© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713617

Investigating Data Leakage–Induced Over-Confidence and Explanation Faithfulness in Transformer-Based Text and Audio Models

RAHUL VAKITI

I. INTRODUCTION

Transformer-based models now sit at the center of modern artificial intelligence, powering systems that read, listen, respond, and increasingly act on behalf of humans. From large-scale language models to voice-driven assistants, these systems exhibit a level of fluency and responsiveness that would have seemed implausible only a few years ago. Their success has been driven largely by scale Larger models, larger datasets, and longer training regimes resulting in impressive performance across text and audio tasks.

However, this apparent intelligence masks a growing tension between performance and trust. As these models are deployed in real-world settings, failures are rarely obvious or catastrophic, instead, they are subtile, confident, and silent. Models often produce answers that sound correct, explanations that appear reasonable, and confidence estimates that suggest certainty—even when the underlying reasoning is fragile or absent. This shift has transformed the central question of AI research from "Can the model perform the task?", "Why does it matter"? A particularly concerning contributor to this gap is data leakage and over-training. In large text and audio models, repeated exposure to similar samples, duplicated corpora, or identity-specific cues can lead models to rely on memorization rather than generalizable reasoning. In such cases, confidence is no longer a reflection of understanding, but of familiarity. The problem is amplified in audio and voice-based systems, where speaker identity, acoustic patterns, or recurring background signals can become shortcuts that inflate confidence without improving semantic comprehension.

Compounding this issue, existing explainability methods often fail to reveal these failure modes. Explanations may remain stable and visually convincing even when predictions are driven by leaked or memorized information. This raises a critical concern: if explanations do not change when the causal basis of a prediction changes, they risk becoming narratives rather than diagnostics. Understanding when and why such explanations lose faithfulness is essential for building AI systems that are not only powerful, but also transparent and safe.

This work focuses on examining how training influences model confidence and explanation behavior, and on identifying when interpretability tools succeed or fail in exposing the true decision-making processes of modern AI systems.

II. MOTIVATION AND OBSERVED FAILURE MODES

During practical experimentation with fine-tuning a speech recognition model based on the Whisper architecture, I encountered behaviors that motivated this research direction. The model was fine-tuned on approximately ten hours of audio data, which was segmented into around 1,900 audio files with corresponding transcriptions extracted for supervised training. The initial training run showed promising behavior: when evaluated on unseen samples, the model produced reasonable transcriptions, and qualitative errors appeared consistent with expected generalization limits for a dataset of this size.

However, as the same dataset was reused for repeated fine-tuning cycles, unexpected pat- terns began to emerge. After training the model multiple times—three to four successive fine- tuning runs on the same data—the model's behavior changed noticeably. While standard evaluation metrics showed only marginal improvements (on the order of one to two percentage points), the qualitative behavior of the model became increasingly rigid. Outputs grew more predictable, error patterns repeated across samples, and the model began producing confident transcriptions even in cases where the input audio was noisy or ambiguous.

More concerningly, the model appeared to integrate fragments of memorized transcriptions rather than relying on acoustic cues alone. In several instances, partial phrases from the training data surfaced in predictions for unrelated audio segments, suggesting that the model had begun to rely on memorized patterns rather than genuine speech understanding. This behavior became more pronounced as the number of fine-tuning iterations increased, despite the dataset remaining unchanged.

These observations highlight a subtle but important failure mode: repeated exposure to a limited dataset can inflate model confidence without improving semantic understanding. From a numerical standpoint, the model appeared to improve slightly from a behavioral standpoint, it became less adaptable and more reliant on familiarity. This gap between quantitative metrics and qualitative behavior suggests the presence of data leakage like effects, where memorization masquerades as learning.

Similar patterns are not limited to speech recognition. In text-based transformer models, repeated fine-tuning on narrow or overlapping corpora can produce highly confident and stylistically consistent outputs that fail to generalize beyond familiar patterns. In both text and audio domains, confidence becomes a weak proxy for correctness once memorization dominates the learning process.

These real world observations motivate a deeper investigation into how training dynamics particularly repeated fine-tuning on limited or overlapping data effect model confidence and explanation behavior. They also raise critical questions about whether existing evaluation and interpretability tools are sufficient to detect such failure modes before deployment in real-world systems.

III. RELATED WORK

Data Leakage and Memorization in Neural Models
Data leakage and memorization have been widely studied as critical challenges in modern neural network training, particularly in large-scale transformer models. Data leakage broadly refers to situations where information that should not be available during training or inference—such as test data, duplicated samples, or indirect label cues—implicitly influences model behavior. In large datasets collected from heterogeneouss sources, leakage can arise unintentionally through dataset overlap, repeated examples, or near-duplicate samples.

Memorization is a closely related phenomenon, where neural models store specific training instances rather than learning generalizable patterns. Prior studies have shown that deep neural networks, especially over-parameterized models, are capable of memorizing rare or unique examples from their training data. This behavior is amplified in scenarios involving repeated training, fine-tuning on limited datasets, or exposure to highly correlated samples. While memorization does not always degrade benchmark performance, it can significantly affect model reliability and privacy.

In natural language processing, dataset overlap across pretraining and fine-tuning corpora has been identified as a major source of leakage. Large language models trained on web- scale data may encounter the same documents multiple times across different training stages, making it difficult to distinguish between genuine generalization and recall. Similar concerns arise in speech and audio models, where repeated exposure to the same speakers, acoustic environments, or transcription patterns can cause models to rely on identity-specific cues rather than semantic understanding.

Importantly, memorization-driven behavior often manifests without clear signals in conventional evaluation metrics. Models may continue to improve marginally in accuracy or loss while increasingly depending on memorized structures. This creates a deceptive appearance of progress, where confidence and output stability increase despite limited improvements in true generalization. As a result, data leakage and memorization represent failure modes that are difficult to detect through standard validation pipelines. Data Leakage most occurs in the real time model training These findings suggest that leakage and memorization are not isolated issues but systemic risks inherent to large-scale training and fine-tuning practices. Understanding how such effects influence downstream behavior—particularly model confidence and explanation reliability remains an open research challenge, motivating further investigation into diagnostic and evaluation-oriented approaches.
Model Confidence and Calibration
Model confidence is commonly used as a proxy for prediction reliability in neural systems however, prior work has shown that confidence estimates produced by deep neural networks are often poorly calibrated.

A well-calibrated model is expected to assign high confidence only when predictions are likely to be correct, and lower confidence when uncertainty is high. In practice, modern transformer-based models frequently violate this assumption, exhibiting over-confidence even in cases of incorrect or ambiguous predictions.

Over-confidence is particularly prevalent in large, over-parameterized models trained on extensive or overlapping datasets. Such models can assign near-certain probabilities to pre- dictions that arise from memorized patterns rather than robust inference. This issue becomes more pronounced during fine-tuning, where repeated exposure to limited or highly correlated data can reduce output variability while artificially inflating confidence scores. As a result, confidence values may reflect familiarity with training distributions rather than true epistemic certainty.

Miscalibration poses significant risks in real-world applications, especially in safety critical settings. In text-based systems, over-confident responses can obscure uncertainty in reasoning or factual grounding. In audio and speech-based models, confidence may be influenced by non- semantic cues such as speaker identity, acoustic consistency, or background patterns, leading to confident outputs even when semantic understanding is weak. These behaviors undermine the reliability of confidence as a signal for downstream decision-making.

Uncertainty estimation methods have been proposed to address these challenges, including calibration techniques and probabilistic modeling approaches. However, their effectiveness depends heavily on the underlying training dynamics and data quality. When over-confidence is driven by memorization or leakage, post-hoc calibration may fail to fully correct the mismatch between confidence and correctness. This suggests that confidence misalignment is not merely a surface level issue, but a deeper consequence of training-induced shortcut learning.
Understanding how confidence, calibration, and uncertainty interact under data leakage and over-training is therefore essential. Without such understanding, models may appear reliable based on confidence alone while remaining vulnerable to silent failure modes that escape standard evaluation protocols.

Explainability and Faithfulness
Explainability methods are commonly used to interpret the predictions of neural networks by highlighting influential inputs, internal activations, or attention patterns. In practice, these methods aim to provide insight into why a model produces a particular output, thereby improving transparency and trust. Popular approaches include attention-based visualizations, gradient- based attribution methods, perturbation techniques, and post-hoc surrogate models. While such tools often produce intuitive explanations, their ability to faithfully reflect the model's true decision-making process remains an open question.

A central limitation of many explainability techniques is that they operate at the level of input relevance rather than intent understanding. That is, they may correctly identify which tokens, audio segments, or features influenced a prediction, but fail to capture how or why those elements were used by the model. In models affected by memorization or data leakage, explanations can remain stable and plausible even when predictions are driven by shortcut patterns rather than semantic reasoning. As a result, explanations may appear correct while masking underlying failure modes.

This issue is particularly evident in text and audio transformer models. In text-based systems, attention weights or token attributions may highlight linguistically meaningful words even when the model relies on repeated phrase structures encountered during training. In audio models, saliency maps may emphasize prominent acoustic regions without revealing whether the prediction is driven by semantic speech content or by non-semantic cues such as speaker identity, pitch consistency, or background noise. In both cases, the explanation aligns with surface-level patterns but not necessarily with the true intent of the input data.

Faithfulness, in this context, refers to whether an explanation accurately reflects the causal mechanisms underlying a model's prediction. A faithful explanation should change meaning- fully when the decision basis of the model changes, and remain sensitive to perturbations that alter the model's reasoning process. However, when models rely on memorized or leaked in- formation, explanations may fail this criterion, remaining invariant even as the causal drivers shift. This disconnect limits the effectiveness of explainability tools as diagnostic instruments. These limitations highlight the need to evaluate explainability methods not only on their interpretability or visual clarity, but on their faithfulness to the model's true behavior. Under- standing when explainability tools succeed or fail in capturing intent and reasoning is critical for deploying reliable and trustworthy AI systems, particularly in settings where confident but incorrect predictions carry significant risk.

IV. PROBLEM FORMULATION AND SCOPE

Despite the growing capabilities of transformer-based models in text and audio domains, their reliability increasingly depends on factors beyond raw performance metrics. In practice, models are often evaluated based on accuracy, loss, or benchmark scores, while behavioral characteristics such as confidence, stability, and explanation consistency receive comparatively less attention. This creates a gap between how models are assessed and how they behave in real- world deployment.

The core problem addressed in this work is the emergence of over-confident behavior in transformer models trained under data leakage or over-training conditions, and the difficulty of detecting such behavior using existing interpretability tools. Specifically, when models are repeatedly fine-tuned on limited, overlapping, or highly correlated datasets, they may develop strong confidence signals that are driven by memorization rather than robust semantic reason- ing. In such cases, confidence ceases to be a reliable indicator of correctness.

This problem is compounded by the behavior of explainability methods applied to these models. Many interpretability techniques aim to highlight influential inputs or internal activations, yet they often fail to distinguish between genuine reasoning and shortcut learning. As a result, explanations may appear plausible and stable even when predictions are influenced by leaked or memorized information. This raises a critical concern: explanations that are not faithful to the true decision-making process can reinforce false trust rather than expose hidden failure modes.

Formally, this work studies transformer-based text and audio models that produce predictions accompanied by confidence estimates and post-hoc explanations. The focus is not on improving predictive performance, but on understanding the relationship between training dynamics, model confidence, and explanation behavior. The scope is limited to identifying when and how confidence becomes decoupled from generalization, and when interpretability tools succeed or fail in revealing this decoupling.

By framing the problem in this way, the objective is to shift attention from outcome-based evaluation toward behavior-based analysis. Understanding these failure modes is a necessary step toward developing safer, more trustworthy AI systems, particularly in applications where confident but incorrect predictions can lead to significant downstream consequences

Problem Formulation

We consider transformer-based models applied to text and audio understanding tasks, including language modeling, speech recognition, and sequence-to-sequence prediction. These models operate on input sequences derived from either textual tokens or acoustic representations and produce structured outputs such as predicted text, labels, or semantic interpretations. In addition to the primary prediction, modern systems often expose a confidence estimate—explicitly through probability distributions or implicitly through stable, deterministic outputs across re- peated inference.

Formally, let x denote an input instance, where x may represent a sequence of text tokens or an audio signal transformed into acoustic features. Given a trained transformer model f_θ, the model produces an output $y = f_\theta(x)$ along with an associated confidence measure c, typically derived from the model's output probabilities or entropy. In deployed systems, this confidence is frequently used as a proxy for reliability, guiding downstream decisions or user trust.

The failure modes of interest arise when the confidence estimate c becomes misaligned with the true generalization capability of the model. In particular, under conditions of data leakage, repeated fine-tuning, or exposure to highly overlapping datasets, the model may rely on memorized patterns rather than robust inference. In such cases, predictions may appear highly confident and stable even when the input deviates from genuine training

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713617

distributions or contains ambiguity.

This phenomenon is especially relevant in large text and audio models, where memorization can be subtle and difficult to detect. In text-based models, leakage may stem from repeated exposure to similar documents, phrases, or structural templates. In audio models, repeated speaker identities, acoustic environments, or transcription patterns can introduce identity-specific shortcuts. These effects can lead to inflated confidence without corresponding improvements in semantic understanding.

An additional challenge arises when interpretability tools are applied to models exhibiting these failure modes. Explanations generated for predictions influenced by memorization may still highlight linguistically or acoustically meaningful features, giving the appearance of valid reasoning. As a result, the presence of an explanation does not guarantee that the underlying decision process is causal or generalizable.

This work formulates the problem as understanding the interaction between training dynamics, confidence estimation, and explanation behavior in transformer-based text and audio models. The goal is to characterize when confidence becomes unreliable due to memorization and leakage, and to assess whether existing interpretability methods are capable of exposing these failure modes.

V. FROM DATA TO INFERENCE: SYSTEM-LEVEL ANALYSIS

To understand how data leakage and memorization influence model confidence and explanation behavior, it is necessary to examine the end-to-end pipeline through which transformer-based text and audio models are developed and deployed. Leakage does not typically originate from a single failure point; rather, it emerges through interactions across multiple stages of the data and training pipeline, often remaining undetected until inference-time behavior is observed.

At the data collection stage, large-scale text and audio datasets are commonly aggregated from diverse and heterogeneous sources. In text-based systems, this aggregation may include web documents, conversational logs, transcripts, or curated corpora that partially overlap in content or structure. In audio systems, datasets may contain repeated speakers, similar acoustic environments, or consistent recording setups. Even when explicit duplication is avoided, near- duplicate samples and correlated patterns can introduce implicit leakage by repeatedly exposing the model to highly similar information.

Preprocessing steps can further amplify these effects. Tokenization, normalization, segmentation, and transcription processes often standardize inputs in ways that reduce variability while preserving repeated structures. For audio data, transcription pipelines may introduce consistent linguistic patterns that mirror training annotations, while noise reduction and feature extraction can suppress variation that would otherwise encourage robust generalization. These transformations, while beneficial for optimization, can inadvertently strengthen shortcut signals that models exploit during training.

During training and fine-tuning, over-parameterized transformer models can readily absorb these patterns. Repeated fine-tuning on limited or overlapping datasets increases the likelihood that the model encodes specific examples or structural templates rather than learning generalizable representations. Because training objectives typically reward predictive accuracy without penalizing memorization, models can improve marginally on standard metrics while becoming increasingly reliant on familiarity with the training distribution.

At inference time, the consequences of these dynamics become visible in model behavior. Predictions may exhibit high stability across diverse inputs, confidence estimates may remain elevated even under ambiguity, and output variability may decrease. These behaviors create the appearance of reliability while masking underlying brittleness. Importantly, such failure modes are difficult to identify using conventional evaluation protocols that focus solely on accuracy or loss.

When interpretability methods are applied at this stage, they operate on representations shaped by the entire upstream pipeline. If leakage or memorization has influenced the model's internal decision process, explanations may faithfully reflect these internal states while still failing to capture true semantic reasoning. As a result, explanations can appear coherent and consistent even when the causal basis of predictions is misaligned with the intended task.

This system-level perspective highlights that data leakage is not merely a dataset issue, but a pipeline-wide phenomenon. Understanding how leakage propagates from data collection through preprocessing, training, and inference is essential for diagnosing over-confidence and evaluating the faithfulness of explanations in transformer-based text and audio models.

Data Collection and Preprocessing

The construction of training datasets plays a critical role in shaping the behavior of transformer- based text and audio models. In practice, large datasets are often assembled by aggregating data from multiple sources to achieve scale and diversity. While this approach improves coverage, it also introduces challenges related to repetition, overlap, and correlation that can contribute to implicit data leakage.

In text-based datasets, overlap can arise from repeated documents, paraphrased content, or shared structural templates across sources. Web-crawled corpora, conversational logs, and instructional datasets frequently contain near-duplicate passages or recurring linguistic patterns. Even when exact duplicates are removed, semantic similarity across samples can expose models to the same informational content multiple times, increasing the likelihood of memorization rather than abstraction.

Audio datasets present analogous but distinct challenges. Speech corpora often include re- peated speakers, consistent recording environments, or similar background conditions. When the same speaker appears across multiple samples, models may associate speaker-specific acoustic features—such as pitch range, cadence, or timbre—with particular transcriptions or labels. This introduces identity-specific cues that can serve as shortcuts during training, allowing models to rely on familiarity with the speaker rather than on semantic speech content.

Preprocessing pipelines can further reinforce these effects. Text normalization, tokenization, and segmentation reduce surface variability while preserving repeated structures. In audio pipelines, transcription, silence trimming, noise reduction, and feature extraction can standardize inputs in ways that suppress natural variation. While these steps are essential for stable training, they may inadvertently amplify repeated patterns and reduce the diversity of signals available to the model.

As a result, the training data presented to the model may contain less effective diversity than intended, even when nominal dataset size is large. This mismatch between apparent and functional data diversity creates conditions under which models can achieve high confidence through repeated exposure rather than through robust generalization. Such effects are difficult to detect during dataset construction and may only become evident through downstream behavior analysis.

Understanding how repetition, overlap, and identity-specific cues enter the pipeline at the data and preprocessing stages is therefore essential for diagnosing leakage-related failure modes. These early-stage factors can have cascading effects on training dynamics, inference-time confidence, and the reliability of subsequent explanations.

Training and Fine-Tuning Dynamics

Training and fine-tuning dynamics play a central role in determining whether transformer-based models develop robust representations or rely on shortcut learning. Modern transformers are highly over-parameterized, enabling them to fit complex patterns in data with relative ease. While this capacity is essential for modeling rich linguistic and acoustic structures, it also increases susceptibility to memorization when training conditions are not carefully controlled. Repeated fine-tuning on limited or highly correlated datasets can exacerbate this issue.

When the same data is reused across multiple training cycles, the model is repeatedly ex- posed to identical input–output mappings. Over time, this can reduce gradient diversity and encourage the model to encode specific examples or structural templates rather than learning generalizable features. As a result, improvements in training loss or validation accuracy may not correspond to meaningful gains in semantic understanding.

Shortcut learning emerges when models exploit spurious correlations that are predictive within the training distribution but do not reflect the intended task. In text models, such short- cuts may involve recurring phrasing patterns, formatting cues, or positional regularities. In audio models, shortcut signals can include speaker-specific characteristics, consistent acoustic environments, or transcription conventions. Because these cues are easier to learn than abstract semantic relationships, models may

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713617

preferentially rely on them during optimization.

Over-training amplifies these effects by reinforcing high-confidence internal representations associated with familiar patterns. As training progresses, the model's outputs often be- come more stable and less sensitive to input variation, even when such variation is semantically meaningful. This stability can be misinterpreted as improved reliability, while in reality it may indicate reduced adaptability and increased dependence on memorized structures.

Importantly, these dynamics are not always visible through standard evaluation metrics. Models may continue to exhibit marginal improvements in accuracy while becoming increasingly brittle under distributional shift or ambiguity. This disconnect highlights the need to analyze training behavior not only in terms of performance, but also in terms of representation diversity, sensitivity to perturbations, and confidence alignment.

Understanding how over-training and repeated fine-tuning influence shortcut learning and memorization is essential for interpreting downstream model behavior. These training-induced effects form a critical link between data characteristics and inference-time confidence, setting the stage for further analysis of prediction stability and explanation faithfulness

Inference-Time Behavior

The effects of data leakage, over-training, and shortcut learning become most apparent during inference, where model behavior is exposed outside the training loop. At this stage, transformer-based text and audio models often exhibit patterns that are not easily explained by standard performance metrics but have significant implications for reliability and trust.

One prominent behavior is confidence inflation. Models trained under repeated exposure to similar or overlapping data distributions tend to assign disproportionately high confidence to their predictions. This elevated confidence persists even when inputs are ambiguous, incomplete, or degraded. In such cases, confidence reflects familiarity with learned patterns rather than genuine certainty derived from robust inference. As a result, confidence scores lose their effectiveness as indicators of prediction reliability.

Prediction stability is another characteristic frequently observed at inference time. Over- trained models often produce highly consistent outputs across a wide range of inputs, including those that differ meaningfully at the semantic level. While stability is sometimes interpreted as robustness, excessive invariance can indicate reduced sensitivity to informative input variations. In text models, this may manifest as repeated phrasing or stylistic uniformity across responses. In audio models, similar transcriptions may be produced despite variations in speaker, noise level, or articulation.

Behavior under ambiguous or noisy inputs further highlights these issues. When presented with unclear text prompts or audio signals containing background noise, overlapping speech, or distortions, well-generalizing models are expected to reflect uncertainty through lower confidence or variable outputs. However, models influenced by memorization may instead produce confident and stable predictions, effectively masking uncertainty. This behavior is particularly concerning in real-world deployments, where input conditions are rarely ideal.

These inference-time patterns are difficult to detect through conventional validation procedures, which typically rely on clean test sets drawn from distributions similar to the training data. As a result, models may appear reliable during evaluation while exhibiting brittle or misleading behavior in practical settings. Understanding confidence inflation and prediction stability at inference time is therefore essential for diagnosing hidden failure modes that arise from earlier stages of the training pipeline.

By examining inference-time behavior in conjunction with training dynamics and data characteristics, it becomes possible to identify when model confidence and stability no longer correspond to meaningful understanding. This analysis provides a critical foundation for assessing the limitations of interpretability methods applied to such models.

Implications for Explainability

System-level effects arising from data leakage, over-training, and shortcut learning have direct implications for the reliability and faithfulness of explainability methods applied at inference time. Interpretability tools are typically designed to reflect a model's internal decision process; however, when that process itself is influenced by memorization or non-semantic cues, explanations may faithfully represent an undesirable or misleading reasoning pathway.

At inference time, explanations are generated based on learned representations that have been shaped by the entire upstream pipeline. If confidence inflation and prediction stability are driven by familiarity rather than understanding, explainability methods may highlight features that are correlated with memorized patterns instead of features that are causally relevant to the task. In such cases, explanations can appear coherent and intuitively satisfying while failing to reflect meaningful reasoning.

This limitation is particularly pronounced in transformer-based text and audio models. In text systems, attribution or attention-based explanations may consistently emphasize linguistically salient tokens even when predictions rely on repeated structural templates encountered during training. In audio models, saliency maps or feature attributions may focus on prominent acoustic regions without distinguishing whether the model's decision is based on semantic speech content or identity-specific cues such as speaker characteristics or recording conditions. A key challenge is that explanation stability is often interpreted as a positive signal. How- ever, when models are over-trained, explanations may remain stable precisely because the model's internal representations have become rigid. In such scenarios, stability reflects reduced sensitivity rather than reliable reasoning. Consequently, explanations may fail to change even when inputs are perturbed in ways that should alter the model's decision basis.

These observations suggest that explainability methods cannot be evaluated in isolation from training dynamics and inference-time behavior. Faithfulness should be assessed with respect to whether explanations respond appropriately to changes in the causal drivers of pre- dictions. Without this perspective, interpretability tools risk reinforcing misplaced trust by providing plausible narratives that do not expose underlying failure modes.

Understanding how system-level effects influence explanation behavior is therefore essential for using interpretability as a diagnostic instrument rather than a post-hoc justification. This motivates the need for evaluation frameworks that consider confidence alignment, sensitivity to perturbations, and consistency between explanations and genuine model reasoning.

VI. RESEARCH APPROACH

The research approach adopted in this work is diagnostic and analysis-driven rather than solution-centric. Instead of proposing new architectures or optimization strategies, the focus is on systematically studying how transformer-based text and audio models behave under conditions that are known to encourage memorization and data leakage. The goal is to understand why certain failure modes emerge, how they manifest during inference, and when existing tools fail to expose them.

At a high level, the approach treats model confidence and explanation behavior as observable signals that reflect underlying training dynamics. By analyzing these signals across con- trolled settings, it becomes possible to identify discrepancies between apparent performance and genuine generalization. This perspective shifts emphasis away from aggregate accuracy metrics toward behavioral characteristics such as confidence alignment, sensitivity to input variation, and explanation consistency.

The research strategy is structured around comparative evaluation rather than absolute performance measurement. Models trained under different data conditions—such as limited data, repeated fine-tuning, or increased overlap—are compared to examine how changes in training exposure affect confidence, stability, and interpretability. This comparative lens allows failure modes to be isolated without attributing them to specific architectures or hyperparameter choices.
Evaluation is designed to be intervention-based, relying on controlled input perturbations and behavioral probes rather than static test sets alone. By observing how predictions, confidence estimates, and explanations respond to small but meaningful changes in text or audio inputs, the analysis aims to distinguish between robust reasoning and shortcut-driven behavior. These interventions serve as diagnostic tools rather than stress tests, revealing the sensitivity of models to variations that should matter semantically.

Overall, this research approach prioritizes understanding over optimization. By framing over-confidence and explanation behavior as symptoms of deeper system-level effects, the work seeks to provide clarity on the limitations of current evaluation and interpretability practices. Such understanding is intended to inform future efforts in building safer and more trustworthy AI systems, rather than to offer immediate corrective mechanisms.

Behavioral Analysis of Model Confidence

Model confidence is analyzed as a behavioral signal that reflects how strongly a transformer model commits to its predictions under different input conditions. Rather than treating confidence as an inherent measure of correctness, this work examines how confidence evolves in response to training exposure, data overlap, and input variation. The analysis focuses on identifying conditions under which confidence becomes inflated or decoupled from meaningful uncertainty.

Confidence is quantified using standard probabilistic outputs derived from the model's predictive distribution, including maximum predicted probability and entropy-based measures. Low entropy outputs are interpreted as high confidence, while higher entropy indicates un- certainty. These measures are evaluated across inputs that vary in clarity, completeness, and similarity to the training distribution, allowing confidence behavior to be compared across con- trolled scenarios.

To assess confidence inflation, models are evaluated on inputs that are semantically ambiguous or partially degraded. In text-based settings, this includes prompts with incomplete context, paraphrased structures, or conflicting cues. In audio settings, inputs may contain back- ground noise, overlapping speech, or variations in speaker characteristics. Confidence inflation is identified when models assign consistently high confidence despite increased ambiguity or reduced semantic signal.

Prediction stability is analyzed by measuring the consistency of model outputs across small perturbations of the same input. Stability is assessed both at the output level—whether predicted labels or transcriptions remain unchanged—and at the confidence level—whether confidence values remain invariant despite perturbations that should meaningfully affect interpretation. Excessive stability is treated as a potential indicator of reliance on memorized patterns rather than adaptive inference.

By comparing confidence behavior across models trained under different data conditions, this analysis seeks to characterize how training dynamics influence uncertainty expression. The emphasis is not on determining whether confidence is numerically correct, but on under- standing whether confidence behaves sensibly in response to uncertainty-inducing inputs. This behavioral perspective provides a foundation for interpreting confidence as a diagnostic signal rather than as a direct measure of reliability.

Perturbation-Based Evaluation

Perturbation-based evaluation is used as a primary diagnostic tool to probe the sensitivity of transformer-based text and audio models to meaningful changes in their inputs. Rather than relying solely on static test sets, this approach introduces controlled variations that are designed to alter semantic content or signal quality while preserving the overall structure of the input. The objective is to observe how predictions and confidence estimates respond to these changes.

In text-based models, perturbations include paraphrasing, token substitutions, reordering of non-critical phrases, and partial removal of contextual information. These modifications are designed to preserve high-level meaning while altering surface-level structure. Sensitivity is assessed by measuring changes in predicted outputs and associated confidence values. A model exhibiting robust reasoning is expected to adjust its confidence or output when semantic cues are weakened or altered.
For audio models, perturbations are applied at the acoustic level. These include the intro- duction of background noise, variation in speaking rate, changes in pitch, and segmentation alterations. Additional perturbations may involve speaker changes or mixing speech segments with different acoustic profiles. These interventions are intended to disrupt non-semantic cues while preserving core linguistic content, allowing analysis of whether predictions depend on speech understanding or on identity-specific patterns.

Sensitivity analysis focuses on two key aspects: output variability and confidence response. Output variability captures whether predictions change appropriately when perturbations affect meaning or signal clarity. Confidence response measures whether uncertainty increases under degraded or ambiguous conditions. A lack of sensitivity—manifested as unchanged predictions and stable confidence across perturbations—suggests reliance on memorized or shortcut signals.

By applying identical perturbation strategies across models trained under different data conditions, this evaluation enables comparative analysis of behavioral robustness. The results are interpreted diagnostically rather than as indicators of superiority, with the aim of identifying when models fail to reflect uncertainty in the presence of meaningful input changes. This perturbation-based perspective provides a principled framework for linking training dynamics to inference-time behavior

Explanation Sensitivity and Faithfulness

Explanation sensitivity is evaluated by examining whether interpretability outputs respond meaningfully to changes in the causal basis of model predictions. Rather than assessing explanations solely on clarity or visual coherence, this analysis treats explanations as behavioral signals that should vary when the underlying reasoning process of the model changes. Faith- fulness, in this context, is defined as the degree to which explanations reflect the true factors driving a prediction.

The evaluation strategy aligns explanations with the perturbation-based analysis of model behavior. For a given input, explanations are generated alongside predictions and confidence estimates. Controlled perturbations are then applied to the input, and the resulting changes in predictions, confidence, and explanations are observed jointly. If a perturbation alters the semantic content or removes shortcut cues, a faithful explanation is expected to change correspondingly.

In text-based models, explanation sensitivity is examined through attribution patterns over tokens or phrases. When perturbations disrupt semantic cues, explanations should shift to- ward alternative relevant tokens or exhibit reduced attribution strength. If explanations remain largely invariant despite changes in prediction behavior or confidence, this suggests that the interpretability method is capturing surface-level correlations rather than causal reasoning.

In audio models, explanation faithfulness is evaluated using saliency or feature-attribution maps over time–frequency representations. Perturbations that remove or distort identity-specific acoustic cues are expected to alter explanation patterns if the model genuinely relies on semantic speech content. Stable explanations under such perturbations may indicate that the model's decisions are driven by non-semantic shortcuts that are not exposed by the explanation method.

A critical aspect of this analysis is the comparison between explanation stability and prediction stability. While stable explanations may appear desirable, excessive invariance can signal a lack of sensitivity to meaningful input changes. Explanations that do not respond to shifts in confidence or output behavior risk functioning as post-hoc narratives rather than diagnostic tools.

By evaluating explanations through the lens of sensitivity and alignment with causal changes, this work aims to characterize when interpretability methods succeed and when they fail to expose hidden failure modes. This perspective emphasizes faithfulness over plausibility and positions explainability as an evaluative instrument for understanding model behavior rather than as a mechanism for justification.

VII. EXPERIMENTAL DESIGN

The experimental design is structured to systematically analyze how training conditions influence model confidence, prediction stability, and explanation behavior. Rather than optimizing for peak performance, the experiments are designed to isolate behavioral effects arising from data overlap, repeated fine-tuning, and controlled input perturbations. All experiments are con- ducted in controlled settings to ensure that observed behaviors can be attributed to specific design choices rather than confounding factors.

The overall setup follows a comparative framework. Multiple instances of transformer- based text and audio models are trained under varying data exposure conditions, including differences in dataset size, degree of overlap, and number of fine-tuning iterations. These models are then evaluated using identical inference-time protocols, allowing behavioral differences to be analyzed consistently

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713617

across conditions.

Experiments are organized into three stages. First, baseline models are evaluated on clean inputs drawn from distributions similar to their training data to establish reference confidence and explanation behavior. Second, controlled perturbations are applied to inputs to assess sensitivity of predictions, confidence estimates, and explanations. Third, comparative analysis is performed across models trained under different conditions to identify systematic trends in behavior.

To minimize confounding effects, all models within a given experiment share the same architecture, optimization settings, and evaluation pipeline. Variations are introduced only at the data and training-exposure level. This design choice ensures that observed differences in behavior are attributable to training dynamics rather than architectural or implementation differences.

Importantly, the experimental setup emphasizes reproducibility and interpretability over scale. Experiments are designed to be feasible within constrained computational budgets while still capturing meaningful behavioral patterns. This allows for iterative refinement of experimental conditions and careful qualitative inspection of model outputs, confidence signals, and explanations.

By structuring experiments as diagnostic probes rather than performance benchmarks, the design supports a deeper understanding of how and when transformer-based models exhibit over-confidence and explanation failure modes. The results of these experiments are intended to inform evaluation practices rather than to establish new state-of-the-art performance

Datasets and Model Selection

The experimental analysis is conducted using a combination of text and audio datasets selected to support controlled investigation of data exposure, repetition, and limited-data regimes. Rather than relying on large benchmark corpora alone, the datasets are chosen to allow precise manipulation of training conditions and to observe behavioral changes under constrained set- tings.

For text-based experiments, datasets consist of curated textual corpora designed to vary in size, overlap, and structural diversity. These include subsets constructed to simulate limited- data scenarios, where the number of unique samples is intentionally restricted, as well as set- tings with controlled repetition or paraphrased content. Such configurations enable analysis of how repeated exposure to similar linguistic patterns influences confidence and explanation be- havior. The datasets are preprocessed using standard tokenization and normalization pipelines to reflect common training practices while maintaining consistency across experimental conditions.

For audio-based experiments, speech datasets are selected to enable fine-grained control over speaker identity, acoustic conditions, and transcription patterns. The primary focus is on datasets with a manageable duration and a limited number of speakers, allowing repeated fine- tuning and targeted perturbation analysis. Audio samples are segmented into shorter utterances with aligned transcriptions, reflecting typical supervised training setups used in speech recognition models. Controlled subsets are created to vary speaker overlap and recording conditions across training runs.

Model selection prioritizes widely used transformer architectures representative of modern text and audio systems. For text, pre-trained transformer language models are used as the base, while for audio, transformer-based speech recognition models are selected. All models are fine-tuned under identical optimization settings within each modality to ensure comparability. Architectural modifications are avoided to keep the focus on training dynamics and behavioral effects rather than model design.

By combining controlled dataset construction with consistent model selection, this experimental setup enables systematic comparison across training conditions. The emphasis on limited-data and repeated-exposure scenarios reflects real-world fine-tuning practices, providing a realistic context for analyzing over-confidence, memorization, and explanation faithful- ness

Evaluation Metrics

The evaluation framework employs a set of complementary metrics designed to capture behavioral characteristics of transformer-based text and audio models beyond conventional accuracy measures. These metrics focus on confidence alignment, prediction stability, and explanation sensitivity, enabling a structured analysis of how model behavior changes under varying training and input conditions.

Confidence alignment is assessed by examining the relationship between model confidence and input uncertainty. Metrics include maximum predicted probability and entropy of the out- put distribution, evaluated across inputs with varying degrees of ambiguity or degradation. Alignment is analyzed by measuring whether confidence appropriately decreases as semantic clarity or signal quality is reduced. Discrepancies between confidence levels and input uncertainty are used as indicators of confidence inflation.

Prediction stability is measured by analyzing the consistency of model outputs under con- trolled input perturbations. For classification tasks, stability is quantified as the proportion of inputs for which predicted labels remain unchanged across perturbations. For sequence- generation tasks, such as text generation or speech transcription, stability is evaluated using similarity measures between outputs, including token-level overlap or edit distance. Excessive stability under meaningful perturbations is interpreted as potential reliance on memorized patterns rather than adaptive inference.

Explanation sensitivity evaluates whether interpretability outputs respond to changes in the causal basis of predictions. Sensitivity is assessed by comparing explanation maps or attribution scores before and after input perturbations. Metrics include overlap ratios between attribution distributions, changes in ranked feature importance, and shifts in saliency concentration. Low sensitivity—where explanations remain largely invariant despite changes in predictions or confidence—is treated as evidence of limited faithfulness.

These metrics are analyzed jointly rather than in isolation. For example, stable predictions accompanied by unchanged confidence and invariant explanations under perturbation are interpreted differently from cases where confidence decreases and explanations shift. This multi-dimensional evaluation approach allows nuanced characterization of model behavior and supports a diagnostic interpretation of confidence and explainability signals

Feasibility and Practical Considerations

The proposed research is designed with practical feasibility as a core consideration. Rather than relying on large-scale training from scratch, the experimental setup leverages pre-trained transformer models and controlled fine-tuning procedures. This approach significantly reduces computational requirements while still enabling meaningful analysis of training dynamics, confidence behavior, and explanation faithfulness.

To maintain a controlled scope, experiments focus on limited-data and repeated-exposure settings that are representative of common real-world fine-tuning practices. Dataset sizes, number of training iterations, and perturbation strategies are intentionally constrained to allow systematic comparison without introducing excessive variability. This controlled design ensures that observed behavioral effects can be attributed to specific experimental conditions rather than uncontrolled scaling factors.

Iterative experimentation is central to the research methodology. Experiments are structured to allow gradual refinement based on intermediate observations, enabling targeted exploration of specific failure modes as they emerge. This iterative process supports careful qualitative inspection of model outputs, confidence signals, and explanations, complementing quantitative metrics with human-in-the-loop analysis.

From an infrastructure perspective, the experiments are feasible on modest GPU resources commonly available in academic and industrial research environments. Training and evaluation pipelines are designed to be reproducible and modular, allowing individual components—such as perturbation strategies or explanation methods—to be adjusted independently. This modular- ity also supports extensibility, enabling future experiments to build upon the same framework. Overall, the emphasis on computational efficiency, controlled scope, and iterative analysis ensures that the proposed research can be executed reliably within practical constraints. This design balances rigor with feasibility, enabling focused investigation of model behavior without the need for excessive computational scale.

VIII. CONCLUSION AND FUTURE DIRECTIONS

his work investigates the behavioral consequences of data leakage, repeated fine-tuning, and shortcut

© JAN 2026 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713617

learning in transformer-based text and audio models, with a particular focus on model confidence and explanation faithfulness. Rather than evaluating models solely through performance metrics, the study frames confidence, stability, and interpretability as diagnostic signals that reveal deeper properties of model behavior. Through this lens, the research emphasizes understanding how models behave under uncertainty, ambiguity, and distributional variation.

The analysis highlights that over-confidence can emerge even when improvements in se- mantic understanding are limited. Under conditions of repeated exposure or overlapping data, models may exhibit stable predictions and confident outputs that mask underlying brittleness. These behaviors challenge the assumption that confidence is a reliable proxy for correctness and underscore the need for behavioral evaluation beyond standard benchmarks.

A central insight of this work is that explainability methods must be interpreted with caution in such settings. Explanations can remain coherent and visually persuasive even when predictions are influenced by memorized patterns or non-semantic cues. This disconnect raises important questions about the faithfulness of interpretability tools and their ability to surface hidden failure modes in modern AI systems.

At the same time, this study is intentionally scoped and exploratory. The analysis is con- ducted in controlled experimental settings using limited data regimes and specific model families. The findings are not intended to generalize across all architectures or applications, but rather to illuminate patterns that warrant closer examination. By focusing on diagnostics and behavioral characterization, the work aims to contribute to a deeper understanding of how trust in AI systems can be misplaced when evaluation focuses narrowly on performance.

Looking forward, this line of inquiry opens several directions for future research. Further work could explore broader model families, more diverse data modalities, and alternative interpretability techniques. Most importantly, it motivates the development of evaluation frameworks that treat confidence and explanations as dynamic signals, helping to ensure that powerful models remain transparent, reliable, and aligned with human expectations.