From Performance to Process: Temporal Information Dynamics in Language Model Fine-tuning

Frida Hæstrup[1,2] and **Ross Deans Kristensen-McLachlan**[2]

[1]Dept. of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark
[2]Center for Humanities Computing, Aarhus University, Aarhus, Denmark
frihae@clin.au.dk, rdkm@cc.au.dk

Abstract

Large language model performance has advanced rapidly in recent years, driven by technical improvements in areas like model architecture, scaling, and reinforcement learning. However, much of our understanding of these models remains rooted in static evaluations calculated post-training. While informative, these snapshots offer limited insight into how models learn, adapt, and transform internally during training, overlooking dynamic processes and representational shifts that occur throughout fine-tuning, potentially concealing important aspects of model behavior. We aim to contribute to ongoing efforts to open the 'black box' of language models by analyzing temporal information dynamics during fine-tuning. Our findings suggest that tracking these internal dynamics demonstrates both training-regime-specific and task-specific differences in learning and may eventually contribute to applications such as change point detection or adaptive training strategies. Ultimately, this work moves toward a more nuanced, mathematical formulation of what learning does to a model, highlighting the constant flux of representational change that underlies seemingly stable performance improvements.

1 Introduction

The rapid development and widespread deployment of large language models (LLMs) have amplified interest in understanding how these models function internally. In pursuit of improved model performance and generalization, the development of pre-trained LLMs has led to models that are increasingly becoming larger and more complex (Simon, 2021; Brown et al., 2020). Such complexity, often driven by millions or even billions of parameters, enables these models to capture and learn intricate patterns within the training data, allowing them to achieve state-of-the-art results across a wide array of tasks (Devlin et al., 2019; Wang et al., 2018; Rozière et al., 2024; Wang et al., 2020).

However, this power comes at a significant cost: it obscures the internal mechanisms by which models arrive at their predictions, rendering the path from input to output difficult to interpret and explain. As these models are increasingly adopted in sensitive and high-stakes domains, the need for transparency into their internal processes becomes not just desirable, but essential (Hassija et al., 2024; Embarak, 2023; Chen et al., 2025). To better understand what these models are actually learning — and how their internal states evolve during training— we must look beyond static evaluations and examine the learning process itself. Standard evaluation metrics such as accuracy, perplexity, or F1-score provide only static snapshots of model behavior. These metrics reflect *what* a model achieves but offer little insight into *how* it learns.

In this paper, we propose analyzing the temporal dynamics of learning in language models within an information-theoretic framework (MacKay, 2002), conceptualizing a model's internal state as a dynamic information system (**?**). Rather than focusing solely on final performance, we track how internal representations evolve during fine-tuning. This allows us to characterize learning as a continuous sequence of representational shifts, offering a more granular and process-oriented perspective on model behavior.

Our proposed framework builds on a growing body of research that has used information theory to study the evolution of complex, dynamic systems. In particular, several studies have modeled cultural and linguistic phenomena by analyzing the balance between how much new information is being introduced and that information's longevity within the system (Barron et al., 2018; Nielbo et al., 2021a,b; Vrangbæk and Nielbo, 2021; Wevers et al., 2021; Krisensen-McLachlan et al., 2024). These studies used windowed relative entropy to quantify the **novelty** of a system - the extent to which a given time period diverges from preced-

ing time periods - and the **resonance** of a system, which captures how information persists over time.

We extend this framework to the context of deep learning by treating the internal states of a language model as a dynamic information system. We investigate the evolution of internal information structure in models from the English BERT family (Devlin et al., 2019) as they are fine-tuned across various classification tasks. Through a series of controlled experiments, we continuously extract internal representations from different BERT models throughout the fine-tuning process. We adopt an exploratory approach, examining whether tracking the dynamics of internal representations over time can reveal novel insights into the mechanisms of learning within these models.

We argue that this approach offers a rich perspective on what it means for a model to learn and opens the door to future applications, such as tracking learning trajectories, identifying shifts in representational focus, or detecting meaningful change points during training. Ultimately, we aim to bridge the gap between surface-level performance and deeper representational change, providing insight into the temporal structure of learning itself.

1.1 Related Work

Prior research has explored how fine-tuning affects the internal structure of transformer-based models such as BERT. A common approach involves probing internal layers to identify which aspects of the model change during adaptation to downstream tasks (Phang et al., 2021; Hao et al., 2020; Merchant et al., 2020; Zhou and Srikumar, 2022; Voita and Titov, 2020; Liu et al., 2019; Tenney et al., 2018; Voita and Titov, 2020). Hao et al. (2020) employ divergence-based measures to track shifts in attention patterns and find that fine-tuning primarily alters the attention modes of higher layers. This is consistent with observations from Merchant et al. (2020) who use probing classifiers and ablation experiments to show that representational change during fine-tuning is concentrated in upper layers. Furthermore, they find variations in this effect across fine-tuning tasks. For example, tasks such as dependency parsing produce deeper representational shifts than tasks like natural language inference or reading comprehension.

Further analyses have investigated the spatial structure of learned representations (Coenen et al. (2019); Hernandez and Andreas (2021). Comparing the spatial structure of class-level embeddings before and after fine-tuning, Zhou and Srikumar (2022) observe that class representations are pushed further apart in the embedding space after fine-tuning, even in cases where the classes were already linearly separable. Extending the findings of Merchant et al. (2020), they also report that while higher layers change more than lower ones, these changes preserve structural similarity with the pre-trained model, suggesting that fine-tuning reshapes but does not fully overwrite earlier representations.

While these studies offer valuable insight into how models change across fine-tuning, they are typically limited to static comparisons between pre-trained and post-trained states. In contrast, our work adopts a dynamic perspective, examining internal representations at every step *during* the fine-tuning process. Moreover, rather than analyzing intermediate encoder layers, we focus on prediction-layer outputs, treating class-level output vectors as a dynamic system whose evolution reflects learning in real time. This allows us to capture transient changes and transitions that static snapshots may miss, offering a more granular view of representational dynamics during training.

2 Methods

We base our analysis on information signals extracted from 24 experiments: four pre-trained large language models fine-tuned on three classification tasks under two conditions. Details of this process are laid out in the following sections.[1]

2.1 Model architectures

We fine-tune four different pre-trained BERT-style models, namely **BERT** (Devlin et al., 2019), **distilBERT** (Sanh et al., 2020), **roBERTa** (Zhuang et al., 2021), and multilingual BERT (**mBERT**) (Devlin et al., 2019). The models are all core models that have been trained across many language-understanding tasks. Each model is based primarily on the BERT architecture, although they each display variations across different parameters such as size or training regime, allowing for a range of possible comparisons across models. An overview of the key differences across model types can be found in Appendix A.2. The pre-trained model weights of

[1]The code-base for the project can be found at https://github.com/frillecode/BERT-infodynamics

all four models were retrieved from HuggingFace.[2]

2.2 Classification tasks

We fine-tune the above-mentioned pre-trained models across three different language classification tasks from the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). GLUE comprises a collection of resources for evaluating the performance of natural language understanding systems across a wide range of linguistic tasks. GLUE consists of nine different language understanding tasks, each built on established English-language text datasets, that are widely accepted as standard benchmarks for assessing how well models can understand and process natural language (Devlin et al., 2019; Radford et al., 2019). In the present study, a subset of three tasks from the GLUE benchmark is used, namely:

- **MNLI**: The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018).
- **MRPC**: The Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005).
- **SST-2**: Stanford Sentiment Treebank (Socher et al., 2013).

The choice of using a subset of tasks is motivated by the following reasons. Firstly, the GLUE benchmark is typically used to assess how well models generalize across tasks and text genres, often with the ultimate goal of driving the development of robust natural language understanding systems (Wang et al., 2018). In contrast, this study seeks to explore the underlying processes of the models as they learn rather than assessing their final performance. Secondly, the experiments in this study make for 24 different fine-tuning processes and subsequent analyses, with the windowed relative entropy calculation adding substantial computational load. Thirdly, the choice of these tasks ensures that the study encompasses both binary and multi-class classification problems, as well as different dataset sizes. Furthermore, the tasks cover a wide range of linguistic phenomena as they represent each of the three general categories of the benchmark (Wang et al., 2018). As such, the tasks provide a sufficient variety of linguistic challenges to, within the scope of this study, explore how models process and learn throughout different natural language understanding tasks.

[2] https://huggingface.co/google-bert/bert-base-uncased
https://huggingface.co/distilbert/distilbert-base-uncased
https://huggingface.co/FacebookAI/roberta-base
https://huggingface.co/google-bert/bert-base-multilingual-uncased

Hyperparameter	Values
Batch size	16, 32
Learning rate	$2e^{-5}$, $3e^{-5}$, $5e^{-5}$
N epochs	2, 3, 4

Table 1: Search space for hyperparameter optimization.

2.3 Training procedures

The fine-tuning process of each model on each task is carried out under two conditions differing in parts of the training setup.

In the **fixed condition**, the hyperparameters of the training process are kept fixed across all experiments to allow for a more direct comparison. The models are trained for 5000 steps using a batch size of 64. All other hyperparameters are kept at default values.

In the **optimized condition**, hyperparameter optimization is incorporated in the training process to explore the effects of optimizing the models' learning process to the task. We perform a simple grid search over pre-defined values for batch size, learning rate, and number of epochs. We use the search space recommended in the original BERT paper (Devlin et al., 2019), as seen in Table 1. For each experiment, we run a total of 10 trials. Based on this, we define the best configuration for each experiment. These can be found in the Appendix A.1. All other hyperparameters are kept at default values.

In both conditions, a standard pipeline for fine-tuning machine-learning models was employed using the HuggingFace Transformers library (v. 4.42.4) (Wolf et al., 2020), and the datasets for the different GLUE tasks were retrieved using the Datasets class. All models are fine-tuned using a Cross-Entropy loss function, and the standard training and validation splits are retrieved automatically upon accessing the datasets from GLUE. All analysis is performed using Python (v. 3.12.3).

2.4 Feature extraction

During the fine-tuning process for each of the experiments, we save the logits at every training step by extracting the output of the last layer of the neural network. By passing the logits through the softmax function, they are converted to vectors rep-

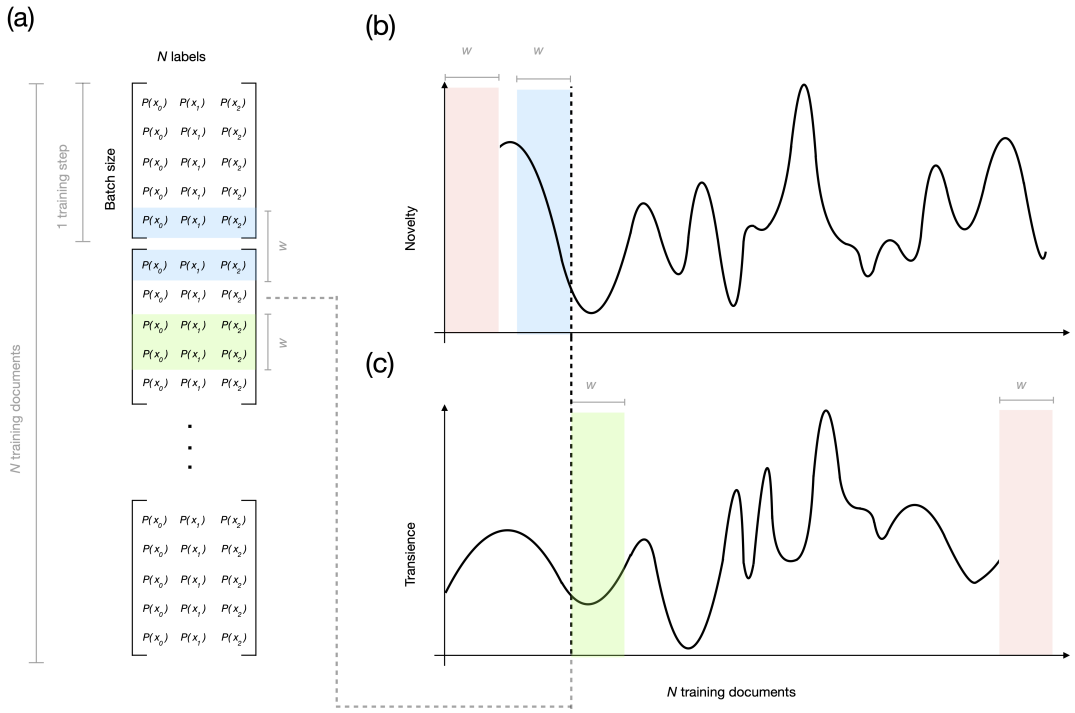


Figure 1: Overview of pipeline for extracting information signals from logits. Fictional example for a classification task with 3 classes and a window size of $w = 2$. **(a)** illustrates the matrix with probability scores based on which the information signals will be extracted. **(b)** illustrates a novelty signal, with the blue area representing the window size within which it is calculated and the light red area representing the documents that are removed. **(c)** illustrates a transience signal, with the green area representing the window size within which it is calculated and the red area representing the documents that are removed.

resenting a probability distribution across labels. As the fine-tuning process continues over training steps, the resulting matrix becomes a temporally sorted series of probability distributions representing the model's predictions. Since these matrices (one for each experiment) capture how the models' predictions evolve over time, this can be used as a proxy reflecting the learning process as the models update their internal representations in response to the data. These probability distribution matrices hence serve as the input from which to extract information signals, as described in the following section. A visual representation of the process of extracting the information signals from the logits can be seen in Figure 1.

2.5 Information dynamics

Based on the temporally sorted probability scores for each of the experiments, we employ methods from information theory to extract information signals (novelty, resonance, transience). Using windowed relative entropy, we can measure the similarity (or 'surprise') between the information patterns in a series of probability distributions (Cover and Thomas, 2006). **Novelty** serves as a measure of how surprising the probability distribution patterns in a document are given past documents, **transience** measures the extent to which those patterns persist in future documents, and **resonance** measures the degree to which patterns in future documents conform to the novelty.

Information signals are extracted for each document using a window size of 160 (w=160). A document in this context refers to a document from the training data (i.e. an input sentence) of the given GLUE task that the model sees during fine-tuning. As such, a window size of 160 means that the information signals are extracted by comparing the model's representation of the current input sentence to the previous 160 input sentences and the

following 160 input sentences.

For the implementation of relative entropy, Jensen-Shannon divergence was used:

$$JSD(s^{(j)} \mid s^{(k)}) = \frac{1}{2}D(s^{(j)} \mid M) + \frac{1}{2}D(s^{(k)} \mid M) \quad (1)$$

where $M = \frac{1}{2}(s^{(j)} + s^{(k)})$ and D is the Kullback-Leibler divergence (Cover and Thomas, 2006):

$$D(s^{(j)} \mid s^{(k)}) = \sum_{i=1}^{K} s_i^{(j)} \times \log_2 \frac{s_i^{(j)}}{s_i^{(k)}} \quad (2)$$

Novelty ($\mathcal{N}$) is defined as a document $s^{(j)}$'s reliable difference from past documents $s^{(j-1)}, s^{(j-2)}, \ldots, s^{(j-w)}$ in window w:

$$\mathcal{N}_w(j) = \frac{1}{w}\sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j-d)}) \quad (3)$$

Resonance ($\mathcal{R}$) is defined as the degree to which future documents $s^{(j+1)}$, $s^{(j+2)}$, ...,$s^{(j+w)}$ conform to the Novelty of document $s^{(j)}$:

$$\mathcal{R}_w(j) = \mathcal{N}_w(j) - \mathcal{T}_w(j) \quad (4)$$

where $\mathcal{T}$ is the Transcience of $s^{(j)}$:

$$\mathcal{T}_w(j) = \frac{1}{w}\sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j+d)}) \quad (5)$$

Given the definitions outlined above, we can see that these information theoretic measures neatly translate into easily interpretable descriptions of the learning process over time. **Novelty** in our setup describes by how much the predictions of a given model at a particular training step differ from those which have come immediately before, indicating a substantial shift in model behavior. **Resonance**, on the other hand, considers to what extent this novelty persists in the system during subsequent training steps. This further allows the characterization of individual (per-experiment-level) signals as information dynamics profiles based on internal representation change. These information patterns can then be analyzed to see how the dynamics of the internals of a language model system evolve over time (i.e. during fine-tuning).

2.6 Signal processing

Due to the granularity of the experiments, the generated information signals are very long (as determined by batch size multiplied by number of training steps). As such, some processing must be done to analyze and interpret the signals meaningfully.

First, the first 160 and last 160 (i.e., the window size) documents are removed from the novelty and resonance signals. Second, following existing research into information dynamics (Nielbo et al., 2021a; Wevers et al., 2021; Nielbo et al., 2021b), non-linear adaptive filtering is performed to extract global trends in the novelty and resonance signals. In broad terms, the algorithm identifies a globally smooth trend signal by 'stitching' together locally best-fitting polynomials in overlapping partitions of the time series, allowing identification of broad trends while preserving local variations within the data. Following Riley et al. (2012), we define the span value (size of the partitions) by visually inspecting the results across a range of values to identify the best fit to extract the globally smooth trend across the different signals. In this study, this is done by comparing the smoothed signal produced by adaptive filters with varying span values to a moving average (see Appendix C.1 for an example). Based on this procedure, the span value for the partitions is set to 92.

3 Results

Figure 2 depicts the smoothed, normalized novelty and resonance signals for the 12 experiments in the fixed group (2a) and the optimized group (2b). Across both groups, the resonance signals show more frequent and periodic oscillations compared to the novelty signals. The trajectories of both novelty and resonance signals in the fixed group show a higher degree of similarity across experiments compared to those of the optimized group.

In the fixed group (Figure 2a), both novelty and resonance signals appear smoother and more coherent with slower oscillations, and we observe visible patterns that correlate across the different experiments. The novelty signals show closely aligned trajectories during the initial training phase, but begin to diverge after seeing approximately 20% of the documents. The divergence is apparent in the magnitude of the fluctuations, with some models having more or less pronounced variance. However, the overall direction of the changes - either increasing or decreasing - remains largely consistent across experiments. Though more variable from the outset, the resonance signals show similar patterns of divergence over time; they exhibit somewhat aligned trajectories in the initial training phase, but the magnitude of the oscillations grows more unsynchronized as the training progresses.

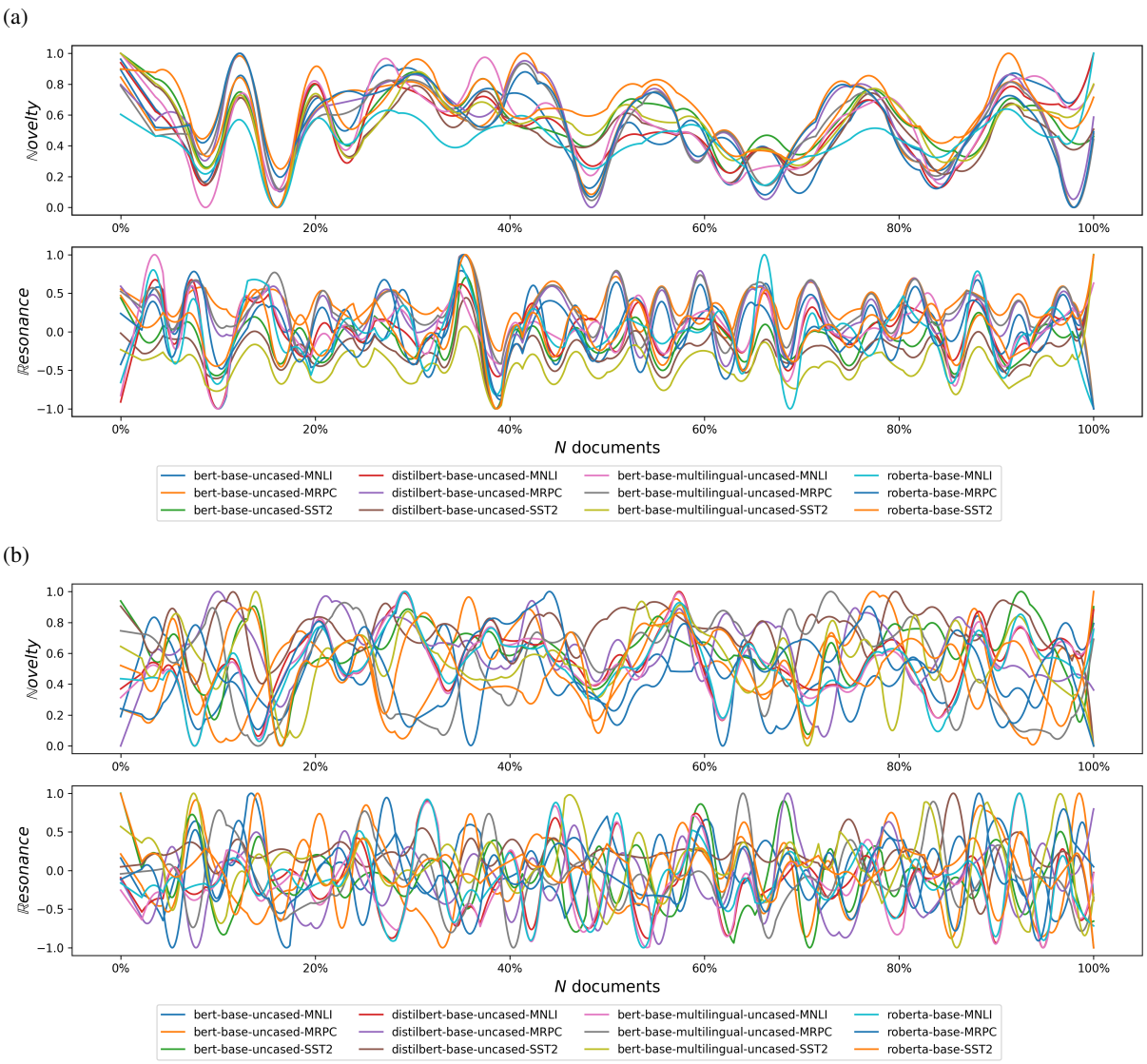


Figure 2: Normalized, smoothed novelty and resonance signals for experiments in (a) the fixed group and (b) the optimized group. The signals are visualized over fine-tuning time with the percentage of training documents seen by the model on the x-axis.

The situation is markedly different in the optimized group (Figure 2b), where the signals are more chaotic and noisy overall, with more rapid fluctuations and less apparent structure. Both novelty and resonance signals show high variability from the beginning of training and remain unsynchronized throughout. We observe less alignment across experiments, with more rapid fluctuations and no clear common direction of changes between experiments.

Figure 3 displays the novelty and resonance signals of the fixed group grouped by fine-tuning task. We can observe clear task-specific patterns in the trajectories of the signals, with high within-task alignment, especially for the MRPC task (Figure 3b). The same is not evident for the optimized group, nor do we find visible shared patterns in either group when grouping signals by model type (see Appendix B.1).

4 Discussion

Our findings reveal variations in information dynamics during the learning process across all experiments, suggesting that BERT models process and handle new information in distinct ways as they learn. Most notably, we observe a high degree of similarity in the signals from experiments in the fixed group. Despite divergences in magnitude, the overall directions of the changes in novelty and resonance remain largely consistent across exper-

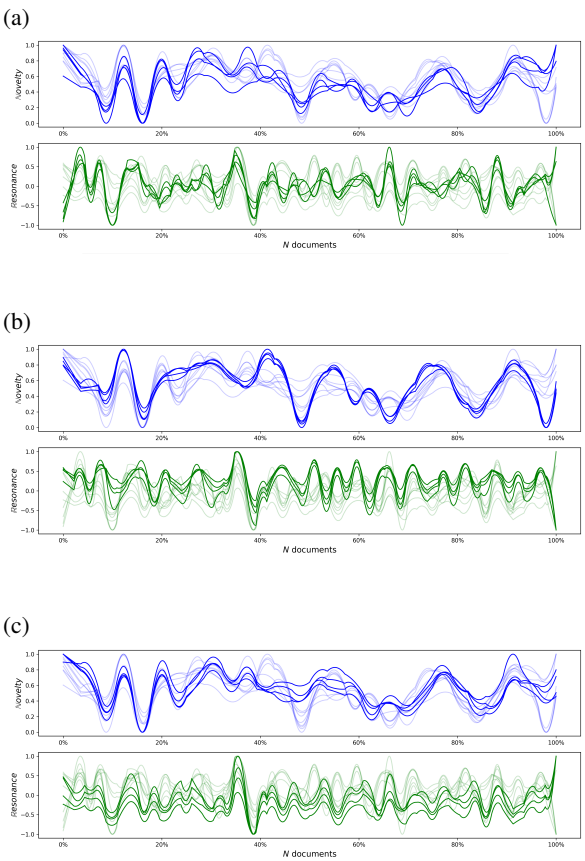


Figure 3: Normalized, smoothed novelty (blue) and resonance (green) signals for experiments in the fixed group grouped by task. The highlighted lines depict the signals from fine-tuning on the (a) MNLI, (b) MRPC, and (c) SST-2 task, respectively. The transparent lines show the remaining signals, i.e., signals from those tasks not highlighted in each plot.

iments, suggesting somewhat stable information structure and shared underlying trends in the evolution of those signals over time. This contrasts with the optimized group, where the signals show more variability and noise across experiments, implying less consistency in information dynamics in that group. Contrary to previous methodologically related research in other data domains (e.g. Nielbo et al. (2021b) and Vrangbæk and Nielbo (2021)), this study does not find clear temporal change points in the information signals that correspond to key events, such as shifts in learning curves (see Appendix B.2).

The consistency we observe in the novelty signals in the fixed group suggests that, across model types, new information is being integrated in a stable and comparable way. The resonance curves show similar trends across models and tasks, indicating that when new information is introduced, its influence tends to persist consistently across experiments. This illustrates a shared structure of learning dynamics, where the models steadily adapt to incoming training data in a similar manner. In contrast, while generally achieving better classification task performance (see Appendix B.2), the optimized group exhibits less consistent information integration. Frequent and high fluctuations in novelty signals in this group suggest that the models are encountering more abrupt changes in their internal representations, likely due to different optimal hyperparameters (e.g. learning rate or batch size). Resonance signals are also less uniform, implying that the influence of novel information on future representations is less predictable and more specific to the given experiment. These observations suggest that hyperparameter optimization introduces variability in how models process and retain information, possibly due to faster convergence, more aggressive adaptation, and divergent learning regimes across runs. However, it remains unclear whether these fluctuations reflect meaningful learning phenomena — such as adaptive capacity or sensitivity to task complexity — or are artifacts introduced by tuning. Distinguishing between the two remains a challenge and motivates future work involving finer-grained ablation studies and statistical analysis. Overall, these findings indicate that stability in training procedure (i.e., fixed hyperparameters) leads to more uniform information dynamics, while optimization increases variability in novelty and resonance, even if it may improve downstream task performance.

These results are aligned with prior work investigating fine-tuning dynamics in BERT models. For instance, as previously introduced, Hao et al. (2020) use divergence-based methods to assess shifts in attention patterns and find that fine-tuning affects the higher layers of BERT more substantially than lower layers. Their findings suggest that learning-induced changes tend to concentrate in specific architectural regions of the model and vary by downstream task — a conclusion that aligns with our observation that models under fixed training conditions exhibit consistent internal changes with observable task-specific patterns, while those under optimized regimes display greater variance. Given these earlier findings, the present study's focus on the prediction layer is a natural starting point for capturing salient representational changes during fine-tuning. However, while this level offers

tractable insight into the model's learning behavior, it may not fully capture the dynamics occurring in earlier layers. Extending the analysis to intermediate representations could provide a more nuanced understanding of how internal structures evolve across the network.

While the results may already conform with expert intuition about how models are learning over time, the explicitly information-theoretic approach can provide a new vocabulary and conceptual framework for explaining how and why certain learning dynamics occur during fine-tuning on different tasks.

For example, Figure 3 illustrates the information signals with fixed hyperparameters grouped by classification task. For all three tasks, there is an initial spike in novelty around 10% into training, indicating that significant, consecutive representational changes are occurring at this stage. This may reflect initial learning in the early stages of fine-tuning where the models make more sporadic or uninformed predictions, thus increasing novelty. Subsequently, novelty decreases, suggesting that the changes become more permanent, perhaps as the models have learned useful patterns from the training data. This is notably followed by a series of oscillations that manifest themselves consistently within each task, perhaps reflecting episodic shifts in representations as the models adjust to task-specific data.

The resonance signals show similarly pronounced regularity with structured, repeating resonance peaks, especially for the MRPC task (Figure 3b). This periodicity might emerge from uniform training dynamics across runs with fixed training regimes; the same types of examples tend to retain influence throughout training. The prominent resonance fluctuations in the MRPC task may correspond with overfitting tendencies observed in the learning curves of models fine-tuned on this task (see Appendix B.2). This suggests that certain training examples in MRPC repeatedly shape model behaviour, potentially leading to memorization rather than generalization.

These discussions highlight how the perspective introduced here offers not only exploratory or descriptive insights but also opens up for practical applications, such as change point detection. This may allow us to identify critical transitions in learning, e.g. sudden shifts in model behavior, convergence phases, or the onset of overfitting, potentially offering a more nuanced view of the training progress. While qualitative patterns suggest links between signal fluctuations and learning phenomena (e.g., spikes in novelty during early training), we do not currently quantify these relationships. The scope of this study is primarily descriptive and comparative; we focus on establishing the plausibility and interpretability of the proposed signals across training conditions. Future work could build on this foundation by investigating formal change point detection techniques or correlating signal dynamics with shifts in validation loss (Appendix B.2) to strengthen causal interpretations. We leave these directions for future research.

5 Conclusion

This paper presented a novel method demonstrating how information-theoretic signals can offer insights into the dynamics of how language models process and integrate information during fine-tuning. While traditional evaluation metrics provide static snapshots of model performance, our findings underscore the value of examining temporal learning dynamics to uncover how internal representations evolve over time. Across fixed training settings, models exhibit synchronous and structured changes, while optimized training regimes introduce greater variability, thus revealing how different learning conditions shape information flow.

For the purpose of this study, we focused only on BERT-style models, but the methods proposed here can be extended to other architectures. both the information-theoretic framework and the format of the GLUE benchmark can be model-agnostic, meaning that this analysis could feasibly be extended to different architectures, training regimes, and tasks. By quantifying how models react to and retain new information, this moves beyond performance outcomes to illuminate *how* models learn, not just how well. It captures the learning process as a sequence of representational shifts, offering a mathematical perspective on learning as continuous adaptation rather than discrete updates. Our work contributes a new layer of transparency to model behavior, bridging performance metrics with internal state changes, and advancing our understanding of learning as an unfolding, temporal process.

Limitations

Signal processing and window size

The generated information signals are inherently dependent on the chosen window size, as this defines the context for measuring 'surprise'. In this study, the choice of window size was intended to balance the trade-off between capturing sufficient context from the surrounding documents while maintaining computational feasibility. Though meaningfully defining an optimal window size for a problem as such remains a complex challenge, a sensitivity analysis (see Appendix C.2) showed that varying the window size within a small range had minimal impact on results. Still, all tested sizes were relatively short compared to the full signal. Future work could explore larger windows to examine long-term trends, though comparing distributions over broader spans may introduce limitations due to memory constraints in the current information-theoretic measures. Additionally, as previously discussed, the choice of adaptive filter span value was guided by visual inspection due to the lack of a standardized quantitative criterion for adaptive filter tuning. Though a range of values were tested for each experiment (Appendix C.1), its effects on signal smoothing could be explored more systematically in future work.

Model and task diversity

The classification tasks were carefully selected to span a variety of differing scenarios. However, extending this work to include more complex classification problems, such as with imbalanced data or a wide number of classes, could offer additional insights. Likewise, our current work has been confined to English language tasks. While we found minimal differences between multilingual and monolingual BERT models, further investigation could clarify how language diversity shapes information dynamics. Similarly, while the models examined in this study have notable differences in architecture and training regimes, they all share the same BERT-style model at their core. Comparing information dynamics across more diverse model types could reveal alternative learning patterns and deepen our understanding of how different architectures integrate and retain information. While our model and task selection ensure a manageable comparison scope, extending this framework to other architectures (e.g., T5, GPT) and task types (e.g., generation, multilingual classification) would help assess the generalizability of information signals across broader learning paradigms.

Ethics Statement

This study aimed to aid in opening the 'black box' of LLMs and enhance transparency by exploring the information dynamics in their internal representations. It takes an exploratory and analytical approach in nature and does not involve model deployment, private user data, or human subjects. The dataset used is publicly available and widely used in the research community. While our work contributes to model transparency research, it does not provide definitive explanations of model decisions. We caution against potential misuse, such as over-interpreting signals or applying our framework to justify opaque model behavior without sufficient validation. Finally, we must consider the environmental impact of our work, with 24 fine-tuning experiments and subsequent generation of information signals.

Acknowledgments

All computations for this project were performed on the UCloud interactive HPC system, managed by the eScience Center at the University of Southern Denmark. Frida Hæstrup is supported by a grant from the Lundbeck Foundation (Grant No. R344-2020-1073).

References

Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2025. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert.

Thomas M Cover and Joy A Thomas. 2006. *Elements of Information Theory*, 2nd edition. John Wiley & Sons, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ossama Embarak. 2023. Decoding the Black Box: A Comprehensive Review of Explainable Artificial Intelligence. In *2023 9th International Conference on Information Technology Trends (ITT)*, pages 108–113.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of bert fine-tuning. In *AACL*.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 16(1):45–74.

Evan Hernandez and Jacob Andreas. 2021. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online. Association for Computational Linguistics.

Ross Deans Krisensen-McLachlan, Rebecca M.M. Hicke, Márton Kardos, and Thunø Mette. 2024. Context is Key(NMF). *CHR 2024: Computational Humanities Research Conference*, pages 829–847.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

David J. C. MacKay. 2002. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Cambridge.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Kristoffer L. Nielbo, Rebekah Brita Baglini, Peter Bjerregaard Vahlstrup, Kenneth C. Enevoldsen, Anja Bechmann, and Andreas Roepstorff. 2021a. News information decoupling: An information signature of catastrophes in legacy news media. *CoRR*, abs/2101.02956.

Kristoffer L. Nielbo, Frida Hæstrup, Kenneth C. Enevoldsen, Peter B. Vahlstrup, Rebekah B. Baglini, and Andreas Roepstorff. 2021b. When no news is bad news – Detection of negative events from news media content. *arXiv:2102.06505 [cs]*. ArXiv: 2102.06505.

Jason Phang, Haokun Liu, and Samuel R. Bowman. 2021. Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers. ArXiv:2109.08406 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Michael A. Riley, Scott Bonnette, Nikita Kuznetsov, Sebastian Wallot, and Jianbo Gao. 2012. A tutorial introduction to adaptive fractal analysis. *Frontiers in Physiology*, 3.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Julien Simon. 2021. Large Language Models: A New Moore's Law?

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan

Das, and Ellie Pavlick. 2018. What do you learn from context? Probing for sentence structure in contextualized word representations.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. ArXiv:2003.12298 [cs].

Eva Elisabeth Houth Vrangbæk and Kristoffer Laigaard Nielbo. 2021. *Composition and Change in De Ciuitate Dei: A Case Study of Computationally Assisted Methods*, volume 14: Augustine of Hippo's De ciuitate Dei: Content, Transmission, and Interpretations of *Studia Patristica*, pages 149–164. Peeters.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Melvin Wevers, Jan Kostkan, and Kristoffer L. Nielbo. 2021. Event flow – how events shaped the flow of the news, 1950-1995.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2022. A Closer Look at How Fine-tuning Changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Appendices

A Additional methods

A.1 Hyperparameters for the optimized group

Before fine-tuning for experiments in the optimized group, we performed hyperparameter tuning, as described in the paper. The resulting hyperparameter configurations can be found in Table 2. Hyperparameters not specified in the table were kept at default values.

Model	Task	*N* epochs	LR	Batch size
BERT	MNLI	3	$3e^{-5}$	16
BERT	MRPC	2	$5e^{-5}$	32
BERT	SST-2	2	$2e^{-5}$	16
distilBERT	MNLI	4	$5e^{-5}$	32
distilBERT	MRPC	3	$3e^{-5}$	16
distilBERT	SST-2	2	$5e^{-5}$	64
roBERTa	MNLI	4	$2e^{-5}$	32
roBERTa	MRPC	4	$2e^{-5}$	64
roBERTa	MRPC	4	$5e^{-5}$	64
mBERT	MNLI	4	$5e^{-5}$	32
mBERT	MRPC	3	$5e^{-5}$	64
mBERT	SST-2	3	$2e^{-5}$	64

Table 2: Hyperparameter configurations for each experiment in the optimized group. LR is the learning rate.

A.2 Model architectures and pre-training details

In Table 3, we highlight some of the main differences between the four models in terms of architecture and pre-training details.

Model	*N* layers	*N* parameters	*N* languages
BERT	12	110M	1
distilBERT	6	66M	1
roBERTa	12	125M	1
mBERT	12	110M	102

Table 3: Overview of architecture and training details for pre-trained versions of BERT, distilBERT, roBERTa, and mBERT.

B Additonal results

B.1 Grouped signals

To explore patterns in the extracted information signals, different groupings of the signals were visualized. As discussed in the paper, the analysis revealed task-specific patterns in the information signals from the experiments in the fixed group. In Figure 4, the information signals from the optimized group are shown grouped by task. All subfigures display all the same signals; however, each subfigure highlights the novelty and resonance signals for a respective task, while the remaining signals are depicted in transparent lines for comparison.

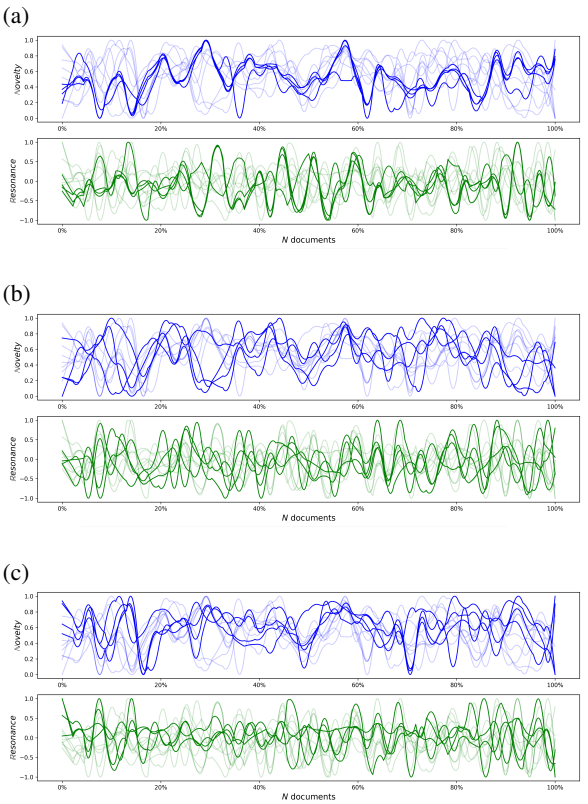


Figure 4: Normalized, smoothed novelty (blue) and resonance (green) signals for experiments in the optimized group grouped by task. The highlighted lines depict the signals from fine-tuning on the (a) MNLI, (b) MRPC, and (c) SST-2 task, respectively. The transparent lines show the remaining signals, i.e., signals from those tasks not highlighted in each plot.

Similarly, the information signals from the experiments were grouped by model type to investigate potential patterns. This is depicted in Figure 5, with each row of subfigures highlighting the signals of the four different models, respectively. The left column shows experiments from the fixed group, and the right column shows experiments from the optimized group.

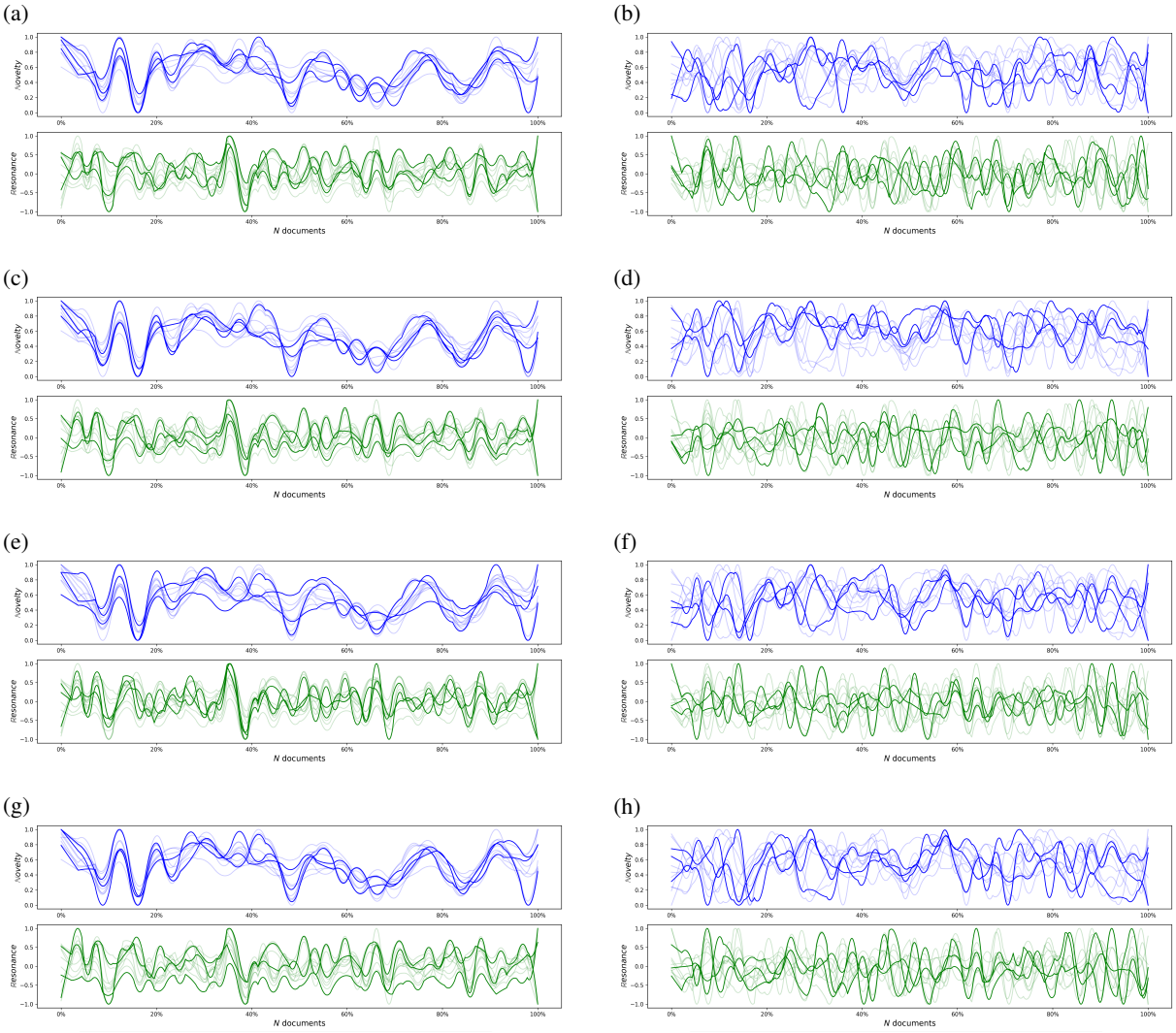


Figure 5: Normalized, smoothed novelty (blue) and resonance (green) signals grouped by model. The left column shows the fixed group and the right column shows the optimized group. Each row corresponds to a model: (a–b) BERT, (c–d) distilBERT, (e–f) roBERTa, and (g–h) mBERT. Highlighted lines show signals for each model-task combination; transparent lines show the rest.

B.2 Learning curves

In Figure 6, the learning curves for the various experiments are presented, illustrating the models' performances on the classification tasks during the fine-tuning process. Each subfigure represents an experiment, displaying the learning curves for each of the models fine-tuned on a task. The purple line represents validation accuracy, the red line represents validation loss, and the yellow line represents training loss. Note that differing training durations in the optimized group led to uneven checkpoint sampling across experiments. As a consequence, some plots — such as those for roBERTa — are missing or incomplete (e.g., if training terminated before enough checkpoints were saved).

C Sensitivity analyses

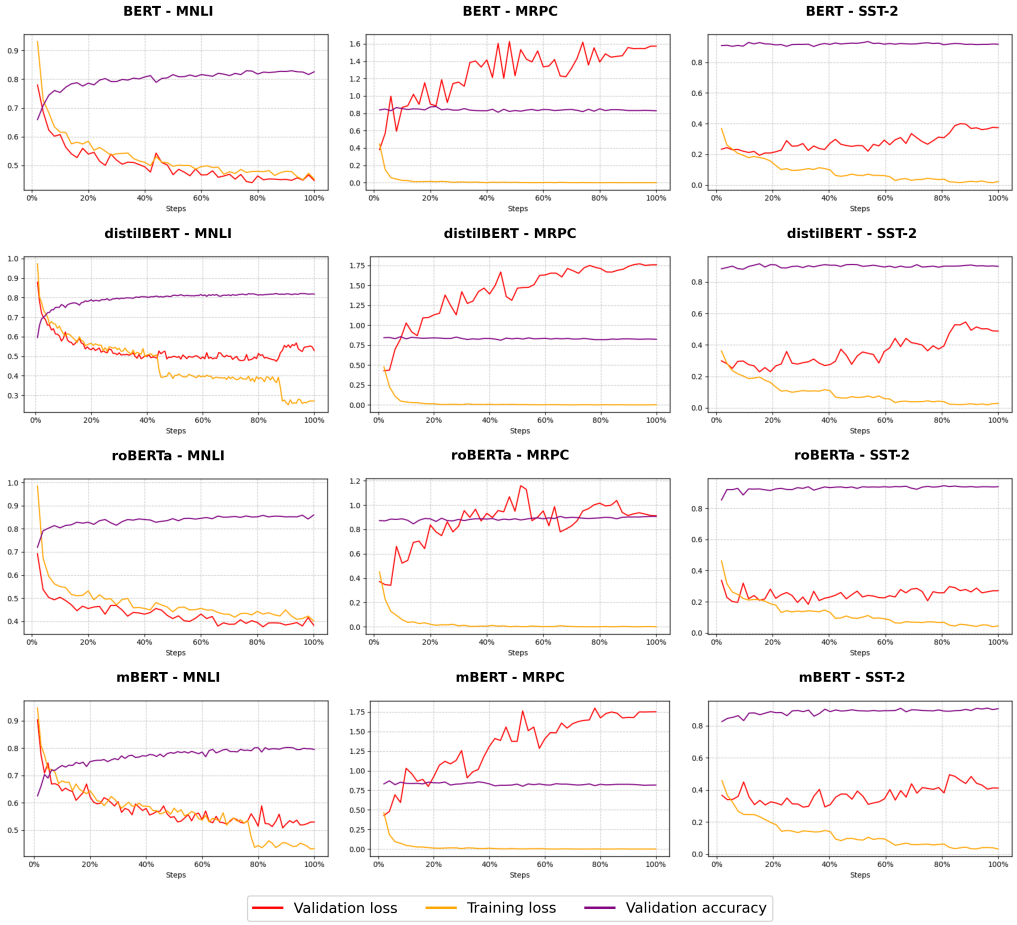
C.1 Defining the adaptive filter span

As discussed in Section 2.6, we follow the proposed method for defining the span value for the adaptive filter (Riley et al., 2012); namely, visual inspection of the fit of the smoothed signal produced by varying span values. Figure 7 displays an example of this.

C.2 Defining the window size

As mentioned in the Limitations, a sensitivity analysis was also performed to investigate the effect of varying the window size in which to calculate the information signals. An example of this can be seen in Figure 8.

(a)



(b)

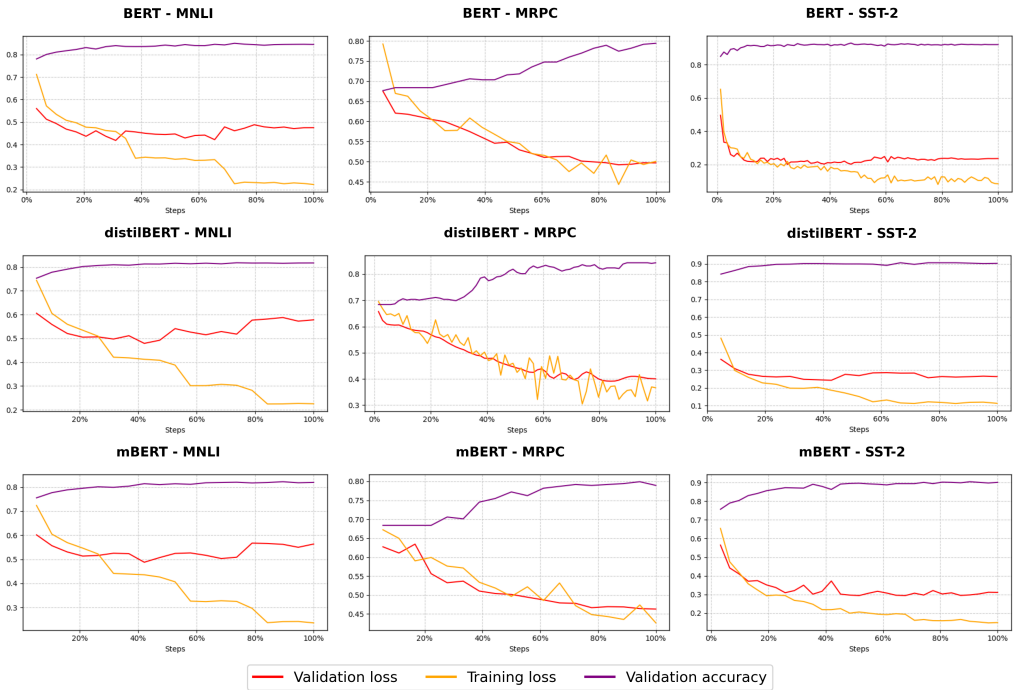


Figure 6: Learning curves for experiments in (a) the fixed group and (b) the optimized group. The red line represents the validation loss, the yellow line represents the training loss, and the purple line represents the validation accuracy.

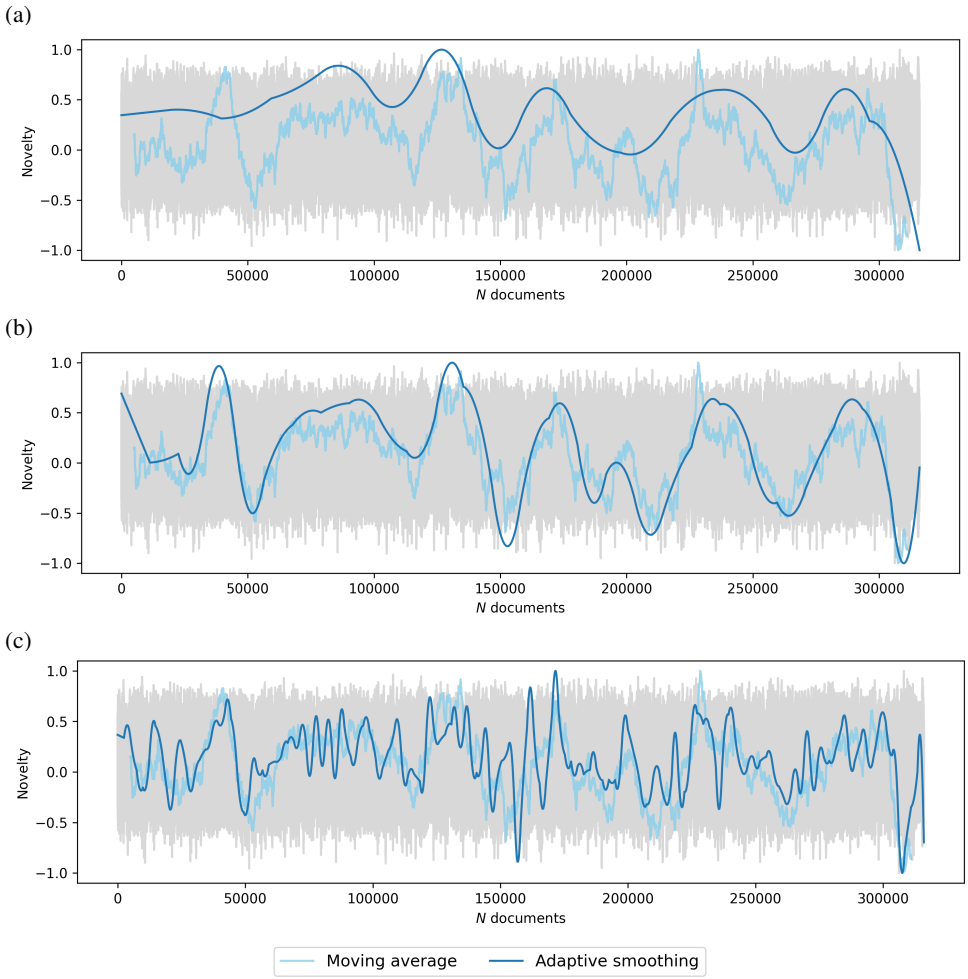


Figure 7: Example of the effect of different span values for the adaptive filter. The signal depicted here is the novelty signal from BERT fine-tuned on the MRPC task with fixed hyperparameters. The grey line depicts the original, unsmoothed novelty signal. The light blue line depicts the novelty signal's moving average (w=10000). The dark blue line depicts the smoothed signal from the adaptive filter using span values of (a) 32, (b) 56, and (c) 128, respectively. All signals are normalized.

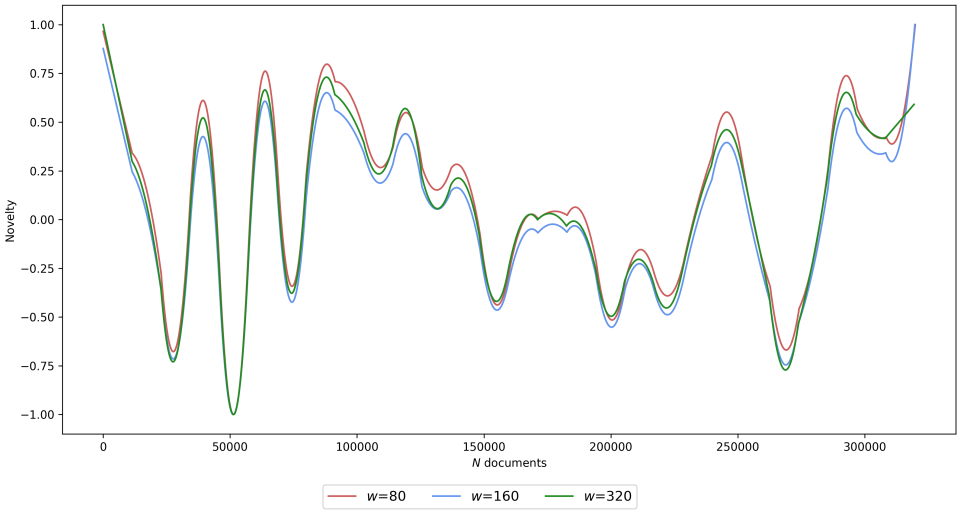


Figure 8: Example of the effect of varying the window size for which to calculate novelty, transience, and resonance in. The signal depicted is the normalized, smoothed novelty signal from distilBERT fine-tuned on the MNLI task with fixed hyperparameters. The different lines represent different window sizes (80, 160, 320)