**	**91.6 ± 0.10**	**83.8 ± 0.05**	**89.9 ± 0.05**
∼1.60×	PCEE-BERT	92.2 ± 0.10	90.9 ± 0.14	82.5 ± 0.05	89.6 ± 0.05
	F-PABEE	92.0 ± 0.14	90.9 ± 0.15	83.1 ± 0.08	89.6 ± 0.06
	E-LANG	92.1 ± 0.13	91.2 ± 0.14	83.4 ± 0.06	89.6 ± 0.05
	Ours	**92.4 ± 0.11**	**91.5 ± 0.13**	**83.6 ± 0.07**	**89.7 ± 0.04**
∼1.90×	PCEE-BERT	91.6 ± 0.15	90.2 ± 0.12	81.9 ± 0.06	89.3 ± 0.06
	F-PABEE	91.7 ± 0.13	90.5 ± 0.16	82.6 ± 0.06	89.4 ± 0.05
	E-LANG	92.0 ± 0.16	90.9 ± 0.14	83.1 ± 0.05	89.5 ± 0.07
	Ours	**92.4 ± 0.15**	**91.4 ± 0.14**	**83.3 ± 0.04**	**89.7 ± 0.06**

Note: The best results are marked in bold.

TABLE 9 | Test results for each early exiting method using ALBERT as the backbone at a 2.00× (±5%) speed-up ratio.

Method	QQP	SST-2	QNLI	MNLI	AVG
ALBERT-base	70.5	93.3	92.0	85.2	85.3
PCEE-BERT	69.8	92.3	90.7	83.9	84.2
F-PABEE	69.9	92.4	90.9	84.0	84.3
E-LANG	70.1	92.3	91.1	84.5	84.5
Ours	**70.4**	**93.0**	**91.8**	**84.7**	**85.0**

Note: We report the F1-score for QQP and accuracy for other tasks. The best results are marked in bold.

5 | Conclusion

In this paper, we propose a novel two-stage early exiting method using PCA and three-way decisions. Our method effectively enhances the reliability of exiting decisions by integrating global information from residual scores with local information from energy scores at both the decision and feature levels, yielding a better trade-off between task performance and inference efficiency for PLMs.Our method is simple yet effective. Extensive experiments on the GLUE benchmark validate the superiority, generality and interpretability of our method, with negligible additional computational or storage costs.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376198), the National Key Research and Development Program of China (No. 2022YFB3104700) and the Shanghai Baiyulan Pujiang Project (No. 08002360429).

Conflicts of Interest

Duoqian Miao is an editorial board member for the journal, and was not involved in peer review process or the decision to publish this article. The authors declare that they have no conflict of interest.

Data Availability Statement

The data that supports the findings of this study is openly available in GLUE-baselines at https://github.com/nyu-mll/GLUE-baselines.git.

References

1. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT (1)* (Association for Computational Linguistics, 2019), 4171–4186.

2. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," in *ICLR* (OpenReview.net, 2020).

3. A. Radford, J. Wu, R. Child, et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI blog* 1, no. 8 (2019): 9.

4. Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR* (2019): abs/1907.11692.

5. Y. Huang, Z. Li, W. Deng, G. Wang, and Z. Lin, "D-BERT: Incorporating Dependency-Based Attention Into BERT for Relation Extraction," *CAAI Transactions on Intelligence Technology* 6, no. 4 (2021): 417–425, https://doi.org/10.1049/cit2.12033.

6. B. Zhou, J. Chen, Q. Cai, Y. Xue, C. Yang, and J. He, "Cross-Domain Sequence Labelling Using Language Modelling and Parameter Generating," *CAAI Transactions on Intelligence Technology* 7, no. 4 (2022): 710–720, https://doi.org/10.1049/cit2.12107.

7. X. Ma, Y. Liu, and C. Ouyang, "Capturing Semantic Features to Improve Chinese Event Detection," *CAAI Transactions on Intelligence Technology* 7, no. 2 (2022): 219–227, https://doi.org/10.1049/cit2.12062.

8. Y. Kaya, S. Hong, and T. Dumitras, "Shallow-Deep Networks: Understanding and Mitigating Network Overthinking," in *ICML, Volume 97 of Proceedings of Machine Learning Research* (PMLR, 2019), 3301–3310.

9. J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, "DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference," in *ACL* (Association for Computational Linguistics, 2020), 2246–2251.

10. W. Zhou, C. Xu, T. Ge, J. J. McAuley, K. Xu, and F. Wei, "BERT Loses Patience: Fast and Robust Inference With Early Exit," in *NeurIPS*, (2020).

11. K. Liao, Y. Zhang, X. Ren, Q. Su, X. Sun, and B. He, "A Global Past-Future Early Exit Method for Accelerating Inference of Pre-Trained Language Models," in *NAACL-HLT* (Association for Computational Linguistics, 2021), 2013–2023.

12. J. Xin, R. Tang, Y. Yu, and J. Lin, "BERxiT: Early Exiting for BERT With Better Fine-Tuning and Extension to Regression," in *EACL* (Association for Computational Linguistics, 2021), 91–104.

13. N. Balagansky and D. Gavrilov, "Palbert: Teaching Albert to Ponder," *Advances in Neural Information Processing Systems* 35 (2022): 14002–14012.

14. Z. Zeng, Y. Hong, H. Dai, H. Zhuang, and C. Chen, "ConsistentEE: A Consistent and Hardness-Guided Early Exiting Method for Accelerating Language Models Inference," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 17 (2024): 19506–19514, https://doi.org/10.1609/aaai.v38i17.29922.

15. J. He, Q. Zhang, W. Ding, et al., "DE3 - BERT: Distance-Enhanced Early Exiting for BERT Based on Prototypical Networks," *arXiv preprint arXiv:2402.05948* (2024).

16. J. He, Q. Zhang, H. Zhang, X. Huang, U. Naseem, and D. Miao, "COSEE: Consistency-Oriented Signal-Based Early Exiting via Calibrated Sample Weighting Mechanism," *arXiv preprint arXiv:2412.13236* (2024).

17. R. Schwartz, G. Stanovsky, S. Swayamdipta, J. Dodge, and N. A. Smith, "The Right Tool for the Job: Matching Model and Instance Complexities," in *ACL* (Association for Computational Linguistics, 2020), 6640–6651.

18. M. Akbari, A. Banitalebi-Dehkordi, and Y. Zhang, "E-Lang: Energy-Based Joint Inferencing of Super and Swift Language Models," *arXiv preprint arXiv:2203.00748* (2022): 5229–5244, https://doi.org/10.18653/v1/2022.acl-long.359.

19. L. Li, Y. Lin, D. Chen, et al., "CascadeBERT: Accelerating Inference of Pre-Trained Language Models via Calibrated Complete Models Cascade," in *EMNLP (Findings)* (Association for Computational Linguistics, 2021), 475–486.

20. X. Gao, W. Zhu, J. Gao, and C. Yin, "F-PABEE: Flexible-Patience-Based Early Exiting for Single-Label and Multi-Label Text Classification Tasks," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023), 1–5.

21. Y. Yao, "Three-Way Decision: An Interpretation of Rules in Rough Set Theory," in *Rough Sets and Knowledge Technology: 4th International Conference, RSKT 2009, Gold Coast, Australia, July 14-16, 2009. Proceedings 4* (Springer, 2009), 642–649.

22. W. Liu, P. Zhou, Z. Wang, Z. Zhao, H. Deng, and Q. Ju, "FastBERT: A Self-Distilling BERT With Adaptive Inference Time," in *ACL* (Association for Computational Linguistics, 2020), 6035–6044.

23. J. Zhang, M. Tan, P. Dai, and W. Zhu, "Leco: Improving Early Exiting via Learned Exits and Comparison-Based Exiting Mechanism," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, (2023), 298–309, https://doi.org/10.18653/v1/2023.acl-srw.43.

24. W. Zhu, P. Wang, Y. Ni, G. Xie, and X. Wang, "BADGE: Speeding up BERT Inference After Deployment via Block-Wise Bypasses and Divergence-Based Early Exiting," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, (2023), 500–509, https://doi.org/10.18653/v1/2023.acl-industry.48.

25. Z. Zhang, W. Zhu, J. Zhang, P. Wang, R. Jin, and T. Chung, "PCEE-BERT: Accelerating BERT Inference via Patient and Confident Early Exiting," in *NAACL-HLT (Findings)* (Association for Computational Linguistics, 2022), 327–338.

26. T. Sun, X. Liu, W. Zhu, et al., "A Simple Hash-Based Early Exiting Approach for Language Understanding and Generation," in *ACL (Findings)* (Association for Computational Linguistics, 2022), 2409–2421.

27. S. Mangrulkar, A. MS, and V. Sembium, "BE3R: BERT Based Early-Exit Using Expert Routing," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (2022), 3504–3512.

28. Y. Ji, J. Wang, J. Li, Q. Chen, W. Chen, and M. Zhang, "Early Exit With Disentangled Representation and Equiangular Tight Frame," in *Findings of the Association for Computational Linguistics: ACL 2023*, (2023), 14128–14142.

29. W. Zhu, "LeeBERT: Learned Early Exit for BERT With Cross-Level Optimization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (2021), 2968–2980, https://doi.org/10.18653/v1/2021.acl-long.231.

30. W. Zhu, X. Wang, Y. Ni, and G. Xie, "GAML-BERT: Improving BERT Early Exiting by Gradient Aligned Mutual Learning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (2021), 3033–3044.

31. Y. Yao, "Three-Way Decisions With Probabilistic Rough Sets," *Information Sciences* 180, no. 3 (2010): 341–353, https://doi.org/10.1016/j.ins.2009.09.021.

32. Y. Yao, "An Outline of a Theory of Three-Way Decisions," in *International Conference on Rough Sets and Current Trends in Computing* (Springer, 2012), 1–17.

33. N. Wang and P. Zhu, "A Three-Way Decision Model Associated With Decision-Theoretic Rough Fuzzy Sets Based on Computing With Words," *Journal of Intelligent and Fuzzy Systems* 45, no. 1 (2023): 285–304, https://doi.org/10.3233/jifs-224215.

34. H. Gou and X. Zhang, "Three-Level Models of Compromised Multi-Granularity Rough Sets Using Three-Way Decision," *Journal of Intelligent and Fuzzy Systems* 46, no. 3 (2024): 6053–6081, https://doi.org/10.3233/jifs-236063.

35. K. Yuan, D. Miao, W. Pedrycz, W. Ding, and H. Zhang, "Ze-HFS: Zentropy-Based Uncertainty Measure for Heterogeneous Feature Selection and Knowledge Discovery," *IEEE Transactions on Knowledge and Data Engineering* 36, no. 11 (2024): 7326–7339, https://doi.org/10.1109/tkde.2024.3419215.

36. K. Yuan, D. Miao, W. Pedrycz, H. Zhang, and L. Hu, "Multigranularity Data Analysis With Zentropy Uncertainty Measure for Efficient and Robust Feature Selection," *IEEE Transactions on Cybernetics* 55, no. 2 (2024): 740–752, https://doi.org/10.1109/tcyb.2024.3499952.

37. Y. Li, Y. Liu, H. Zhang, C. Zhao, Z. Wei, and D. Miao, "Occlusion-Aware Transformer With Second-Order Attention for Person Re-Identification," *IEEE Transactions on Image Processing* 33 (2024): 3200–3211, https://doi.org/10.1109/tip.2024.3393360.

38. Y. Li, D. Miao, H. Zhang, J. Zhou, and C. Zhao, "Multi-Granularity Cross Transformer Network for Person Re-Identification," *Pattern Recognition* 150 (2024): 110362, https://doi.org/10.1016/j.patcog.2024.110362.

39. Y. Wang, J. Pu, D. Miao, L. Zhang, L. Zhang, and X. Du, "SCGRFuse: An Infrared and Visible Image Fusion Network Based on Spatial/Channel Attention Mechanism and Gradient Aggregation Residual Dense Blocks," *Engineering Applications of Artificial Intelligence* 132 (2024): 107898, https://doi.org/10.1016/j.engappai.2024.107898.

40. M. Han, Y. Qu, N. MacParthaláin, C. Shang, Z. Yao, and Q. Shen, "Representation-Based Decision-Theoretic Rough Sets for Three-Way Classification," *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).

41. L. Guo, J. Zhan, C. Zhang, and Z. Xu, "A Large-Scale Group Decision-Making Method Fusing Three-Way Clustering and Regret Theory Under Fuzzy Preference Relations," *IEEE Transactions on Fuzzy Systems* 32, no. 9 (2023): 4846–4860, https://doi.org/10.1109/tfuzz.2023.3335965.

42. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *ICLR (Poster)* (OpenReview.net, 2019).

43. T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP (Demos)* (Association for Computational Linguistics, 2020), 38–45.

44. I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *ICLR (Poster)* (OpenReview.net, 2019).

45. M. D. Gerst, "Revisiting the Cumulative Grade-Tonnage Relationship for Major Copper Ore Types," *Economic Geology* 103, no. 3 (2008): 615–628, https://doi.org/10.2113/gsecongeo.103.3.615.

46. K. Zhao and J. Zhang, "Incipient Fault Diagnosis Based on Moving Window Cumulative Sum Principal Component Analysis and Reconstructed Contribution Plot," in *2024 36th Chinese Control and Decision Conference (CCDC)* (IEEE, 2024), 4349–4354.

47. L. van der Maaten and G. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *JMLR* 9, no. nov (2008): 2579–2605.

48. D. J. Bajpai and M. K. Hanawal, "CEEBERT: Cross-Domain Inference in Early Exit BERT," *arXiv preprint arXiv:2405.15039* (2024): 1736–1748, https://doi.org/10.18653/v1/2024.findings-acl.101.