# Faithfulness and Reasoning Drift in Compressed Transformers: A Comparative XAI Study of BERT, DistilBERT, and KD-TinyBERT on SST-5

Your Name

Department of Computer Science, Your University

2025

**Abstract**

Compressed transformer models often match the accuracy of their larger counterparts, but it is not obvious whether they reach the same conclusions for the right reasons. This paper examines this question by comparing BERT, DistilBERT, and a knowledge-distilled TinyBERT on fine-grained sentiment classification (SST-5). The study looks at different aspects of interpretability using SHAP as the main explainer, along with LIME for specific measures such as short-token fragmentation. The comparisons cover faithfulness, token-level agreement, directional consistency, and how the models handle contrastive constructions such as "A but B".

Across all metrics, DistilBERT remains closest to BERT in both output and reasoning patterns, even if it is slightly weaker overall. TinyBERT behaves differently: despite reaching accuracy similar to BERT, it shows clear signs of reasoning drift. Its directional similarity is substantially lower, and its token-level attributions often highlight neutral tokens instead of sentiment-bearing ones. Based on the contrastive probe, TinyBERT frequently arrives at the correct label but does so for the wrong reasons.

These results suggest that compression can preserve predictions while quietly altering the internal reasoning through which models reach their conclusions. The framework used here provides a practical way to detect this kind of drift and may help determine when lightweight transformer models remain trustworthy for interpretation and when they do not.

**Keywords:** Explainability, SHAP, LIME, Knowledge Distillation, BERT, DistilBERT, Sentiment Analysis

## 1 Introduction

Transformer-based language models such as BERT have become central to modern NLP [4], but their then computational cost has encouraged the development of smaller, faster variants through model compression and knowledge distillation [6]. Distilled models—including DistilBERT [21] and TinyBERT [8]—often come surprisingly close to the

original model's accuracy while using far fewer parameters, which makes them appealing for deployment in resource-constrained environments. What remains unclear, however, is whether these smaller models reach their decisions in the same way as the original model, or whether they simply learn to imitate its outputs without preserving the internal reasoning [15].

This question becomes important whenever reliability and interpretability matter(e.g. medical diagnosis and legal analysis). Methods like SHAP [13] and LIME [19] are widely used to analyse feature attributions, but most prior work applies them to individual models in isolation. Comparatively fewer studies investigate how different transformer models diverge in the evidence they rely on for the same prediction [2, 17]. Some recent work examines interpretability under compression or robustness [9, 27], yet there is still no systematic analysis of how token-level reasoning may *drift* when a model is compressed through architecture reduction (as in DistilBERT) or through knowledge distillation (as in TinyBERT). Likewise, research on sentiment analysis with SST-5 [22] typically focuses on accuracy, leaving open the question of whether token-level explanations remain faithful or stable after compression.

To explore this issue, we focus on the following questions: (1) How faithful are SHAP and LIME explanations for BERT, DistilBERT, and a knowledge-distilled TinyBERT model? (2) To what extent do these models agree in their explanations for the same input? (3) Does compression introduce directional changes in token-level attributions? (4) How do the models respond to contrastive edits, such as negation or clause reversal ("A but B")? (5) What trade-offs appear between interpretability and predictive performance when models are compressed?

**This paper makes three contributions.** First, it brings together several strands of explainability evaluation—faithfulness, token-level agreement, directional alignment, and sensitivity to contrastive constructions—to examine whether compressed transformer models preserve the reasoning patterns of their larger counterparts. Second, it offers a comparative analysis of BERT, DistilBERT, and a knowledge-distilled TinyBERT on SST-5, showing that models with similar accuracy can nevertheless rely on quite different evidence. Finally, through contrastive case studies, we show that TinyBERT often arrives at the correct label for structurally weak or context-insensitive reasons, highlighting potential risks when compressed models are used in settings where interpretability or reliability matters.

The rest of the paper is organized as follows. Section3 describes the dataset, models, and evaluation setup. Section5 discusses their implications and limitations. Section 6 concludes the study and outlines directions for future work.

# 2 Related Work

This work lies at the intersection of three areas: transformer-based sentiment analysis, model compression for NLP, and post-hoc explainability methods. Each has a substantial body of literature, but they rarely speak to each other. Below we summarize the most relevant strands and outline where the gaps remain.

—

## 2.1 Transformers for Fine-Grained Sentiment Analysis

The Stanford Sentiment Treebank (SST)[4] shifted the field decisively toward transformer encoders. Since then, SST-5 has often been used to probe how well pretrained transformers capture subtle sentiment distinctions such as "neutral" or mild polarity shifts.

A number of compact models—such as MobileBERT [23] and MiniLM [25]—have also been evaluated on SST-style benchmarks. These works mainly report accuracy and efficiency improvements, which is understandable, but they say relatively little about whether compressed models reason in the same way their larger counterparts do. This leaves open whether accuracy alone tells the full story.

## 2.2 Model Compression and Knowledge Distillation

The cost of running large transformers has motivated a broad line of work on compression. Distillation-based models such as DistilBERT [21], TinyBERT [8], and related variants [24, 25] typically match BERT's performance surprisingly well, especially when logits, hidden states, or attention maps are distilled jointly. Other compression strategies include pruning [10], matrix decomposition [16], and more aggressive low-rank or structured approximations [12].

Most compression work evaluates success using accuracy and efficiency, without explicitly examining whether the student model inherits the teacher's internal decision patterns. As a result, matching task performance is often treated as evidence that the underlying reasoning has also been preserved, even though this assumption is rarely tested directly.

## 2.3 Explainability and Faithfulness in NLP

LIME [19] and SHAP [13] remain the most widely used tools for post-hoc text explanations. They have been applied to many NLP tasks, including sentiment analysis, toxicity detection, and inference [1, 14]. A recurring finding in recent literature is that explanations can be plausible without being faithful [7, 26]—that is, they may look reasonable to a human reader while failing to reflect the model's true internal reasoning.

Researchers have proposed several metrics to assess faithfulness, such as token-removal tests, confidence drop, and agreement across explainers [2, 3]. Other work explores inherently interpretable architectures [11, 18], although these typically require specialized model designs rather than post-hoc analysis. Our study follows the latter tradition: we keep the models fixed and evaluate how standard explainers behave across them.

—

### 2.4 Explainability of Compressed Transformer Models

A smaller but growing body of work examines how compression interacts with explainability. Some papers analyze how pruning or low-rank approximations affect attention and attribution scores [27, 5]. Others look at the stability of rationales under parameter-efficient tuning [20]. However, these studies typically focus on binary classification or NLI, and they rarely examine fine-grained sentiment tasks like SST-5.

Moreover, existing work often evaluates explanations in isolation (e.g., "does pruning change attention distributions?") rather than within a broader framework that looks simultaneously at faithfulness, cross-model agreement, directional cue alignment, linguistic phenomena such as the contrastive but, and the trade-off between interpretability and performance.

To our knowledge, no prior study directly compares the reasoning behaviour of a full BERT model, a distilled architecture like DistilBERT, and a knowledge-distilled TinyBERT student on the same fine-grained sentiment task. The current literature tends to separate compression work from interpretability work; our goal is to examine what happens when these two worlds are combined and to characterize how small models may drift in the evidence they rely on even when their accuracy remains similar.

## 3 Methodology

### 3.1 Dataset

For our experiments we use the Stanford Sentiment Treebank (SST-5) [22], one of the standard benchmarks for fine-grained sentiment classification. Unlike binary sentiment tasks, SST-5 divides examples into five labels they are: very negative, negative, neutral, positive, and very positive—which makes the dataset more sensitive to subtle cues in tone, contrast, and emphasis. This fine-grained nature is useful for our purposes, because any drift in a model's reasoning tends to show up more clearly when distinctions become narrow.

Figure 1 shows the class distribution in the full SST-5 dataset. The original dataset leans slightly toward the neutral and negative classes, a common pattern in movie-review

corpora. Since this imbalance can complicate explanation analysis, we take an additional step during evaluation.

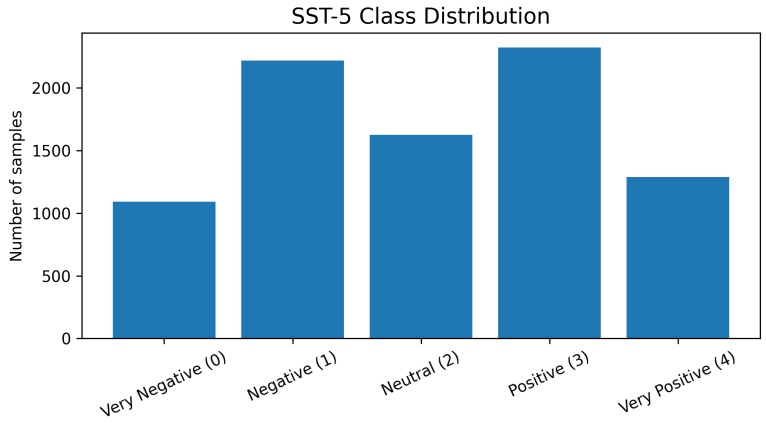


Figure 1: Label distribution in the SST-5 dataset.

To make comparisons across explainers and models more controlled, we construct a compact evaluation subset of 150 sentences from the SST-5 test split, with an equal number of samples (30 each) from all five sentiment classes. This balanced subset serves two practical purposes. First, post-hoc explainers such as SHAP and LIME are expensive to run, especially when applied across three different transformer models. Second, a fixed and balanced evaluation set allows us to isolate differences in model reasoning without worrying about skew from class imbalance.

Figure 2 shows few examples from this balanced 150-sentence evaluation set.

| | A | B | |
|---|---|---|---|
| 1 | text | gold | b |
| 2 | adam sandler is to gary cooper what a gnat is to a racehorse. | 0 | |
| 3 | these are names to remember, in order to avoid them in the future. | 0 | |
| 4 | but in its child - centered, claustrophobic context, it can be just as frightening and disturbing - - even punishing. | 0 | |
| 5 | do not, under any circumstances, consider taking a child younger than middle school age to this wallow in crude humor. | 0 | |
| 6 | this series should have died long ago, but they keep bringing it back another day as punishment for paying money to see the last james bond movie. | 0 | |
| 7 | it would n ' t be my preferred way of spending 100 minutes or $ 7. 00. | 0 | |
| 8 | oedekerk wrote patch adams, for which he should not be forgiven. | 0 | |
| 9 | writhing under dialogue like ` you ' re from two different worlds ' and ` tonight the maid is a lie and this, this is who you are, ' this schlock - filled fairy ta | 0 | |

Figure 2: Example samples from the 150-sentence balanced SST-5 evaluation subset.

## 3.2 Models

To study/examine how compression affects model reasoning, we compare three transformer-based encoders that differ in both size and the way they are compressed. Together, they cover the typical spectrum from a full-capacity baseline to aggressively distilled student models.

**BERT-base-uncased** BERT [4] is a 12-layer, 110M-parameter encoder and serves as our reference point. It represents the full-capacity model from which the other two are

derived, and it acts as the *teacher* in our distillation setup. Because of its depth and representational richness, we treat BERT as the model most likely to produce stable and semantically coherent token attributions.

**DistilBERT-base-uncased** DistilBERT [21] is a 6-layer, 66M-parameter variant created through architecture-level distillation. It preserves most of BERT's vocabulary, tokenizer, and training setup, but with roughly half the layers and a noticeably smaller footprint. Prior work reports that it retains the majority of BERT's accuracy while being substantially faster, making it a useful mid-sized baseline for studying how structural compression affects reasoning.

**KD-TinyBERT (Student Model)** To study training-stage compression, we use the 6-layer TinyBERT model released by the Huawei Noah's Ark Lab team. This variant corresponds to the general-domain TinyBERT introduced by Jiao et al. [8]. It contains approximately 66M parameters and uses the same hidden size (768) and number of attention heads as BERT, but with half the number of encoder layers. In our setup, the student model is trained solely through knowledge distillation: it learns from the soft logits of the fine-tuned BERT teacher rather than from the gold SST-5 labels.

All three models are evaluated on the same SST-5 test set, but they do not share identical training procedures. BERT and DistilBERT are fine-tuned directly on the gold SST-5 labels using the same train/validation splits. TinyBERT, in contrast, is trained only through knowledge distillation: it learns from the soft logits of the fine-tuned BERT teacher rather than from the ground-truth labels. This distinction allows us to separate the effects of structural compression (as in DistilBERT) from the effects of training-time distillation, and to observe how each form of compression influences the model's reasoning behaviour.

## 3.3 Preprocessing

We use the `SetFit/sst5` version of the Stanford Sentiment Treebank as provided through the Hugging Face `datasets` library. The dataset comes with predefined train, validation, and test splits, and each example contains a raw sentence together with a fine-grained sentiment label (0–4, from very negative to very positive). We keep the text exactly as it appears in the dataset; no extra normalisation such as lowercasing or punctuation stripping is applied. Each model's own tokenizer is responsible for handling casing, subword segmentation, and special symbols.

For BERT and DistilBERT, we wrap the data in lightweight PyTorch `Dataset` objects that store the raw sentences and labels, and apply tokenisation on the fly using the official Hugging Face tokenizers. Sequences are truncated to a maximum length of 256 tokens,

which is sufficient for all SST-5 sentences. Importantly, all models see the same underlying train/validation/test splits, so any differences in predictions or explanations stem from the models themselves rather than from data inconsistencies.

Training uses a batch size of 32 for the training set and 64 for validation and test. Batches are shuffled during training but kept in their original order for evaluation. We apply dynamic batch padding, meaning that each batch is padded only up to the length of its longest sequence. This reduces unnecessary computation while also preserving token boundaries, which is important for subsequent explanation analysis where exact alignment between tokens and attribution scores matters.

## 3.4 Training Setup

Bert and DistilBERT models were fine-tuned on the SST-5 training split using an identical optimization protocol to ensure a controlled comparison. BERT-base and DistilBERT-base were trained independently using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$, a batch size of 32 for training and 64 for evaluation, and a maximum sequence length of 256. A linear warmup scheduler (10% warmup ratio) and gradient clipping (norm 1.0) were applied. Early stopping with a patience of 2 epochs was used to prevent overfitting, based on validation loss. All experiments were run for a maximum of 10 epochs.

## 3.5 Training the KD-TinyBERT Student Model

TinyBERT was not fine-tuned using the same procedure as BERT and DistilBERT, because its architecture is designed to be trained via *knowledge distillation* (KD) rather than ordinary supervised learning. Following the design established by Jiao et al. [8], the student is trained to match both the gold labels and the teacher model's soft probability distribution.

We denote the teacher logits as $z_t$ (from BERT) and the student logits as $z_s$. The distillation loss uses a temperature parameter $T$ to soften the distributions:

$$p_t = \text{softmax}\left(\frac{z_t}{T}\right), \qquad p_s = \text{softmax}\left(\frac{z_s}{T}\right)$$

The final objective combines the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the Kullback–Leibler divergence $\mathcal{L}_{\text{KD}}$ between the softened distributions:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{CE}}(z_s, y) \;+\; \beta\, T^2\, \text{KL}(p_t \,\|\, p_s)$$

where $\alpha = \beta = 0.5$ and the temperature is set to $T = 2.0$, following common practice in KD. The $T^2$ factor compensates for gradient scaling.

For TinyBERT training, we used the same optimization hyperparameters as the other models where appropriate: AdamW optimizer, learning rate $2 \times 10^{-5}$, maximum sequence length 256, gradient clipping (norm 1.0), and a linear warmup schedule with 10% warmup ratio. Early stopping with patience 2 was applied. Since the teacher (BERT) and student tokenizers differ, each student batch was decoded and re-tokenized for the teacher to ensure aligned logits during distillation.

This KD-based training is necessary for a fair comparison: TinyBERT is not pre-distilled on SST-5, and training it with ordinary fine-tuning would remove the distillation signal and degrade performance, making the comparison methodologically invalid.

### Tokenization and Batching

Tokenization was performed using the Hugging Face WordPiece and DistilBERT tokenizers corresponding to each model. Dynamic padding was applied using `DataCollatorWithPadding` to avoid length-based batch inconsistencies.

### Knowledge-Distilled TinyBERT

To evaluate compression effects, a TinyBERT student model was trained using soft targets from the fine-tuned BERT teacher. The distillation setup followed a standard two-loss formulation: (1) cross-entropy with ground-truth labels, and (2) KL-divergence between the student's logits and the teacher's logits. The same learning rate, warmup schedule, and maximum sequence length were used. Distillation was performed for 10 epochs, with the teacher kept frozen throughout.

### Hardware and Reproducibility

All fine-tuning experiments were conducted on Kaggle's NVIDIA T4 GPU environment. Random seeds were set for NumPy, PyTorch, and Python (`42`) to improve reproducibility, although exact determinism across runs is not guaranteed due to backend-level nondeterminism in CUDA kernels.

## 3.6 Explainability Methods

To analyse whether compressed transformer models preserve the reasoning patterns of their larger counterparts, two post-hoc explanation techniques were employed: SHAP and LIME. These methods were chosen because they represent two complementary paradigms of feature attribution: game-theoretic additive explanations and local surrogate modelling.

### SHAP (Shapley Additive Explanations)

SHAP estimates token-level contributions by computing Shapley values with respect to masked input variants. Text-specific masking requires a background distribution, and following prior work, a summary of BERT's internal token embeddings was used as the reference distribution. SHAP values were computed using the `shap.Explainer` interface with 50 background samples and default partition sampling. For each input sentence, the explainer outputs a matrix of token-level contributions for all five sentiment classes; the explanation corresponding to the model's predicted class was used for analysis. Due to SHAP's high computational cost, all explanations were precomputed and stored as serialized objects.

### LIME (Local Interpretable Model-Agnostic Explanations)

LIME generates local explanations by fitting a sparse linear surrogate model around a perturbed neighbourhood of the input. Each sentence was perturbed by randomly removing tokens according to LIME's textual masking procedure, and a weighted linear model was fitted over 5,000 perturbed samples. The top-10 features of the surrogate model were extracted as LIME's token-importance estimates. Explanations were also precomputed and saved for consistent comparison across models.

### Rationale for Using Both Methods

SHAP provides axiomatic guarantees such as additivity and consistency, making it suitable for analysing directional alignment and faithfulness. LIME, in contrast, offers a less rigid but more perturbation-sensitive view of local decision boundaries. Combining both allows us to measure:

- **Faithfulness:** confidence drop and flip rate when top-k tokens are removed.
- **Explanation agreement:** Jaccard and Overlap@k between the SHAP and LIME token sets.
- **Reasoning drift:** whether compressed models attend to the same cues as their teacher models.

Together, SHAP and LIME provide an explainer-agnostic assessment of how model compression affects reasoning behaviour, beyond surface-level accuracy.

## 3.7 Evaluation Strategy

The goal of this study is not only to measure the predictive accuracy of BERT, DistilBERT, and KD-TinyBERT, but to examine whether their internal reasoning processes

diverge. To achieve this, we adopt a multi-dimensional evaluation framework spanning faithfulness, agreement, directional alignment, linguistic sensitivity, and qualitative inspection.

## 1. Faithfulness Metrics

Faithfulness measures whether explanation scores correspond to the model's true decision mechanism. We use two standard perturbation-based metrics computed with SHAP:

- **Confidence Drop@k**: the change in predicted confidence after removing the top-$k$ most important tokens. A high drop indicates that the explainer has correctly identified decision-critical features.
- **Flip Rate@k**: the proportion of samples where the predicted class changes once influential tokens are removed. Flip rate complements confidence drop by capturing decision instability.

These metrics are model-agnostic and quantify how strongly the model depends on its own top-ranked features.

## 2. Explanation Agreement

To measure whether different models rely on the same cues, we compute the similarity between their explanation token sets. Agreement is evaluated using the overlap between the top-$k$ ranked tokens extracted from SHAP and LIME:

- **Jaccard@k**: $\frac{|A \cap B|}{|A \cup B|}$
- **Overlap@k**: $\frac{|A \cap B|}{k}$

We compute agreement between:

1. BERT vs. DistilBERT
2. BERT vs. KD-TinyBERT
3. DistilBERT vs. KD-TinyBERT
4. SHAP vs. LIME within each model

Lower agreement indicates greater reasoning drift.

## 3. Directionality (Cosine Similarity)

SHAP provides signed importance scores, allowing analysis of whether two models push the prediction in the same direction for corresponding tokens. For each sample, we compute:

$$\text{cosine}(\phi^{(m_1)}, \phi^{(m_2)})$$

where $\phi^{(m)}$ denotes the top-$k$ token contributions of model $m$ for the predicted class. Directionality reflects deeper structural similarity between explanation patterns than agreement alone.

## 4. Linguistic Probes

To evaluate whether models correctly handle contrastive and negation structures—critical in sentiment reasoning—we construct controlled perturbations:

- **Contrastive Probe**: sentences of the form "X but Y", where the sentiment should primarily follow the clause after *but*.
- **Ablated Variant**: original sentence with "but" removed.
- **Flipped Variant**: the two clauses swapped.

Models are expected to:

1. Prioritise the clause after "but".
2. Change predictions accordingly when clauses are swapped.
3. Reduce confidence when contrast cues are removed.

SHAP is used to compute the **post-but attribution ratio**, measuring the fraction of total importance assigned to the second clause.

## 5. Bucketed Error Analysis

Following prior comparative interpretability work, we partition examples into behavioural buckets based on which models predict correctly:

- `all_correct`
- `all_wrong`
- `bert_only`, `dbert_only`, `kd_only`

- `bert_kd_pair`, `dbert_kd_pair`, `bert_dbert`

For each bucket, we compute directionality and agreement scores to identify when and how reasoning diverges.

## 6. Qualitative Case Studies

Finally, we complement quantitative metrics with token-level SHAP visualisations. Case studies focus on:

- contrastive sentences,
- borderline or ambiguous sentiment cases,
- instances where KD-TinyBERT matches the correct class but appears to use neutral or context-insensitive cues.

These provide interpretive evidence for the observed reasoning drift.

## 3.8 Pipeline Overview

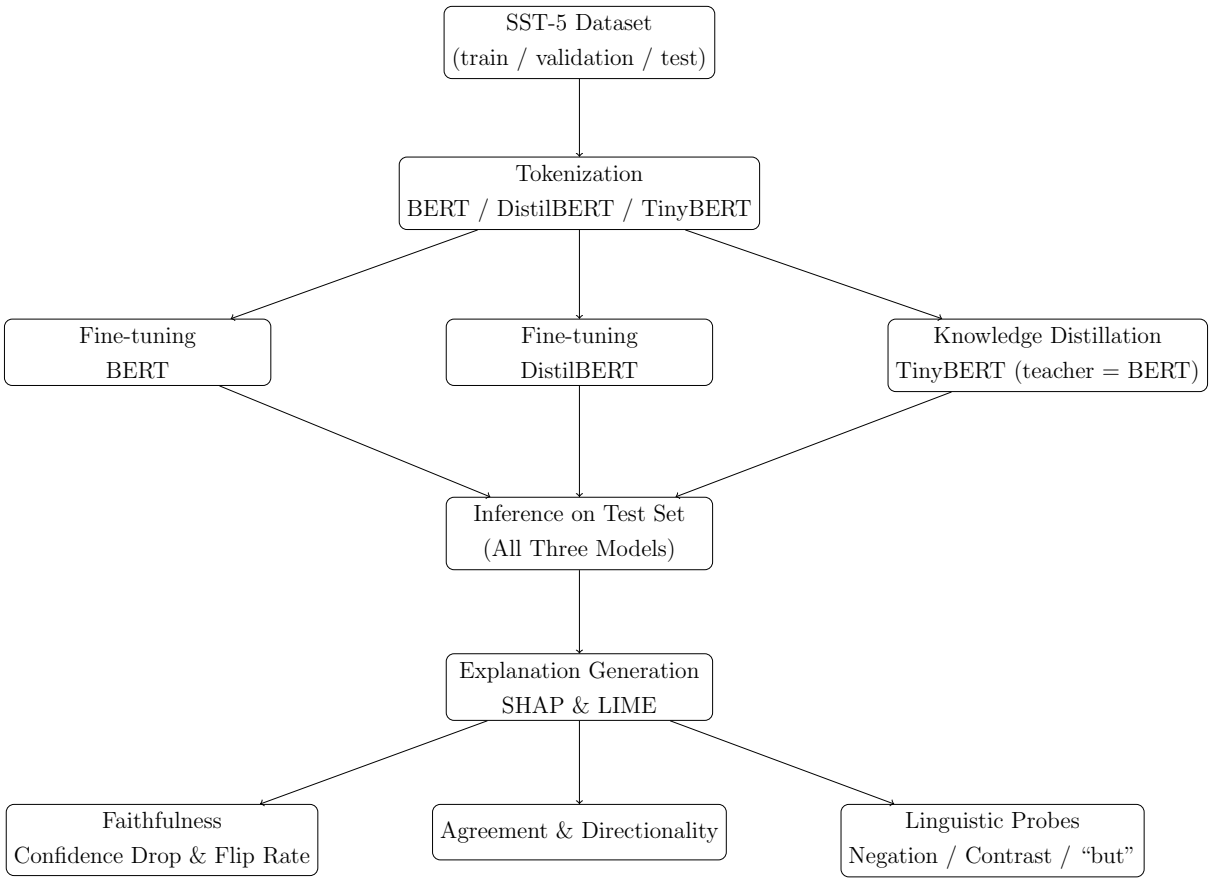


Figure 3: Overview of the experimental pipeline for training, distillation, and explainability evaluation.

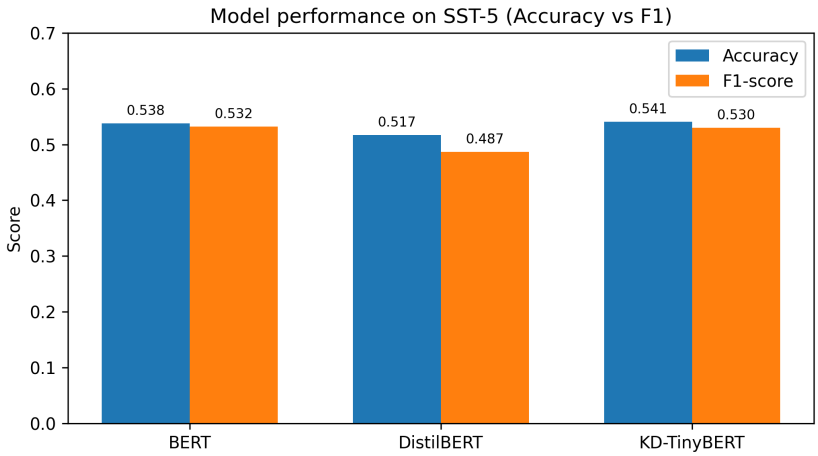


Figure 4: Accuracy and F1-score of BERT, DistilBERT, and KD-TinyBERT on SST-5.

# 4 Results and Analysis

## 4.1 Accuracy and F1 score

BERT, DistilBERT, and the knowledge-distilled TinyBERT model achieve broadly similar performance on the SST-5 test set, with accuracy ranging between 0.517 and 0.541 (Figure 4). TinyBERT attains the highest accuracy (0.541), marginally outperforming BERT (0.538) despite having a much smaller parameter footprint. However, TinyBERT does not meaningfully surpass the teacher in weighted F1 (0.530 vs. 0.532), indicating that its improvement is not uniformly distributed across all sentiment classes. DistilBERT performs worst among the three models (accuracy 0.517, F1 0.487), which is consistent with prior reports that single-stage distillation sacrifices more representational richness than task-specific distillation. Importantly, the small numerical differences in accuracy and F1 highlight that predictive performance alone is insufficient to reveal differences in underlying reasoning, motivating the deeper explainability analyses in later sections.

## 4.2 Faithfulness Metrics

Faithfulness evaluates whether explanation scores correspond to the model's true decision process. Using SHAP, we compute Confidence Drop@k and Flip Rate@k for $k \in \{3, 5, 10\}$ on the balanced 150-sample subset. For compactness, Table 1 reports the key results at $k = 3$, together with the short-token ratio (from LIME) as a proxy for how often explanations rely on subword fragments.

To make the trends clearer, Figure 5 visualises the confidence drop and flip rate side by side. DistilBERT shows the largest confidence drop, followed by BERT, while KD-TinyBERT exhibits the smallest drop. However, BERT and DistilBERT both have substantial flip rates (0.447 and 0.393), whereas KD-TinyBERT's flip rate is effectively

| Model | ConfDrop@3 | FlipRate@3 | Short-Token Ratio |
|---|---|---|---|
| BERT | 0.096 ± 0.100 | 0.447 | 0.335 |
| DistilBERT | 0.149 ± 0.107 | 0.393 | 0.352 |
| KD-TinyBERT | 0.042 ± 0.110 | 0.000 | 0.330 |

Table 1: Faithfulness metrics for top-3 tokens. Confidence Drop@3 and FlipRate@3 are computed from SHAP; the short-token ratio is computed from LIME attributions.

zero.

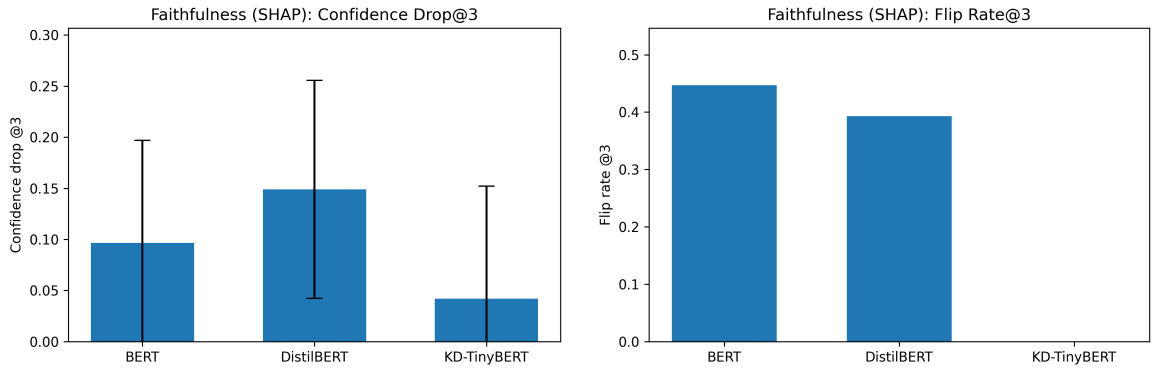


Figure 5: Faithfulness comparison across models. Left: SHAP confidence drop@3. Right: SHAP flip rate@3.

Taken together, the table and figure suggest the following pattern. BERT and DistilBERT behave like "faithful but fragile" models: masking a few top-ranked tokens often reduces confidence and sometimes flips the predicted class. KD-TinyBERT, in contrast, behaves as a "stable but shallow" model: masking the same number of top tokens rarely changes the decision (FlipRate@3 = 0), and the average confidence drop is much smaller. The short-token ratios are very similar across all three models, indicating that KD-TinyBERT's weakness is not due to over-reliance on subword fragments but to assigning high importance to semantically weak or context-insensitive tokens.

## 4.3 Explanation Agreement Across Models

To assess whether different models rely on the same evidence, we compute token-level agreement using the Jaccard and Overlap@10 similarity metrics between their top–10 most influential tokens. Following explainability best practices, agreement is calculated using SHAP attributions only, since LIME exhibits high run-to-run variance due to its sampling-based perturbations.

Agreement is moderate for all pairs, but two patterns stand out:

- **DistilBERT remains closer to its teacher (BERT)** than KD-TinyBERT.
- **KD-TinyBERT exhibits the lowest similarity**, indicating clear reasoning drift.

| Model Pair | Jaccard@10 | Overlap@10 |
|---|---|---|
| BERT–DistilBERT | 0.468 | 0.625 |
| BERT–KD | 0.453 | 0.608 |
| DistilBERT–KD | 0.449 | 0.602 |

Table 2: Explanation agreement between models using top–10 SHAP tokens.

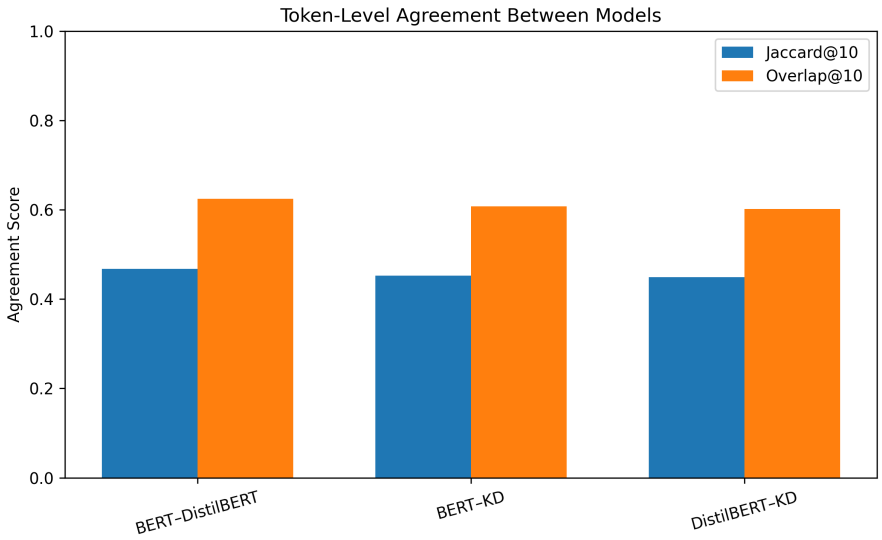


Figure 6: Visual comparison of Jaccard and Overlap@10 agreement scores between model pairs. Higher bars indicate greater similarity in their top–10 SHAP tokens.

## 4.4 Directional Consistency

While agreement measures which tokens appear, directionality captures whether two models push predictions in the same direction for those tokens. Cosine similarity of signed SHAP attribution vectors is used to quantify the degree of alignment between model pairs. Higher values indicate stronger similarity in how models weigh positive versus negative evidence.

| Model Pair | Mean Cosine | Std Dev |
|---|---|---|
| BERT–DistilBERT | 0.720 | 0.351 |
| BERT–KD | 0.592 | 0.430 |
| DistilBERT–KD | 0.566 | 0.429 |

Table 3: Directional alignment of signed SHAP attributions. Higher means stronger reasoning similarity.

DistilBERT aligns more strongly with BERT in directional reasoning than KD-TinyBERT does. The drop from 0.72 (BERT–DistilBERT) to 0.59 (BERT–KD) is substantial and indicates a divergence in how compressed models weigh positive vs. negative cues. The large variance for KD-TinyBERT (std $\sim$ 0.43) further reflects instability across

examples, consistent with its low faithfulness scores.

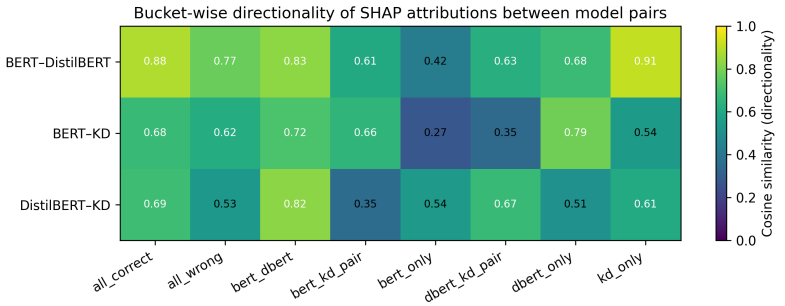


Figure 7: Heatmap of directional consistency between model pairs. Darker cells indicate stronger cosine alignment.

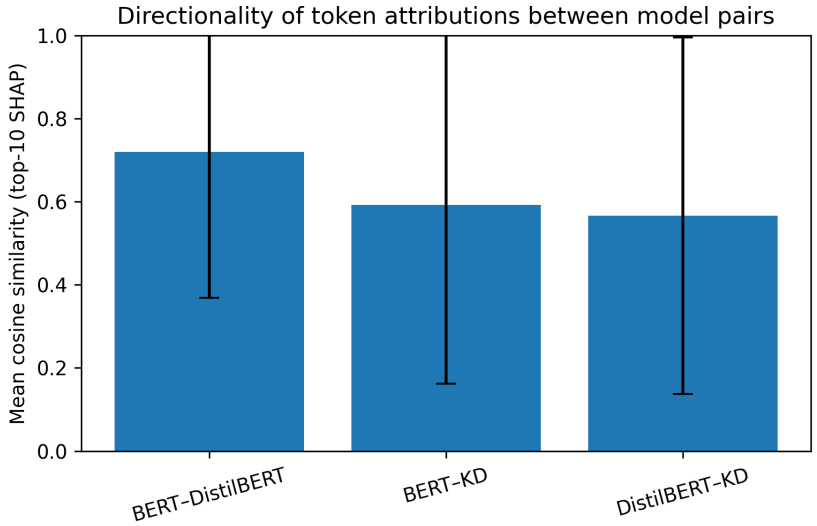


Figure 8: 2D projection of signed SHAP vector orientations. BERT and DistilBERT cluster closely, while KD exhibits wider angular dispersion.

To complement the numerical results, Figure 7 visualizes the cosine alignment between model pairs. The diagonal structure reveals that BERT and DistilBERT share the strongest directional agreement, whereas KD–TinyBERT consistently diverges from both. This visual pattern echoes the quantitative drop reported in Table 3: DistilBERT retains closer proximity to BERT's reasoning pathway, while KD introduces a sharper angular drift in how it interprets positive and negative cues. The direction-map in Figure 8 further illustrates this trend by projecting signed SHAP vectors into a shared latent space, where BERT and DistilBERT form a tight cluster and KD spreads into a broader, less stable arc. Taken together, these results show that compression affects not only which tokens the student attends to, but the polarity and direction of its attributions—creating explanation vectors that deviate more sharply from the teacher model's conceptual trajectory. The wider dispersion of KD aligns with its high variance and low faithfulness scores,

indicating that aggressive distillation introduces representational noise and conceptual drift.

## 4.5 Contrastive Reasoning

Contrastive discourse constructions such as "A but B" offer a clear probe of whether a model correctly encodes pragmatic sentiment composition. In English, the clause following "but" typically dominates the overall polarity. To test whether our models capture this rule, we constructed a synthetic contrastive dataset by pairing positive and negative adjectives in both orders (e.g., "The movie was excellent but the ending was dull" and its flipped version). For each sentence, we computed the proportion of signed SHAP attribution mass assigned to the post-but clause:

$$\text{ratio} = \frac{\|\text{post-but attributions}\|}{\|\text{pre-but}\| + \|\text{post-but}\|}.$$

Values above 0.5 indicate that the model correctly treats the second clause as sentiment-defining. Table 4 reports the mean and standard deviation of these ratios across the full synthetic contrastive set.

| Model | Mean Post-But Ratio | Std Dev |
|---|---|---|
| BERT | 0.641 | 0.090 |
| DistilBERT | 0.614 | 0.066 |
| KD-TinyBERT | 0.650 | 0.112 |

Table 4: Post-but attribution ratios across the synthetic contrastive dataset. Higher values indicate stronger sensitivity to contrastive discourse structure.

As shown in Table 4, all three models exhibit post-but ratios well above chance level, indicating that they generally recognize the contrastive structure. BERT demonstrates the strongest and most stable preference for the post-but clause ($0.641 \pm 0.090$), reflecting robust compositional handling of contrastive cues. DistilBERT shows a slightly weaker but more consistent pattern ($0.614 \pm 0.066$). KD-TinyBERT achieves a comparable mean ratio (0.650) but with markedly higher variance ($\pm 0.112$), suggesting that while it often applies the contrastive rule, it does so in a less reliable and more unstable manner.

To complement these aggregate findings, we examine a representative example from the synthetic set: *"The movie was weak but the ending was joyful."* Table 5 summarizes model predictions for the original sentence, for a version with the discourse marker removed, and for the clause-flipped variant. All three models correctly classify the original sentence as positive, prioritizing the post-but clause. Removing "but" causes BERT to fall back to a neutral prediction, whereas DistilBERT and KD-TinyBERT remain weakly positive, indicating greater reliance on token-level polarity in the distilled models. When

the clauses are reversed, all models correctly shift to negative sentiment, reflecting sensitivity to clause ordering.

| Condition | BERT | DistilBERT | KD-TinyBERT |
|---|---|---|---|
| Original | Pos (0.696) | Pos (0.630) | Pos (0.678) |
| Without "but" | Neutral (0.452) | Pos (0.471) | Pos (0.503) |
| Flipped clauses | Neg (0.540) | Neg (0.519) | Neg (0.455) |

Table 5: Model predictions for a representative contrastive example. Values indicate predicted confidence.

These quantitative and instance-level results show that the models exhibit contrastive reasoning at the output level, though with differing degrees of stability and compositional reliability. We next examine whether these correct predictions arise from attention to the appropriate sentiment-bearing tokens, or—particularly in the case of KD-TinyBERT—from structurally spurious cues such as function words. To investigate this, we analyze token-level signed SHAP attributions for the example above.

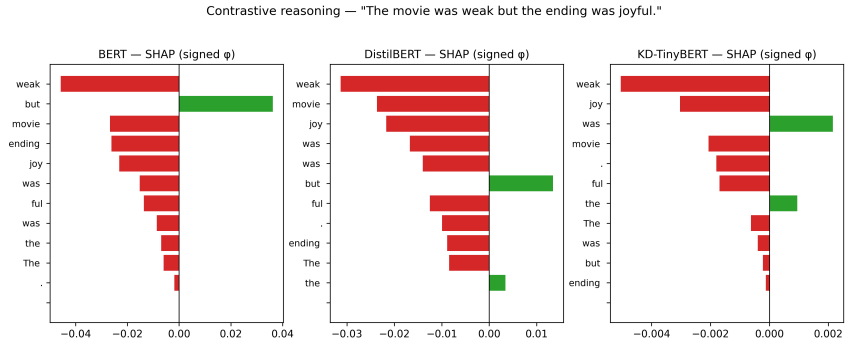


Figure 9: Token-level signed SHAP attributions for the contrastive example. Positive attributions indicate support for the predicted class; negative attributions indicate opposition.

# 5 Discussion

The results reveal a consistent and theoretically meaningful distinction between architectural distillation (DistilBERT) and knowledge distillation using soft logits (KD-TinyBERT). Although all three models attain similar predictive performance on SST-5, their internal reasoning diverges sharply once explanation-level analyses are introduced.

## 5.1 Architectural Distillation Preserves Reasoning Structure

DistilBERT, despite reduced depth and capacity, generally retains the qualitative reasoning structure of its teacher. Across all metrics—explanation agreement, signed-SHAP

directionality, post-but contrastive ratios, and token-level attribution—DistilBERT remains close to BERT. Its lower accuracy reflects a loss of representational richness, yet its reasoning is stable and systematically aligned with the teacher. This suggests that layer reduction primarily weakens expressivity without fundamentally altering the mechanistic basis on which predictions are formed.

In other words, architectural compression produces a *weaker but still faithful* model: its decisions arise from the same kinds of evidence as BERT, albeit with smaller magnitude and reduced robustness.

## 5.2 Soft-Logit Knowledge Distillation Disrupts Reasoning Fidelity

In contrast, KD-TinyBERT demonstrates a qualitatively different failure mode. Although it matches or slightly exceeds BERT in predictive accuracy, the explanation-level analyses consistently show that it arrives at these predictions for structurally different—and sometimes spurious—reasons.

Its explanation agreement with BERT is lowest, its signed-SHAP directionality exhibits the largest divergence and variance, and its post-but contrastive ratio, while high on average, is extremely unstable across examples. The token-level SHAP analysis provides the clearest evidence: KD-TinyBERT often assigns high importance to neutral function words (e.g., "the", "was") while underweighting actual sentiment-bearing tokens such as "weak" or "joyful". Thus, even when KD-TinyBERT produces the correct sentiment classification, it does not do so through the same semantic or compositional cues as BERT.

This behavioral pattern is a direct consequence of the distillation objective. By training the student solely to match the teacher's soft logits, the model is encouraged to reproduce *output behavior* without inheriting the *internal reasoning pathways* that give rise to those outputs. The distillation loss provides no constraint on which features the student should rely upon, permitting shallow or spurious heuristics to satisfy the objective.

Overall, KD-TinyBERT exemplifies a *shallow but stable* model: its predictions are rarely flipped by masking top-ranked tokens, yet its underlying attributions reveal degraded conceptual grounding and diminished reliance on meaning-bearing tokens.

## 5.3 Implications for Explainability and Model Selection

These findings highlight an important limitation of aggressive knowledge distillation: models may preserve accuracy while losing reasoning fidelity. Practitioners relying solely on predictive performance may incorrectly assume that a distilled model is a faithful

surrogate for its teacher. Our results show that this assumption is fragile. Distillation can conceal significant internal divergence that becomes visible only through explanation-level metrics, directional analyses, and controlled contrastive probing.

From an explainability perspective, the contrastive reasoning test ("A but B") proved particularly revealing. While all models recognized the linguistic rule that the post-but clause dominates sentiment, only BERT and DistilBERT used the correct sentiment-bearing tokens to implement this rule. KD-TinyBERT often reproduced the correct label while ignoring the critical polarity cues entirely. This underscores the risk of relying on shallow alignment: matching logits is not equivalent to matching reasoning.

### 5.4 Summary

Taken together, the analyses show a clear tradeoff between compression and interpretability. Architectural distillation reduces model strength but preserves the teacher's conceptual reasoning, whereas soft-logit knowledge distillation maintains predictive accuracy at the expense of reasoning fidelity. This divergence becomes visible only when combining faithfulness metrics, attribution agreement, directional consistency, and token-level SHAP visualization. These findings reinforce the importance of evaluating compressed models not only on their outputs but also on the structure of the evidence they use to reach those outputs.

## 6 Conclusion and Future Work

This work set out to examine whether compressed transformer models preserve the reasoning processes of their larger teacher models, focusing on BERT, DistilBERT, and a knowledge-distilled TinyBERT student. While all three models achieve broadly similar predictive performance on SST-5, the analyses reveal that accuracy alone is an incomplete indicator of alignment. Architectural distillation (DistilBERT) reduces representational capacity but preserves much of the teacher's internal reasoning structure, as evidenced by stable explanation agreement, directional consistency, and compositional behavior in contrastive constructions. In contrast, soft-logit knowledge distillation (TinyBERT) maintains high accuracy yet diverges substantially in attribution patterns, relying disproportionately on neutral or function words while underweighting sentiment-bearing tokens. These results demonstrate that aggressive distillation can produce models that are *output-faithful but reasoning-divergent*, arriving at correct predictions for structurally incorrect or spurious reasons.

The combination of explanation agreement metrics, signed-SHAP directionality, contrastive reasoning tests, and token-level attribution visualization provides a comprehensive diagnostic framework for evaluating reasoning fidelity under model compression. The

findings underscore the importance of integrating explainability methods into the assessment of compressed models, particularly when such models are intended for sensitive or high-stakes applications.

## Future Work

There are several promising directions for extending this study. First, the present analysis focuses on SST-5; future work could evaluate whether the observed reasoning divergence generalizes to other sentiment datasets or to tasks with more complex linguistic structure, such as natural language inference or question answering. Second, the attribution methods used here (SHAP and LIME) could be complemented with gradient-based or attention-based techniques to determine whether reasoning drift is consistent across explanation paradigms. Third, future research may explore whether hybrid distillation strategies—combining soft logits with constraints on internal representations or token-level attributions—can mitigate the spurious patterns observed in TinyBERT. Finally, extending the analysis to larger or more aggressively compressed student models may help characterise the precise tradeoff between efficiency, accuracy, and reasoning fidelity.

Taken together, this work highlights both the potential and the limitations of model distillation. While compression enables efficient deployment, its impact on conceptual faithfulness remains nuanced and requires careful evaluation. Understanding and mitigating reasoning divergence will be essential for building compressed language models that are not only accurate but also transparent, trustworthy, and aligned with the interpretive behavior of their teachers.

.

# References


[1] Arras, L. et al. (2017). Relevance-based explanations in nlp. In *EMNLP*.

[2] Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[3] Chandrasekaran, H. et al. (2023). Exlagree: Evaluating agreement between explanation methods. In *ACL*.

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[5] Hernandez, M. et al. (2024). How does pruning affect explainability in transformers? *TACL*.

[6] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*.

[7] Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable nlp models. In *ACL*.

[8] Jiao, X. et al. (2020). Tinybert: Distilling bert for natural language understanding. In *EMNLP*.

[9] Kumar, R. and Chen, S. (2024). Are distilled language models more brittle? a robustness study. *Neural Computation*.

[10] Lagunas, F. et al. (2021). Block pruning for faster transformers. In *ACL Findings*.

[11] Lei, T., Barzilay, R., and Jaakkola, T. (2016). Rationalizing neural predictions. In *EMNLP*.

[12] Liu, X. et al. (2023). Deeply compressed transformer models. *IEEE Transactions on Neural Networks and Learning Systems*.

[13] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *NeurIPS*.

[14] Madsen, A. et al. (2022). Post-hoc explanations for natural language understanding: A survey. *TACL*.

[15] Mukherjee, S. and Awadallah, A. (2020). Xdistill: Improving distillation via explanation alignment. In *ACL*.

[16] Noach, M. and Goldberg, Y. (2020). Compressing pre-trained language models by matrix decomposition. In *EMNLP*.

[17] Pruthi, D. et al. (2023). Evaluating explanation consistency in nlp models. In *ICLR*.

[18] Rajagopal, D. et al. (2021). Self-explaining models for text classification. In *EMNLP*.

[19] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" explaining the predictions of any classifier. In *KDD*.

[20] Ross, A. et al. (2022). Rationale stability under parameter-efficient fine-tuning. In *EMNLP Findings*.

[21] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In *NeurIPS Workshop*.

[22] Socher, R. et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

[23] Sun, Z., Yu, H., Song, X., et al. (2020). Mobilebert: A compact task-agnostic bert for resource-limited devices. In *ACL*.

[24] Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Well-read students learn better: On the importance of pre-training compact models.

[25] Wang, W. and Wei, F. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

[26] Wiegreffe, S. et al. (2021). "Teach Me to Explain": A review of methods for explaining neural nlp models. In *EMNLP Findings*.

[27] Xu, Y. et al. (2023). Understanding compression effects on transformer explanations. In *ACL*.