When Dimensionality Hurts: The Role of LLM Embedding Compression for Noisy Regression Tasks

Anonymous ACL submission

Abstract

LLMs have shown remarkable success in language modelling due to scaling laws found in model size and the hidden dimension of the model's text representation. Yet, we demonstrate that compressed representations of text can yield better performance in LLM-based regression tasks. In this paper, we compare the relative performance of embedding compression in four different signal-to-noise contexts: financial return prediction, health outcome prediction, writing quality assessment and review scoring. Our results show that compressing embeddings, in a minimally supervised manner using an autoencoder's hidden representation, can mitigate overfitting and improve performance on noisy tasks, such as financial return prediction; but that compression reduces performance on tasks that have high causal dependencies between the input and target data. Our results suggest that the success of interpretable compressed representations, such as sentiment, may be due to a regularising effect.

1 Introduction

Modern machine learning research increasingly relies on LLMs to handle complex real-world tasks (Lin et al., 2025; Rahimikia and Drinkall, 2024; Huang et al., 2024). Recent progress in LLM performance has largely come from scaling models' parameters, training dataset size and the expressivity of the LLM via the model's hidden dimension (Kaplan et al., 2020). In recent years the hidden dimension has scaled from a standard representation size of 768 dimensions (Devlin et al., 2019) to up to 16384 (Grattafiori et al., 2024), and possibly higher in some large closed-source models. While it is not clear whether the relationship between model scaling and linguistic performance will hold indefinitely (Xue et al., 2023), it is generally accepted that modelling language requires a high-dimensional representation space (Grattafiori et al., 2024). This has meant that LLMs have very strong formal linguistic competence (Mahowald et al., 2024). However, some machine learning tasks, like stock returns prediction tasks, have inherently low signal-to-noise relationships between the input and output (Sawhney et al., 2020), which we will refer to as "noisy" tasks in this paper. In the case of predicting stock returns from news, noise arises not only from uninformative articles or weak causality between the article and the stock price but can also come from delayed reactions, market efficiency, and unpredictable macroeconomic influences (Antweiler and Frank, 2004; Mantilla-García and Vaidyanathan, 2017). In such noisy research areas, the link between dimensionality and performance is unclear, and feature selection or compression can act as a regularising component (Tian and Zhang, 2022). When input dimensionality is too large, models risk overfitting by memorising noise rather than learning meaningful patterns. On the other hand, a low-dimensional embedding might underfit by losing critical high-order interactions. This paper explores the relationship between text-embedding dimensionality and downstream performance in tasks where the signal is noisy, focusing on tasks with a differing signal-to-noise ratio.

Not all tasks are well-suited to purely generative LLMs; many tasks benefit more from supervised machine learning (Tang et al., 2024), where labelled data guide classification or regression outcomes by identifying robust dependencies between input text and target outputs (Johan Berggren et al., 2019). Generative models often require extensive computational resources and large datasets (Hoffmann et al., 2022), creating obstacles under computational or data constraints. Problems also emerge when using the output from generative models in larger architectures that fuse textual data with other modalities, such as numerical or structured information (Drinkall et al., 2025), since generative models can produce unpredictable outputs (Wu et al., 2022) and are relatively weak at com-

plex multivariate numerical reasoning (Liu et al., 2024). Discriminative models are also better able to model a probabilistic distribution with a limited run of data, since generative models can overfit to the training distribution (Lee et al., 2022; Carlini et al., 2023). As such, it is often preferable to use the embedding representations from LLMs as input features for a conventional neural network in regression-based tasks as opposed to passing all of the numerical and textual data into a prompt.

There has been work investigating how well text embeddings perform in regression tasks (Tang et al., 2024), but none have investigated the degree to which the noise of a task affects the optimal dimensionality of the text representation. This is despite the widespread adoption and success of interpretable, compressed representations of text in financial return prediction tasks like sentiment (Fazlija and Harder, 2022), emotion (Tilly et al., 2021), and topics (Drinkall et al., 2022; García-Méndez et al., 2023). While many papers have shown that these compressed representations can perform better than raw text embeddings, this paper investigates the degree to which this is due to regularisation rather than the true value of the features. We consider the internal representation from an autoencoder, and show that interpretable features such as sentiment or emotion do not deliver an improvement beyond the minimally supervised autoencoder latent representation.

This paper has the following contributions:

1. Provides some evidence for a link between the optimal representational dimensionality of text in a regression task and the signal-to-noise ratio of the dataset.

2. Demonstrates that gains attributed to interpretable features (e.g., emotion, sentiment) in financial returns tasks may stem from representational compression, rather than from inherently superior feature sets.

3. Identifies the optimal dimensionality in several downstream tasks.

4. A financial news stock returns dataset released under an academic license.[1]

2 Datasets

To explore the relationship between the signal-to-noise ratio and the optimal dimensionality of the text representation in a regression task, we compare four domains of conceivable regression tasks. Stock market return prediction using news articles is a notoriously noisy domain (Black, 1986) since a significant proportion of the news articles are likely to not contain any useful information (Antweiler and Frank, 2004). Health outcome prediction is also a noisy task, with length of stay reflecting many latent factors (bed flow, comorbidities, discharge logistics etc.) that are only partially mentioned in nursing notes. We contrast these noisy domains to customer review and writing quality datasets, which both have a stronger connection between the regression input and the target value.

2.1 Low Signal Datasets

Financial Returns. We combine two data sources to form the financial dataset: CRSP daily price data and news articles. For the 50 most traded U.S. stocks, we use the closing bid/ask average as the daily price. The daily return is defined as the percentage increase between the previous and the next day's closing price. This return is the regression model's target. We use the next and previous day's data as opposed to the current day's price to be sure that the publication of the article intersected the two prices and thus avoid including samples where the article was published after the market's closing time. The underlying assumption is that the news content either causes or reflects the observed price change. More details about the dataset are reported in §B.1.

Health Outcomes. For the health outcome prediction, we used the MIMIC-III dataset (Johnson et al., 2016a,b) and extracted the first nursing notes from a patient entering the hospital to try to predict the length of stay. Since duration is always positive and log-normal, we apply a log transform to achieve a normal distribution for the target variable. In addition to the signal being indirectly affected by latent factors, nursing notes are clinician-dependent and can be highly variable, adding lexical and stylistic noise that can obscure whatever predictive signal does exist. This task is less noisy than financial return prediction, but for the preceding reasons, a slight dislocation remains between the input text and the ultimate target variable. More details about this dataset are outlined in §B.2.

2.2 High Signal Datasets

To compare the degree to which noise affects the optimal dimensionality of a task we selected a

[1]To be included in final version.

dataset with a high causal dependency between the input data and target information. Written product reviews are directly linked to the score assigned to the review, therefore we use the following datasets: Yelp Reviews (Zhang et al., 2015), App Reviews and Amazon Reviews (McAuley and Leskovec, 2013). Writing quality is a more ambiguous task but there is still a direct link between the input text and the target which also makes it a good candidate for comparison against more noisy tasks – as such we use the ELL English Grading dataset (Franklin et al., 2022). Details of each dataset are in §B.6.

3 Methodology

Given the success of generative LLMs, much of the recent research on downstream tasks has focused on how to use LLMs in a prompting setting (Chang et al., 2024). However, there are some domains where encoder-based LLMs are better suited: embedding tasks have been dominated by LLMs pre-trained with bi-directional attention (Song et al., 2020) or uni-directional attention followed by bi-directional fine-tuning (Lee et al., 2024; BehnamGhader et al., 2024). Likewise, recent work shows that generative models perform worse on word meaning comprehension than encoder-based LLMs (Qorib et al., 2024). As such, we encode the textual information using *all-mpnet-base-v2* (Song et al., 2020; Reimers and Gurevych, 2019), an encoder-based model fine-tuned using a contrastive objective function on a series of sentence similarity tasks. The model is widely used and competitive on the MTEB benchmark (Muennighoff et al., 2023). The results were consistent across different models (§E).

3.1 Input Processing and Chunking

Each textual input x is tokenized into a sequence of tokens $(t_1, t_2, \ldots, t_L)$. To handle variable-length inputs, we split the token sequence into M chunks, each of length at most C, where C is the maximum context window of the model. If the final chunk is shorter than C, it is padded; similarly, sequences with fewer than M chunks are padded, ensuring each batch element has uniform shape $[M, C]$. After chunking, we pass the sequences through a pre-trained language model to produce token-level embeddings, and then mean-pool across the final layer token representations of all of the chunks to produce $\mathbf{v}_i \in \mathbb{R}^{d_{\text{LLM}}}$. Where d_{LLM} is the size of the LLM's hidden dimension. The target features are standardized to enable easier interpretation.

3.2 Dimensionality Reduction

To obtain a lower-dimensional representation, we train an autoencoder consisting of an encoder $E : \mathbb{R}^{d_{\text{LLM}}} \to \mathbb{R}^{d_z}$ and a decoder $D : \mathbb{R}^{d_z} \to \mathbb{R}^{d_{\text{LLM}}}$, where d_z is the size of the autoencoder's hidden dimension, and $\mathcal{L}_{\text{AE}}$ is the loss:

$$\mathbf{z}_i = E(\mathbf{v}_i), \quad \hat{\mathbf{v}}_i = D(\mathbf{z}_i).$$

$$\mathcal{L}_{\text{AE}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{v}_i - \hat{\mathbf{v}}_i\|^2.$$

The training runs for 100 epochs with early stopping if validation loss does not improve for 5 epochs. After training, we use the encoder E to compress all embeddings $\mathbf{v}_i$ into $\mathbf{z}_i \in \mathbb{R}^{d_z}$ and vary d_z to identify the optimal compression ratio. This methodology outperforms other compression techniques (§E), but the overall performance profile shown in §4 is consistent across methodologies.

Our methodology is compared to popular techniques used in stock returns analysis, like emotion and sentiment scores. We take the softmax outputs from DistilRoBERTa models that have been fine-tuned on sentiment and emotion classification tasks and pass the outputs into the regression model. This enables us to compare their relative performance to infer whether their strong performance is due to regularisation or valuable feature selection.

3.3 Regression Model

For the regression task, we use a random forest model as it is robust and widely used (Breiman, 2001; Roy and Larocque, 2012), which simplifies the experimental setup so that focus can be directed to the compression methodology. For the same reason, we use the default parameters. We also tried a two-layer MLP in line with Tang et al. (2024) (§C), which did not learn the financial returns task.

4 Results

We report results with the Huber loss (Huber, 1964), which combines the robustness to outliers of MAE with the sensitivity to small errors and smooth gradients of MSE, removing the outlier bias that can dominate MSE. The significance level of $d_z = x$ is determined using a T-Test (Student, 1908) between the Huber error distributions of the best performing latent dimension d_z^* and $d_z = x$.

By varying the hidden dimension of the autoencoder d_z and then passing the input into our regression model, Figure 1 shows that the optimal dimensionality on the financial returns task is 8, but

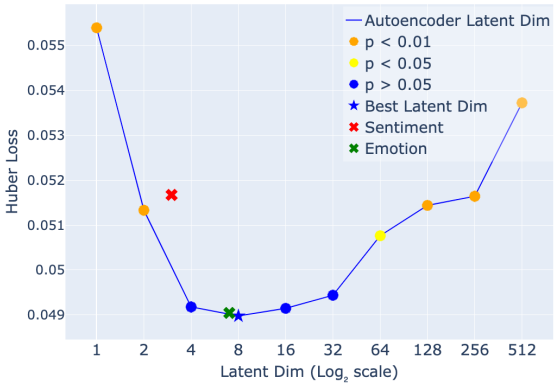


Figure 1: Huber Loss on the fin. returns task for different d_z values. The performance of sentiment and emotion is also reported. The significance of each result compared to the best d_z is displayed using blue ($p > .05$), orange ($p < .01$) and yellow ($p < .05$) colours.

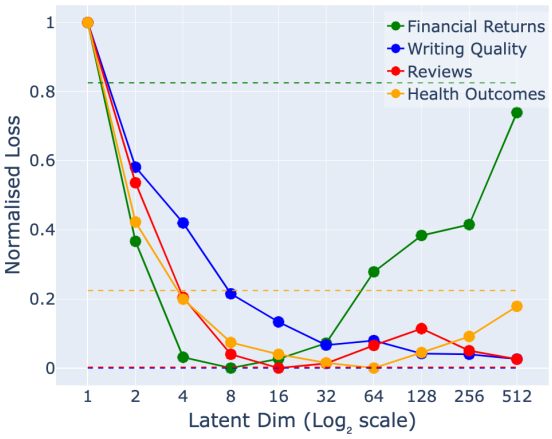


Figure 2: Huber loss averaged over **Review**, **English Writing**, **Health Outcomes** and **Financial Returns** tasks - granular performance in §D. The y-axis represents the loss of each task as a percentage of the maximum and minimum loss on that task. The performance without any compression is marked with the dashed line.

that there is not a statistically significant difference between a $d_z^* = 8$ and $d_z \in \{4, 16, 32\}$. There is also a significant difference between $d_z^* = 8$ and $d_z \in \{1, 2, 128, 256, 512\}$. The result shows that the optimal dimensionality is significantly less than d_{LLM}, showing that for this noisy task some dimensionality reduction is necessary.

Figure 1 also reports the performance of compressing the text representation into a class probability vector of sentiment and emotion scores. Both representations do not exceed the expected performance of the autoencoder features at their respective dimensions d_z. Despite the reported success of emotion and sentiment features in similar tasks (Tilly et al., 2021; Fazlija and Harder, 2022) the findings of this work suggest that some of this performance improvement can be explained due to the regularising effect of dimensionality reduction rather than the inherent value of the features.

4.1 Impact of Noise

To compare the extent to which noise affects the optimal dimensionality of a task, we test on different domains. Fig. 2 shows that for the financial returns and health outcomes tasks, there is a convex relationship between loss and dimensionality, whereas it approximates a negative exponential in strong signal tasks; the performance does not deteriorate at high dimensions. The large difference in error distributions between the different tasks suggests that input dimensionality is a key parameter for regression-based tasks. Also, in all domains, the performance reaches 10% of the minimum loss at a much smaller dimension than d_{LLM}. The dimension at which this performance is achieved can be called the "intrinsic dimension", which ranges between 4 and 32 for all tasks. This suggests that the pertinent signals for regression tasks in general can be compressed to a lower dimension and achieve strong performance. For architectures that have poor time complexity as a function of input length, this is an important finding.

5 Conclusion

Our results suggest that for tasks with high noise, coarser and lower-dimensional features improve performance. The result implies that researchers should consider the noise of a task when making decisions about the dimensionality of text. In particular, the results highlight the importance of dimensionality reduction in financial returns prediction tasks, with an optimal autoencoder latent dimension of $d_z = 8$. The lack of statistical significance for $d_z \in \{4, 16, 32\}$ suggests some flexibility in choosing the dimensionality, while extreme values lead to significant performance deterioration. It is also clear that the results are consistent across domains, with smaller but comparable effects seen in health outcome prediction tasks. It seems that coarse features are more performant than the default granular LLM representation in such regression tasks. The findings also indicate that sentiment and emotion-based representations do not provide inherent advantages over learned latent features in financial contexts, implying that their previous success in similar tasks may be attributed to regularisation effects rather than intrinsic informativeness.

6 Limitations

Although our findings demonstrate the importance of reducing dimensionality in high-noise tasks, some limitations should be noted. Firstly, by using real tasks, it is difficult to measure the intensity of the noise in each task. The lack of a "noise" metric limits the extent to which we can analyse the degree and type of the relationship between noise and dimensionality. However, by using a diverse array of task types, we aim to have minimised the problems associated with this point. Secondly, while we aimed to keep the modelling process simple to not distract from the main thrust of the paper, data compression, a future research direction could be to apply the findings to more complex models.

References

Werner Antweiler and Murray Z Frank. 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *First Conference on Language Modeling*.

Fischer Black. 1986. Noise. *The Journal of Finance*, 41(3):528–543.

L. Breiman. 2001. Random Forests. *Machine Learning*, 45:5–32.

Nicholas Carlini, Daphne Ippolito, et al. 2023. Quantifying Memorization Across Neural Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net. In-person poster; Top 25% paper; Session MH1-2-3-4 #96.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient Prompting Methods for Large Language Models: A Survey. *Preprint*, arXiv:2404.01077.

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*.

Felix Drinkall, Janet B. Pierrehumbert, and Stefan Zohren. 2025. Forecasting Credit Ratings: A Case Study where Traditional Methods Outperform Generative LLMs. In *Proceedings of the FinNLP-FNP-LLMFinLegal Workshop @ COLING 2025*. Association for Computational Linguistics.

Felix Drinkall, Stefan Zohren, and Janet Pierrehumbert. 2022. Forecasting COVID-19 Caseloads Using Unsupervised Embedding Clusters of Social Media Posts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, et al. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv preprint arXiv:2502.13595*.

Bledar Fazlija and Pedro Harder. 2022. Using Financial News Sentiment for Stock Price Direction Prediction. *Mathematics*, 10(13).

Alex Franklin, Maggie, Meg Benner, Natalie Rambis, Perpetual Baffour, Ryan Holbrook, Scott Crossley, and Ulrich Boser. 2022. Feedback Prize - English Language Learning.

Silvia García-Méndez, Francisco de Arriba-Pérez, Ana Barros-Vila, Francisco J. González-Castaño, and Enrique Costa-Montenegro. 2023. Automatic Detection of Relevant Information, Predictions and Forecasts in Financial News Through Topic Modelling with Latent Dirichlet Allocation. *Applied Intelligence*.

Ary L. Goldberger, Luis A. N. Amaral, et al. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A Generic News Crawler and Extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. 2022. Training Compute-optimal Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Baizhou Huang, Shuai Lu, Xiaojun Wan, and Nan Duan. 2024. Enhancing Large Language Models in Coding Through Multi-Perspective Self-Consistency. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1429–1450, Bangkok, Thailand. Association for Computational Linguistics.

Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *Annals of Statistics*, 53(1):73–101.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or Classification? Automated Essay Scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. MIMIC-III Clinical Database (version 1.4).

Alistair E. W. Johnson, Tom J. Pollard, Li-wei Shen, Li-wei H. Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016b. MIMIC-III, A Freely Accessible Critical Care Database. *Scientific Data*, 3:160035.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *Preprint*, arXiv:2001.08361.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *Preprint*, arXiv:2405.17428.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony G. Cohn, and Janet B. Pierrehumbert. 2025. Graph-enhanced Large Language Models in Asynchronous Plan Reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, Bangkok, Thailand. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating Language and Thought in Large Language Models. *Trends in Cognitive Sciences*, 28(6):517–540.

Daniel Mantilla-García and Varadharajan Vaidyanathan. 2017. Predicting Stock Returns in the Presence of Uncertain Structural Changes and Sample Noise. *Financial Markets and Portfolio Management*, 31.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal Components Analysis (PCA). *Computers Geosciences*, 19(3):303–342.

J. McAuley and J. Leskovec. 2013. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 897–908. ACM.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

Eghbal Rahimikia and Felix Drinkall. 2024. Re (Visiting) Large Language Models in Finance. *Available at SSRN*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marie-Hélène Roy and Denis Larocque. 2012. Robustness of Random Forests for Regression. *Journal of Nonparametric Statistics*, 24(4):993–1006.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. 2020. Deep Attentive Learning for Stock Movement Prediction From Social Media Text and Company Correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8415–8426, Online. Association for Computational Linguistics.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel Principal Component Analysis. In *Artificial Neural Networks — ICANN'97*, pages 583–588, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Student. 1908. The Probable Error of a Mean. *Biometrika*, pages 1–25.

Eric Tang, Bangding Yang, and Xingyou Song. 2024. Understanding LLM Embeddings for Regression. *Preprint*, arXiv:2411.14708.

Yingjie Tian and Yuqi Zhang. 2022. A Comprehensive Survey on Regularization Strategies in Machine Learning. *Information Fusion*, 80:146–166.

Sonja Tilly, Markus Ebner, and Giacomo Livan. 2021. Macroeconomic Forecasting Through News, Emotions and Narrative. *Expert Systems with Applications*, 175:114760.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

David Weenink. 2003. Canonical Correlation Analysis. *Institute of Phonetic Sciences, University of Amsterdam, Proceedings*, 25:81–99.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. To Repeat or Not To Repeat: Insights from Scaling LLM Under Token-crisis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

A Future Work

Future research could explore adaptive dimensionality compression methods that dynamically adjust based on the signal-to-noise ratio; however, to do this, a measure of signal-to-noise is required before processing the input features. Researchers could also use this framework to assess the relative performance of new text encoding methodologies in regression tasks to make sure that the value does not just come from model regularisation.

B Dataset Details

B.1 Stock Returns Dataset

As outlined in the main section of the paper, this dataset was curated for this paper. Most of the dataset details are outlined in Section 2.1, but this section contains the details that are missing. We source news articles via CommonCrawl News (Hamborg et al., 2017), scraping articles from Yahoo Finance. Using a pre-trained named entity recognition BERT model (Tjong Kim Sang and De Meulder, 2003; Devlin et al., 2019), we extract all mentioned organisations, then filter them through a dictionary of company synonymous and abbreviations to identify target companies. We then apply another filter to make sure that only one of the target companies is mentioned in each sample to reduce the noise slightly. The test set is the whole of 2023, which contains 17,810 articles, and the training and validation sets are defined using a temporal split, which takes the last 10% of data between 2017 and 2022. The resultant training and validation sets contain 30,115 and 3,346 samples, respectively.

B.2 Health Outcomes Dataset

The health outcomes dataset uses the *"Nursing"* and *"Nursing/other"* categories from the *NOTEEVENTS.csv* in the MIMIC-III database (Goldberger et al., 2000; Johnson et al., 2016a,b). Using the patient IDs, we select the first entry for each patient and remove all entries that are less than 100 tokens. To determine the length of stay, we retrieve the admission and discharge times (*ADMITTIME* and *DISCHTIME*) from the *ADMISSIONS.csv*, and then link with the appropriate nursing notes using the patient ID. The dataset consists of 40,543 samples, and we perform a random train-val-test split of 85.5:4.5:10.

The first author completed the necessary training course "CITI Data or Specimens Only Research", and the research application was approved by the MIT Laboratory for Computational Physiology, Institute for Medical Engineering and Science. The research conducted in this paper complies with the license outlined in §H. The data was previously collected, anonymised and does not include NHS data. As a result, it complies with the authors' institution's ethical policy.

B.3 Yelp Reviews Dataset

The Yelp dataset (Zhang et al., 2015) consists of 700k Yelp reviews with a star rating between 1 and 5. There are 650k training samples and 50k testing samples, and the split is taken from the Huggingface dataset *Yelp/yelp_review_full*.

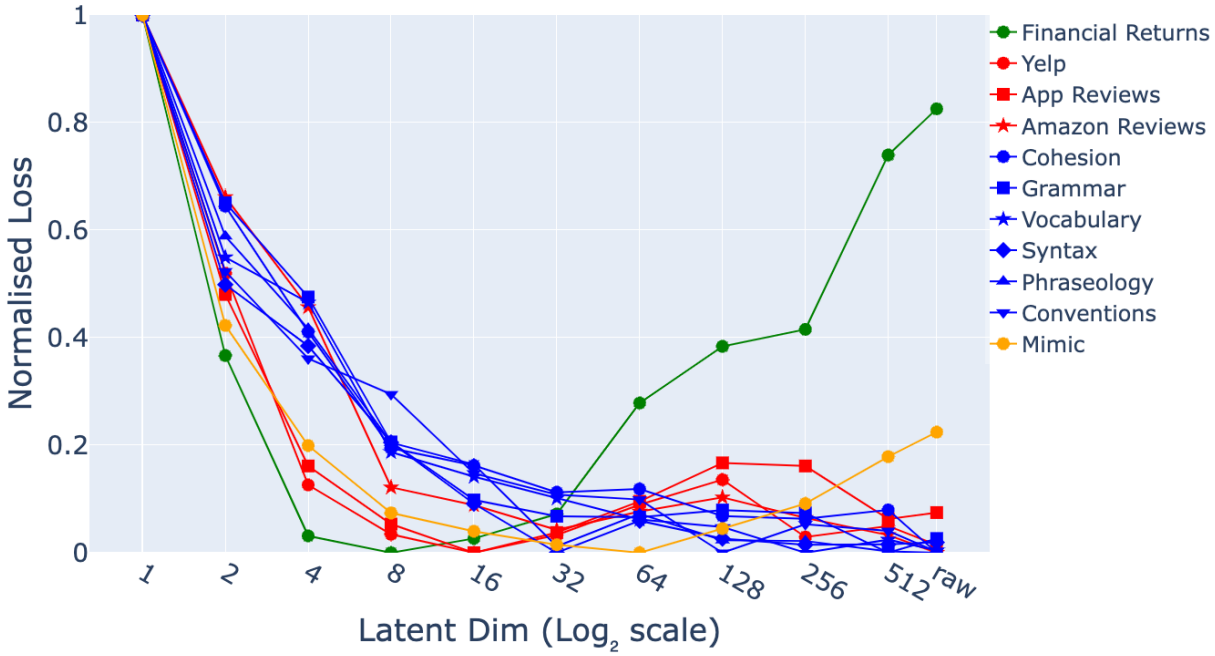


Figure 3: The normalised Huber loss of each dataset that makes up the result in Figure 2. "raw" appears in the same location as 768 in the plot since this is the dimension of the non-compressed embedding of the *all-mpnet-base-v2*.

B.4 App Reviews Dataset

This App Review dataset contains reviews of 395 Android applications, covering 629 versions. It provides the review and the star rating between 1 and 5, and user-reported issues in English. The dataset consists of 288k samples and we perform a random train-val-test split of 85.5:4.5:10. The data came from the *sealuzh/app_reviews* Huggingface dataset.

B.5 Amazon Reviews Dataset

The Amazon Reviews dataset (McAuley and Leskovec, 2013) consists of 568k fine food reviews collected from Amazon over a period of more than 10 years, up to October 2012. Each review includes a product ID, user ID, profile name, rating (1–5), helpfulness votes, timestamp, summary, and full text. The data came from the Huggingface dataset *jhan21/amazon-food-reviews-dataset*, which did not contain any predefined train-test splits so we performed a random train-test split of 85.5:4.5:10.

B.6 Writing Quality Dataset

The writing quality dataset (Franklin et al., 2022) comes from a Kaggle competition set up by Vanderbilt University. The competition aimed to improve automated feedback tools for English Language Learners (ELLs) by developing language proficiency models using real student essays. The dataset assesses English text over six criteria: cohesion, syntax, vocabulary, phraseology, grammar and conventions. We report the results of cohesion, vocabulary, and grammar. The dataset consists of 3.91k samples and we perform a random train-val-test split of 85.5:4.5:10.

C MLP

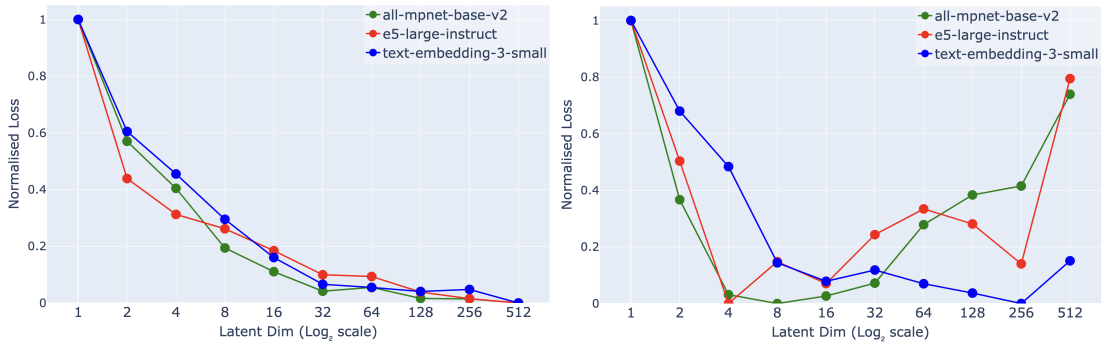
The configuration for the unsuccessful MLP is outlined in this section. The model was not able to learn the financial returns task for any dimensional input. The inadequate overall performance and high variance in prediction errors meant that no statistically significant conclusions could be drawn. We believe that this negative result will aid other researchers in this area. The compressed embeddings $\mathbf{z}_i$ serve as inputs to an MLP with hidden dimension d_{mlp}:

$$\mathbf{h}^{(1)} = \mathcal{D}_p\big(\sigma(W^{(1)}\mathbf{z}_i + \mathbf{b}^{(1)})\big),$$

$$\mathbf{h}^{(2)} = \mathcal{D}_p\big(\sigma(W^{(2)}\mathbf{h}^{(1)} + \mathbf{b}^{(2)})\big),$$

$$\hat{y}_i = W^{(3)}\mathbf{h}^{(2)} + \mathbf{b}^{(3)},$$

where $\mathbf{z}_i \in \mathbb{R}^{d_z}$, $W^{(1)} \in \mathbb{R}^{d_{\text{mlp}} \times d_z}$, and $W^{(2)}, W^{(3)}$ similarly match the required dimensions. Dropout $\mathcal{D}_p(\cdot)$ is applied with probability p, and $\sigma(\cdot)$ is the ReLU activation. We optimize the Huber loss with $\delta = 1.0$:



(a) Writing quality tasks (b) Financial Returns

Figure 4: Performance comparison of ***all-mpnet-base-v2***, ***multilingual-e5-large-instruct***, and ***text-embedding-3-small*** on different domain tasks.

$$\mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} \text{HuberLoss}(y_i, \hat{y}_i, \delta).$$

Most targets y_i are close to zero, so we apply target standardisation and stop training when validation loss does not improve for 5 epochs, restoring the best model state.

D Individual Dataset Results

Figure 3 shows the performance of each dataset that makes up the averaged result in Figure 2.

E Generalisation of Findings

E.0.1 Compression Comparison

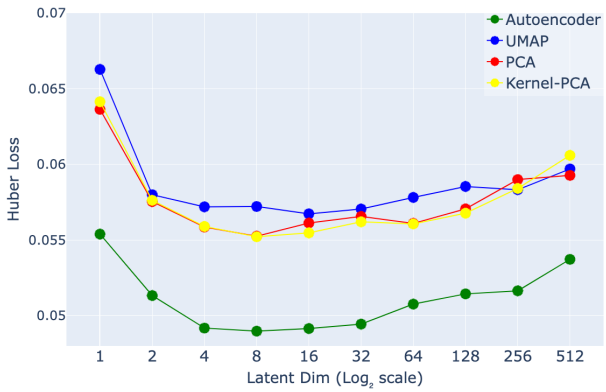


Figure 5: Raw loss for the different compression methodologies: Autoencoder, UMAP, PCA, Kernel-PCA. Initial embedding created using *all-mpnet-base-v2*.

To determine whether the findings in this paper were consistent across compression methodologies, we compared the performance distribution to multiple commonly used compression algorithms: PCA (Maćkiewicz and Ratajczak, 1993), Kernel-PCA (Schölkopf et al., 1997) with an RBF kernel to capture non-linear dependencies, and UMAP (McInnes et al., 2018) as a popular modern comparison. Figure 5 shows that the autoencoder is the best methodology on the financial returns task, and justifies its use in the main body of the paper. The other methodologies exhibit very poor performance, but have the same performance profile.

E.1 Model Comparison

The results of *all-mpnet-base-v2* are reported in the paper, because it is a highly performant model which is widely used in the community. In order to understand if the results of this paper are model agnostic, we also tested the performance of different architectures. The first comparison model was OpenAI's *text-embedding-3-small*, which represents an industry standard. The second comparison model was *intfloat/multilingual-e5-large-instruct*, the best model below 1B parameters on the MTEB leaderboard (Enevoldsen et al., 2025). Figure 4 shows that both models display a similar performance profile to *all-mpnet-base-v2*. Interestingly, the optimal dimensionality is largest for *text-embedding-3-small*. While the model size is unknown as it is closed source, the hidden size of the residual stream representations is the largest of the three models, suggesting that it could likely be the largest of the three models. Future work should look into how model size effects the optimal dimensionality of a regression task.

F Autoencoder Visualisation

Downstream performance on regression tasks provides insight into the quality of the autoencoder's compression. However, Figure 6 offers a more direct comparison between the autoencoder's input and output embeddings, $\mathbf{v}_i$ and $\hat{\mathbf{v}}_i$, respectively. The figure displays the cosine similarity between the raw ($\mathbf{v}_i$) and reconstructed ($\hat{\mathbf{v}}_i$) embeddings for different hidden dimensions. The graph provides us with a further understanding of the reconstructive process; it seems that a $d_z = 256$ is the point at which performance reaches an asymptote. It also implies that there is some semantic loss at the optimal dimensions in Figure 2.

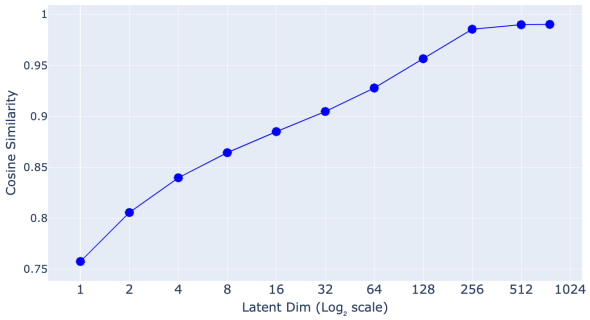


Figure 6: Cosine similarity between $\mathbf{v}_i$ and $\hat{\mathbf{v}}_i$ on the financial returns dataset.

G Feature Comparison

To understand whether the autoencoder latent dimensions were capturing features similar to the emotion and sentiment representations, we encoded the test set of the financial returns task using the emotion/sentiment models used in Section 3.2 and a trained autoencoder with d_z equal to the number of emotions/sentiments.

To determine whether the encoding types were similar we identified the feature pair with the largest absolute correlation across all of the test samples. The maximum Pearson correlation coefficient between the latent encoding and sentiment was 0.2312 and for emotion it was 0.2796. Both scores are low and show that the autoencoder does not compress the text into any single emotion/sentiment feature.

We also wanted to test whether the autoencoder was compressing the text to contain emotion/sentiment information through a combination of the features, so we used Canonical Correlation Analysis (CCA) (Weenink, 2003) to determine the correlation between a linear combination of both feature sets. We fitted the CCA model to the latent features and the emotion/sentiment features to maximise the correlation between the two sets. The maximum canonical Pearson correlation coefficient between the two sets was 0.4644 for emotion and 0.3011 for sentiment. While there is some positive correlation between the two sets, the result is weak and suggests that the encoding methodologies extract different signals. This result suggests that a mixture of the two methodologies could lead to better performance still and that neither compression technique is optimal, which is an exciting finding for future research.

H Licenses

The data has been used for exclusively academic and research purposes and as a result, complies with the Terms of Use for CRSP. The news information was taken from Commoncrawl News datacrawl, with is released under a permissive Apache 2.0 license.

The dataset released alongside this paper is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, allowing for academic sharing and adaptation while prohibiting commercial use. Researchers may use the dataset under Fair Use/Dealing law, as it is intended for non-commercial research and study, aligning with legal exemptions for academic purposes. By applying this license, we ensure open academic access and maintain compliance with Fair Use (US) and Fair Dealing (UK) provisions. Fair Use/Dealing permits the use of copyrighted material for academic purposes because it serves the public interest by enabling research, study, education, and transformative analysis without unfairly impacting the original work's commercial value.

The MIMIC-III database is released by MIT-LCP under PhysioNet's Restricted Data Use Agreement, which permits research use provided that users complete human-subjects/HIPAA training, keep the data secure and confidential, never attempt re-identification, do not redistribute the raw data, and share any publication-related code openly. The terms of this license and use agreement have been kept during the research outlined in this paper.